

Eliciting Inference-time Emergent Abilities of Large Language Models ·

Zhen Wan

March 2026

Abstract

Historically, the development of neural approaches in Natural Language Processing (NLP) established the modification of model parameters as the fundamental pathway for task-specific adaptation. In this classical paradigm, addressing a new task necessitated specialized training or fine-tuning to encode task-relevant knowledge directly into the model’s weights, as known as the “one model, one task” paradigm. However, the emergence of Large Language Models (LLMs) has fundamentally disrupted this trajectory, introducing a “one model, all tasks” approach. In this new landscape, because the massive scale of parameters often remains frozen during deployment, inference-time execution has emerged as the new valid stage for performance optimization. Rather than modifying weights, the model’s performance is now determined by the strategic construction of the inference process itself. We argue that the critical bottleneck of current AI systems lies not in a deficiency of their pre-trained knowledge, but in the inherent unreliability of inference-time execution when facing complex tasks, domain shifts, or noisy multimodal environments. This thesis investigates the inference-time emergent abilities of LLMs when specialized inference-time scaffolding activates the model’s core behaviors of context utilization and reasoning without parameter updates, and introduces advanced mechanisms across the above three key dimensions.

First, we address the performance gap in In-Context Learning (ICL), which means that models learn to perform tasks by observing a few demonstrations provided in the prompt without parameter updates. We argue that vanilla ICL often falls short of supervised models in structured regimes because it lacks the synergy between effective context utilization (selecting semantically aligned demonstrations) and robust reasoning (mapping those demonstrations to strict task

schemas). We propose task-aware retrieval and reasoning-enhanced mechanisms such as GPT-RE and GPT-MM, to bridge this gap by aligning external examples with internal logical constraints.

Second, we tackle the unreliability of domain adaptation under distribution shifts by introducing the Adapt-Retrieve-Revise framework. This mechanism enhances context utilization by using adapted smaller models to anchor precise evidence retrieval and scaffolds the internal reasoning of large models to perform reliable hypothesis revision.

Third, we extend this design to agentic multimodal systems, the most sophisticated frontier of scaffolding, where text serves as the universal medium for coordinating diverse modules. In these scenarios, a perception model (e.g., for speech recognition) translates sensory input into textual artifacts, which are then fed into the context of a downstream LLM for high-level decision-making. To bridge the persistent perception-reasoning gap, we introduce SpeechIQ, a framework for agent-centric evaluation that assesses the model’s capacity to utilize noisy perceptual context beyond verbatim accuracy. Furthermore, we propose SpeechHands, a voice-agentic framework that leverages inference-time self-reflection to resolve conflicts between internal logic and external hypotheses. These contributions demonstrate that agentic coordination can transform LLMs from passive text generators into resilient decision-makers capable of autonomous reasoning and meta-cognitive coordination in complex environments.

In conclusion, through extensive empirical evaluations across diverse benchmarks, this thesis establishes a comprehensive framework for guiding the emergent abilities of LLMs at inference time. This dissertation demonstrates that properly designed inference-time scaffolding can effectively substitute for costly full-parameter retraining, providing a robust path toward high-performance and resilient intelligence. By evolving LLMs into reflective agents, we pave the way for a more unified intelligence capable of navigating the uncertainties of the real world. Looking ahead, this work provides a foundation for the evolution toward inference-time training for real-time adaptation and autonomous agentic systems capable of self-managed scaffolding, ultimately aiming for a unified intelligence that can autonomously reason across all human sensory and cognitive dimensions.

Acknowledgments

As I complete this dissertation, a surreal sensation lingers in my mind. Have five years truly passed? It feels like a blink of an eye, yet the memories of the early days seem so distant, almost blurred. Perhaps it is not the temporal distance, but the fact that the person I was then and the person I am now are fundamentally different.

As a researcher transitioning from a different field into Natural Language Processing, my journey began in a state of persistent trepidation regarding the future. I still recall the exceptional softness of the laboratory sofa during those countless nights spent in the lab. For me, the most significant achievement of these five years has been the process of overcoming fear. I feared that my non-CS background would hinder my coding; I feared that my presentations were not refined enough to escape negative evaluation; and I feared that my painstakingly conceived innovations might ultimately prove futile. In essence, I was paralyzed by the fear of failure.

Yet, absolute success is an illusion. Even the most esteemed scholars face rejections. If success is inherently a probabilistic event, then mathematics teaches us that the only rational response is not to recoil from attempts, but to increase the number of trials. If my code was inadequate, I learned from others; if my presentations were unpolished, I revised them relentlessly; if experiments failed, I delved deeper into the analysis. When the number of trials increases, probability eventually rewards us.

Of course, the direction of effort is equally paramount, and the journey of confronting fear is a solitary one. I would like to express my deepest gratitude to the all those who have accompanied me through these five years:

First, great thanks to my supervisors, Professor Sadao Kurohashi and Senior Lecturer Fei Cheng: My growth and breakthroughs in every moment are inseparable from your sincere and professional guidance. Your strictness allowed me to identify problems with maximum efficiency, while your patience gave me the courage to face challenges and achieve breakthroughs.

I am also grateful to all the teachers and students in the laboratory, including Associate Professor Chenhui Chu, Associate Professor Murawaki Yugo, Assistant Professor Yin-Jou Huang, my mentor Shuichiro Shimizu, Dr. Hirokazu Kiyomaru, Dr. Qianying Liu, Dr. Zhuoyuan Mao, Dr. Haiyue Song, Dr. Nobuhiro Ueda, Dr. Kazumasa Omura, Zhengdong Yang, Yihang Li, Yikun Sun, Yahan Yu, Youyuan Lin, Xinger Fu, Shiho Matta, Chengzhi Zhong, Ruiyi Yan, Yang Liu, Zelin Zhang, Sirou Chen and Sakiko Yahata. A vibrant research atmosphere elevates everyone within it. Thank you for the meaningful academic discussions and the selfless mutual help that we shared. Also, I must express a special thanks to Terumi Kosugi for her meticulous administrative support. As someone who occasionally overlooks the practicalities of daily life, I am deeply grateful for her untiring reminders and assistance.

For external collaborators, I sincerely thank Professor Jiwei Li for his invaluable advice and timely feedback, which were essential to our impactful joint works. I also thank for the guidance and experience shared by Dr. Yating Zhang and Yaxiang Wang from Alibaba Damo Academy, Professor Qian Chen and Professor Haiyuan Liu from SenseTime Japan, Dr. Huck Yang and Dr. Hanrong Ye from Nvidia Research, and Professor Aizawa Akiko from the National Institute of Informatics.

Next, I am deeply indebted to Kyoto University and the Japan Society for the Promotion of Science (JSPS) for providing generous scholarship support and numerous seminar opportunities, which allowed me to broaden my academic network.

Finally, I want to thank my family. Living abroad meant I could not always be by your side, yet your unwavering support and lack of pressure gave me the freedom to pursue my dreams. Your help at every moment I needed it has been my ultimate anchor.

While the journey of doctoral research was demanding, it was, in many ways, a "single-head" chapter of my life—a period of intense, singular focus. In contrast, the future ahead promises to be a "multi-head" landscape, filled with diverse and complex challenges that extend beyond the laboratory. My hope is that the "sedimentation" of these past five years will empower me to step forward. I wish to face this next stage not with the trepidation of my early days, but with an open heart and the courage to embrace the unknown with confidence.

Contents

Abstract	i
Acknowledgments	iii
1 Introduction	1
1.1 Background of Inference-Time Emergent Abilities	1
1.1.1 Two Core Behaviors: The Dual Engines of Inference	2
1.1.2 Three Scaffolding Levers: Practical Pathways to Emergence	3
1.2 Challenges and Our Proposals	4
1.2.1 Challenge 1: Vanilla ICL breaks in complex scenarios	4
1.2.2 Challenge 2: Unreliable domain adaptation via retrieval- augmented generation	5
1.2.3 Challenge 3: Bridging the perception-reasoning gap in mul- timodal multi-agent pipelines	6
1.2.4 Organization of the Dissertation	7
2 GPT-RE	9
2.1 Introduction	9
2.2 Related Work	13
2.3 Methodology: GPT-RE	14
2.3.1 Task Definition	14
2.3.2 Overview	14
2.3.3 Prompt Construction	14
2.3.4 Task-aware Demonstration Retrieval	15

2.3.5	Gold Label-induced Reasoning	17
2.4	Experiment Setup	18
2.4.1	Datasets	18
2.4.2	Baseline Methods	19
2.5	Experimental Results	20
2.5.1	Main Results	20
2.5.2	Ablation Study on Task-aware Retrieval	22
2.5.3	Ablation Study on Reasoning Enhancing	23
2.5.4	Low-resource Scenario	23
2.6	Analysis	24
2.6.1	The Issue of “Overpredicting”	24
2.6.2	Case Study of Demonstration Quality	24
2.7	Limitations and Discussions	25
2.8	Summary of This Chapter	25
3	GPT-MM	28
3.1	Introduction	28
3.2	A Unified Multimodal Extension: GPT-MM	29
3.2.1	Task Definition	29
3.2.2	Overview	29
3.2.3	Prompt Construction	30
3.2.4	Task-aware Multimodal Retrieval	30
3.2.5	Label-induced Reasoning for Multimodal	31
3.3	Multimodal Experiment Results	32
3.3.1	Datasets	32
3.3.2	Baselines	33
3.3.3	Main Results	34
3.4	Summary of This Chapter	35
4	Adapt-Retrieve-Revise	37
4.1	Introduction	37
4.2	Methodology	40
4.2.1	Draft Answer Generation by the Domain-adapted LLM	41

4.2.2	Answer-based Evidence Retrieval	42
4.2.3	GPT-4 Revision	42
4.3	Experiment Settings	43
4.3.1	Model Settings	43
4.3.2	Evaluation of Chinese Legal Tasks	44
4.3.3	Evaluation Metrics	44
4.4	Main Experimental Results	46
4.5	Further Analysis	48
4.5.1	Analysis of Retrieval Methods	48
4.5.2	Case Analysis of the Improvements after the GPT-4 Revision	50
4.5.3	Does the iteration make the generation better?	51
4.6	Related Work	51
4.6.1	Chinese Legal Domain Tasks	51
4.6.2	Chinese Legal LLMs	52
4.6.3	Retrieval-augmented Inference	52
4.7	Summary of This Chapter	53
5	SpeechIQ	55
5.1	Introduction	55
5.2	Related Work	58
5.2.1	ASR Evaluation	58
5.2.2	LLM _{Voice} Understanding Systems	59
5.3	SpeechIQ: Speech Intelligence Quotient	60
5.3.1	Remember: WER Metric	60
5.3.2	Understand: Similarity Metric	61
5.3.3	Apply: QA Accuracy	62
5.3.4	Final SIQ Score	63
5.4	Experiment Setup	64
5.5	QA Generation	64
5.5.1	Datasets	65
5.5.2	Models	65
5.5.3	Setup Details	66

5.6	Experimental Results	69
5.6.1	Remember-Level Results	69
5.6.2	Apply-Level Results	70
5.6.3	SIQ Scores	70
5.6.4	Cascaded vs End-to-end	72
5.6.5	Unanswerable Set: Detecting Hallucination in LLM _{Voice} . .	74
5.6.6	End-to-end Models on the Cascaded Category	75
5.7	Case Study	75
5.8	Limitations and Discussions	77
5.9	Summary of This Chapter	78
6	Speech-Hands	80
6.1	Introduction	80
6.1.1	Preliminary: The Surprising Failure of Multimodal Correc- tion with Omni-LM	82
6.1.2	Prompt Ablations in Preliminary SFT	84
6.1.3	Zero-shot Qwen in Preliminary	85
6.2	Related Work	85
6.2.1	Voice Retrieval and Agentic Framework	85
6.2.2	Omni-Modality and Self-Reflection in Multimodal Models .	86
6.3	Methodology	86
6.3.1	Task Formulation	87
6.3.2	Action Token Construction for ASR	87
6.3.3	Action Token Construction for Audio QA	88
6.3.4	Prompt Formatting and Training	88
6.3.5	Agentic Inference via Action Tokens	89
6.4	Experimental Setup	90
6.4.1	Datasets	90
6.5	Prompt Template	91
6.6	Results	92
6.6.1	Training Details	92
6.6.2	Baselines	93

6.6.3	ASR Results	93
6.6.4	AudioQA Results	94
6.7	Additional Analysis	95
6.7.1	Accuracy of Action Token Prediction	95
6.7.2	Confusion Analysis In AudioQA	96
6.7.3	Category Analysis and Case Study	97
6.8	Limitations and Discussions	98
6.9	Summary of This Chapter	99
7	Conclusions and Future Work	101
7.1	Conclusions	101
7.2	Future Work	103
	Bibliography	106
A	Additional Experimental Details of GPT-RE	139
A.1	Hyperparameters	139
A.1.1	GPT-3 Hyperparameters	139
A.1.2	Fine-tuning Baseline PURE	139
A.1.3	Sentence Embedding Methods	140
A.2	Case Study	141
B	Additional Experimental Details of Adapt-Retrieve-Revise	142
B.0.1	Legal Instruction Tuning	142
B.0.2	Retrieval Modules	143
C	Additional Experimental Details of SpeechIQ	145
C.1	Complete Results	145
C.1.1	WER Results	145
C.1.2	Similarity Results	145
C.1.3	QA Accuracy Results	147
C.2	Scaling Law Normalization	147

D Additional Experimental Details of Speech-Hands	152
D.1 Failure Case of Multimodal GER	152
D.1.1 Amplifying Hallucination	152
D.1.2 Overcorrection	153
D.2 Dataset Details for DCASE2025 AudioQA	153
D.2.1 Bioacoustics QA	153
D.2.2 Temporal Soundscapes QA	154
D.2.3 Complex QA (MMAU)	154
D.3 Cases On ASR	155
D.4 Cases On AudioQA	156
List of Publications	163

List of Figures

2.1	Micro F1 performances on two RE datasets. Previous GPT baselines (<i>GPT-Random</i> : randomly selected demonstrations and <i>GPT-Sent</i> : sentence-level demonstration retrieval) largely underperform fine-tuning baseline PURE while our <i>GPT-RE</i> substantially outperforms all baselines.	10
2.2	Retrieval without considering the task-aware triplet results in noisy demonstrations.	11
2.3	Confusion matrix on Semeval dataset with three selected relation labels. The NULL examples are overpredicted to other relations by GPT-3. CE: Cause-Effect, IA: Instrument-Agency, PP: Product-Producer.	12
2.4	An illustration of GPT-RE. Given a test input, we first leverage two different task-aware retrieval methods to search for highly relevant demonstrations from the training set, and then incorporate the gold label-induced reasoning for each demonstration. Above contents will then be included in the prompt construction to make the prediction.	13
2.5	An illustration of adding reasoning.	17
2.6	Ablation study on the retrieval and reasoning components on Semeval. We sampled a subset from the test data with 300 examples. We show the ‘w/o reasoning’ results with $k = 30$ for comparison.	20

2.7	Low-resource Scenario on Semeval. We limit the percentage of training data for both fine-tuning and retrieval in GPT-RE. . .	21
2.8	Analysis on the effects of NULL examples. w/o NULL refers to the classification setting that NULL examples are excluded from the train and test data. w/ NULL refers to the original extraction setting. We use the full test set for the evaluation.	22
2.9	A case study of demonstration quality on Semeval. [NULL] is the gold label here.	27
4.1	Left: A real translated example of Chinese LegalQA. The square brackets and subscripts are offered for the purpose of clear demonstration, not actually exist in the ground-truth answer or generation. Right: Models' F1 scores on the LegalQA dataset.	38
4.2	Examples of hallucinations of various models. Red denotes the content containing hallucinations. The ground-truth answer refers to the left case in Figure 4.1.	40
4.3	Overview of our proposed method. The example and prompt are translated from Chinese to English for the demonstration purpose. . .	41
4.4	Comparison of retrieval recalls on the LegalQA dataset. . .	47
4.5	We compare performances of the draft answer of 7B legal LLM and our proposed adapt-retrieve-revise model using different contents in retrieval.	47
4.6	Case studies on the GPT-4 revision. The examples are translated from Chinese to English for the demonstration purpose. . . .	49
4.7	Comparison of iterations on LegalQA dataset.	51
5.1	An overview of cognitive levels in voice understanding large language model systems related to the first three foundational hierarchies of Bloom's Taxonomy. We design a corresponding examination to measure SpeechIQ as a detailed pipeline in Figure 5.2. . . .	56

5.2	Overview of Speech IQ (SIQ) test. We compared cognition-inspired categories of LLM _{Voice} represented in three rows, with the right columns denoting the SIQ sub-tests. The truth reference will be used in the sub-tests.	59
5.3	Prompt for QA generation	65
5.4	Prompt for end-to-end models	68
5.5	Prompt for QA generation	68
5.6	Raw remember scores. All cascaded models outperform end-to-end models including Gemini-1.5.	69
5.7	Raw understand scores. Gemini-1.5-flash shows the best semantic capturing.	70
5.8	Raw apply scores. End-to-end-large models perform best while smaller ones perform worst.	71
5.9	Left: Llama3-8B-Instruct, Right: Desta2. Red the errors in answering, we find that Llama-8B-Instruct based on the error ASR even hallucinate to change the option in QA, while Desta2 though answer correctly, still inherit this hallucination.	76
5.10	Left: Llama3-8B-Instruct, Right: Desta2. Red the errors in answering, we find that Llama-8B-Instruct gives the correct answer but based on the error ASR and also hallucinate to change the option in QA, while Desta2 keeps the same hallucination.	76
6.1	<i>Speech-Hands</i> acts as a dynamic orchestrator that predicts a special action token to govern its cognitive strategy.	81
6.2	Overview of our proposed Self-Reflection Multimodal GER framework. A special token is generated at the beginning to decide whether to use audio perception (i.e., transcription hypotheses or caption) or not.	81

6.3	Preliminary results on the cascaded agentic Qwen-omni baseline for generative error correction (GER) with supervised fine-tuning show that both text-only and text-audio GER degrade ASR performance, where the best ASR and LLM combination achieves a low agentic output oracle of 5% WER.	83
6.4	Text-base GER uses only ASR hypotheses. This setup fails to correct deletion or hallucination if all hypotheses are wrong. Multimodal GER include the original audio as grounding to improve error correction.	84
6.5	Confusion matrices of Speech-Hands' agentic action execution for audio QA and reasoning three subsets on (a) bio-acoustic QA, (b) temporal and sound event QA, and (c) complex audio information QA.	96
A.1	More casees.	141
C.1	WER on earning22	146
C.2	WER on medasr	146
C.3	WER on voxpopuli	147
C.4	Similarity on earning22	148
C.5	Similarity on medasr	148
C.6	Similarity on voxpopuli	149
C.7	QA accuracy on earning22	149
C.8	QA accuracy on medasr	150
C.9	QA accuracy on voxpopuli	150

List of Tables

2.1	Statistics of datasets.	18
2.2	Main Results on four RE datasets. All results are given by Micro-F1. * denotes the same k -shot for the comparison with + Reasoning. Due to the costly GPT-3 expense, we conducted Reasoning experiments on the two relatively smaller datasets Semeval and TACRED. ♣ denotes that this performance is not comparable as it evaluates on the entire test set. The <u>underline</u> denotes the results outperforming the fine-tuning baseline PURE.	19
3.1	Main results on VQA and ASR tasks under 3-shot in-context learning. GPT-MM achieves consistent improvements by using task-aware answer-state embeddings for retrieval. Note that due to the limited length of multimodal inputs, we restrict the number of demonstrations to be 3.	34
4.1	Zero-shot Results on four Chinese legal datasets. “Rec.” denotes recall, “Acc.” denotes accuracy, and “Avg.” is computed by the F1 and accuracy scores of all four tasks	45
4.2	Statistics of datasets.	45
4.3	Results of two retrieval setups on the Similar Case Retrieval dataset.	49
5.1	Statistics of datasets. WER results are on whole dataset via Whisper-large-v2.	66

5.2	Results of same ASR models with different LLMs in answering QA.	67
5.3	Main Results. The results of remember and understand levels are from single experiment, while apply level include the majority results of 5-time generations.	72
5.4	Human Voted Rankings & Existing Metrics.	73
5.5	Cascaded v.s. End-to-end including hallucination detection. Cascaded models perform better correspondingly. Dsta2 shows unexpectedly high QA accuracy and hallucinations along with its foundation Llama3 _{8B-ins} .	74
5.6	Take ASR from end-to-end models	75
5.7	Human test on 3 examples.	77
6.1	Prompt ablation results.	85
6.2	Confusion matrix of zero-shot decisions under different prompting strategies. The results show that the model’s arbitration is highly sensitive to prompt wording rather than the ground truth correctness (Oracle), often collapsing into trivial heuristics.	85
6.3	WER (%) results across 7 datasets, with the average WER shown in the rightmost column. Speech-Hands training significantly outperforms both baseline systems (ASR model, Qwen) and prior cascaded GER setups.	89
6.4	Number of training samples used from each dataset. For GigaSpeech, only 9,389 valid samples met our filtering criteria.	90
6.5	AudioQA and acoustic content reasoning accuracy (%) across knowledge-intensive bioacoustic QA [158], multi-sound-object soundscapes, and MMAU-style [154] complex audio QA tasks.	92
6.6	Training distribution and test-time F1 scores for Speech-Hands’ action tokens.	94
6.7	Oracle Statistics of the three DCASE2025 AudioQA test sets used in our experiments.	96
A.1	GPT-3 Hyperparamters.	140

A.2 Search range for each dataset. 140

B.1 **Results of different retrieval seteps.** 144

D.1 Statistics of the three DCASE2025 AudioQA subsets used in our experiments. **<in>/<ex>/<re>** shows the token distribution in training w/ majority sampling. 153

Chapter 1

Introduction

1.1 Background of Inference-Time Emergent Abilities

The evolution of Natural Language Processing (NLP) is characterized by a series of fundamental shifts in how models acquire and apply linguistic knowledge. Historically, the field was governed by statistical approaches, where language understanding relied heavily on manual feature engineering and probabilistic models such as n-grams [92, 16] and Hidden Markov Models (HMMs) [137]. In this era, performance was a direct reflection of human-designed rules, and systems were inherently brittle, lacking the capacity for general semantic representation.

The transition toward the neural network era, marked by the advent of word embeddings (e.g., Word2Vec [121]) and sequence-to-sequence architectures [167] (e.g., LSTMs [56] and CNNs [83]), introducing a paradigm of automated feature learning. However, these models were still task-specific; addressing a new problem required training a dedicated model from scratch on curated datasets. This trajectory culminated in the BERT-era [37, 8, 144, 90, 217], which popularized the "Pre-train then Fine-tune" pipeline. While BERT and its variants enabled powerful context-aware representations, they still operated within a "one model, one task" ecosystem, where the primary lever for optimization remained the modification of model parameters through supervised weight updates for each specific application.

The rise of Large Language Models (LLMs) [17, 173, 29, 142, 57] has fundamen-

tally disrupted this trajectory, ushering in a "one model, all tasks" paradigm. With billions of parameters and vast pre-training on diverse corpora, LLMs demonstrate a remarkable ability to generalize across thousands of tasks through zero-shot or few-shot prompting [17, 122, 214, 98, 162]. In this new landscape, the model parameters are often frozen due to computational costs or API accessibility—shifting, the locus of performance optimization from training-time weight updates to inference-time execution design.

However, this transition also exposes a critical reliability gap. While LLMs possess vast pre-trained knowledge, they remain notoriously brittle in settings that demand strict output constraints, faithful domain grounding, or robust multimodal interaction. Naive prompting often leads to inconsistent format adherence, hallucinations under domain shift, and reasoning failures in noisy environments.

This dissertation argues that many of these limitations are not solely due to insufficient model capacity, but rather a failure of how inference is executed. When parameters are frozen, model performance becomes highly sensitive to the inference process: what context is provided, how it is structured, and how the model's internal reasoning is elicited. I therefore focus on inference-time design as the primary pathway to bridge the performance gap between general-purpose models and task-specific excellence without retraining. I refer to the resulting latent skills as inference-time emergent abilities: capabilities such as high-precision reasoning, context-sensitive adaptation, and autonomous self-reflection that already exist within the model's weights but can only be "unlocked" through appropriate inference-time scaffolding.

1.1.1 Two Core Behaviors: The Dual Engines of Inference

I argue that the successful execution of inference-time design relies on the synergy between two core behaviors, which serve as the primary engines for emergent capabilities.

- **Context Utilization (External Pillar):** This behavior refers to the model's ability to effectively ground its generation in external information provided in the prompt. It requires the LLM to identify relevant signals (such as

few-shot demonstrations, retrieved evidence, or perceptual artifacts) and integrate them consistently into its representation.

- **Reasoning (Internal Pillar):** Complementing external grounding, this behavior involves the internal cognitive processing required to follow task-specific rules and align with required structures. Include logical deduction, format adherence, and the ability to revise hypotheses when new evidence contradicts initial priors.

However, these internal and external engines often operate in isolation or remain underutilized, leading to the unreliability observed in complex tasks. To bridge the gap between pre-trained knowledge and robust execution, this research posits that these behaviors must be systematically activated through specific **inference-time scaffolding**. These scaffolds provide the necessary structural guidance to balance context and reasoning, thereby unlocking the inference-time emergent abilities that are the focus of this dissertation.

1.1.2 Three Scaffolding Levers: Practical Pathways to Emergence

To manifest these core behaviors in real-world systems, I identify three practical inference-time scaffolding levers. Each lever represents a distinct architectural approach to structuring the interaction between the model and its task environment.

1. In-Context Learning (ICL): Learning through Demonstration Scaffolding. ICL [17, 98, 162, 152] represents a paradigm shift where task mappings are taught at inference time by providing structured input-output pairs. This lever primarily targets context utilization, challenging the model to extract and imitate latent relational patterns from external demonstrations. As we will discuss in Challenge 1, the effectiveness of this scaffolding in structured regimes depends on whether the model can align these external signals with its internal logical constraints.

2. Retrieval-Augmented Generation (RAG): Grounding through Evidence Scaffolding. RAG [94] addresses the factuality gap by injecting dynamic, external evidence into the generation process. This scaffolding creates a complex interplay where the model must not only utilize context by filtering retrieved evi-

dence but also employ robust reasoning to perform hypothesis revision when that evidence contradicts its internal priors. Challenge 2 explores how this scaffolding can be made resilient for domain-specific adaptation.

3. Agentic Multimodal Inference: Coordination through Systems Scaffolding. The most advanced form of scaffolding involves placing LLMs as central cognitive hubs within multimodal pipelines. Here, the model must coordinate between noisy external perceptual signals (Context) and its own meta-cognitive reasoning (Self-reflection). This scaffolding transforms the LLM from a passive text processor into an active agent, necessitating new paradigms for evaluation and conflict resolution, as examined in Challenge 3.

The remainder of the introduction is organized around three challenges that arise when these levers are applied in realistic settings.

1.2 Challenges and Our Proposals

1.2.1 Challenge 1: Vanilla ICL breaks in complex scenarios

As the foundational lever of inference-time scaffolding, ICL allows LLMs to infer new tasks from a few demonstrations without parameter updates. However, a significant paradigm gap persists in structured, supervision-favored regimes, such as relation extraction (RE), where BERT-era fine-tuned pipelines historically outperform vanilla ICL [74]. This dissertation argues that the underperformance of vanilla ICL in these domains stems from a dual breakdown in core inference behaviors. First, context utilization is hindered by poor demonstration alignment; when the model fails to identify and utilize external examples that match the query’s semantic and entity-type patterns, it may overfit to surface cues rather than the intended task boundary. This challenge is further amplified in multimodal settings, where the model must not only align textual semantics but also ensure cross-modal consistency between visual features and textual logic. Second, the model’s internal reasoning often lacks explicit, label-grounded logic. Without a structured reasoning trace that maps the utilized context—especially complex multimodal signals—to strict schema constraints, the model relies on shallow pattern matching, making it unreliable for the fine-grained understanding tasks essential for downstream

applications.

To bridge this gap and unlock the latent potential of LLMs in both textual and multimodal tasks, we propose an enhanced inference-time design centered on task-aware retrieval and reasoning scaffolds. We introduce GPT-RE and its multimodal extension GPT-MM, which optimize context utilization by selecting demonstrations based on deep relational similarity and cross-modal relevance rather than mere lexical overlap. By constructing reasoning-heavy prompts that elicit explicit evidence-to-label mappings, these frameworks effectively scaffold the model’s internal reasoning to follow strict task schemas and resolve multimodal ambiguities. Our findings demonstrate that this structured approach allows LLMs to emerge as high-precision reasoners capable of matching or surpassing supervised baselines in both text-only and multimodal environments. Ultimately, this work establishes that the paradigm gap can be closed by sophisticated inference-time mechanisms that effectively align heterogeneous external context with internal logical constraints.

1.2.2 Challenge 2: Unreliable domain adaptation via retrieval-augmented generation

While state-of-the-art Large Language Models (LLMs) demonstrate astonishing capabilities in general tasks, they frequently suffer from hallucinations in specialized domains, such as Chinese law, where they lack specific in-domain knowledge. The conventional pathway for adaptation—continual fine-tuning—is often impractical for large-scale models due to prohibitive computational costs and the difficulty of acquiring exhaustive domain data. Within the framework of inference-time scaffolding, vanilla retrieval-augmented generation (RAG) often fails because the model possesses no prior domain intuition to effectively guide context utilization, leading to inconsistent grounding when retrieved evidence is noisy. Furthermore, without a structured inference procedure, the model’s internal reasoning lacks the necessary scaffolding to reliably assess evidence and revise its initial hypotheses, hindering its ability to achieve the precision required for high-stakes domain applications.

To address these limitations, this dissertation introduces a domain adaptation

framework that reformulates the generation process as an Adapt-Retrieve-Revise procedure. We first adapt an affordable 7B LLM to the target domain to generate an initial draft answer, which then serves as a knowledgeable anchor for retrieving precise supporting evidence from an external knowledge base. This mechanism optimizes context utilization by providing high-quality, domain-relevant signals that vanilla retrieval often misses. Finally, a larger model (e.g., GPT-4) is employed to assess the retrieved context and revise the draft, unlocking an emergent capability for inference-time self-correction. Our empirical results across Chinese legal tasks show that this structured scaffolding improves accuracy by 33.3% over direct generation, demonstrating that the synergy between efficient small-model adaptation and large-model reasoning can effectively bridge the domain gap without the need for full-parameter retraining.

1.2.3 Challenge 3: Bridging the perception-reasoning gap in multimodal multi-agent pipelines

As the most sophisticated frontier of inference-time scaffolding, agentic multimodal systems represent a pivotal paradigm shift: Large Language Models (LLMs) are no longer mere recipients of static text prompts, but serve as the central cognitive hubs within complex, multi-agent ecosystems. This scaffolding requires the model to perform a high-level orchestration, coordinating the flow between heterogeneous external perceptual signals and internal logical processing. However, as we transition from isolated prompting to this agentic environment, two fundamental bottlenecks emerge that challenge the reliability of the model’s emergent capabilities. First, existing evaluation paradigms are largely human-centric, focusing on surface-level transcription fidelity (e.g., WER) rather than the emergent utility of the context for downstream agents. A transcript may be verbatim accurate but semantically insufficient for an agent to act, creating a gap in measuring the model’s context utilization within an agentic workflow. Second, agentic scaffolding exposes a persistent perception-reasoning gap, where noisy external hypotheses (perceptual context) frequently contradict the model’s internal prior knowledge (reasoning). Without a mechanism to navigate these conflicts, the system is prone to cascading errors, as vanilla models lack the meta-cognitive ability to reflect on the reliability

of the external inputs they utilize.

To bridge this gap, this dissertation proposes that agentic scaffolding can unlock sophisticated evaluative and reflective abilities at inference time through targeted design. We first introduce SpeechIQ, which reframes multimodal evaluation as an emergent capability of LLMs to audit external perceptual context across tiered cognitive levels, shifting the metric from human-facing similarity to agent-centric utility. This framework allows the model to utilize noisy context more effectively by assessing whether it provides the semantic grounding necessary for complex application. Furthermore, we propose Speech-Hands, a voice-agentic framework that scaffolds the model’s internal reasoning through a learnable self-reflection primitive. This mechanism enables the model to resolve the conflict between external perception and internal knowledge by autonomously deciding at inference time whether to trust its own reasoning or verify the external signals. Together, these proposals demonstrate that agentic scaffolding, when properly designed, enables LLMs to evolve from passive processors into resilient decision-makers capable of high-level meta-cognitive coordination.

1.2.4 Organization of the Dissertation

This dissertation is organized into seven chapters, tracing the progression from theoretical foundations to the implementation of advanced inference-time scaffolding.

Chapter 1 establishes the conceptual framework, defining the paradigm shift toward inference-time design and identifying the two core behaviors (context utilization and reasoning) that drive emergent abilities in LLMs.

Chapters 2 and 3 focus on the first lever of scaffolding: In-Context Learning (ICL). Chapter 2 introduces GPT-RE, a framework that optimizes context utilization for structured relation extraction. Chapter 3 extends this to multimodal tasks through GPT-MM, demonstrating how task-specific retrieval and reasoning allow LLMs to emerge as high-precision reasoners across different modalities.

Chapter 4 addresses Retrieval-Augmented Generation (RAG) as a second lever for domain-specific adaptation. Using the Chinese legal domain as a case study, this chapter introduces the Adapt-Retrieve-Revise framework. It illustrates the synergy between external grounding and internal reasoning required to maintain

reliability under a significant domain shift.

Chapters 5 and 6 explore the most advanced scaffolding lever: Agentic Multimodal Inference. Chapter 5 introduces SpeechIQ, a cognitive-level evaluation framework that treats the LLM as an agent auditing perceptual signals. Chapter 6 proposes Speech-Hands, a mechanism that employs self-reflection to resolve perception-reasoning conflicts in omni-perception environments. Together, these chapters demonstrate how system-level scaffolding transforms LLMs into autonomous decision-makers.

Finally, Chapter 7 synthesizes the findings of this research and discusses future directions.

Chapter 2

GPT-RE: In-context Learning for Relation Extraction using Large Language Models

2.1 Introduction

The emergence of large language models (LLMs) such as GPT-3 [17, 173, 29, 142, 57] represents a significant advancement in natural language processing (NLP). Instead of following a pretraining-and-finetuning pipeline [37, 8, 144, 90, 217], which finetunes a pre-trained model on a task-specific dataset in a fully-supervised manner, LLMs employ a new paradigm known as in-context learning (ICL) [17, 123] which formulates an NLP task under the paradigm of language generation and makes predictions by learning from a few demonstrations. Under the framework of ICL, LLMs achieve remarkable performance rivaling previous fully-supervised methods even with only a limited number of demonstrations provided in various tasks such as solving math problems, commonsense reasoning, text classification, fact retrieval, natural language inference, and semantic parsing [17, 122, 214, 98, 162].

Despite the overall promising performance of LLMs, the utilization of ICL for relation extraction (RE) is still suboptimal. RE is the central task for knowl-

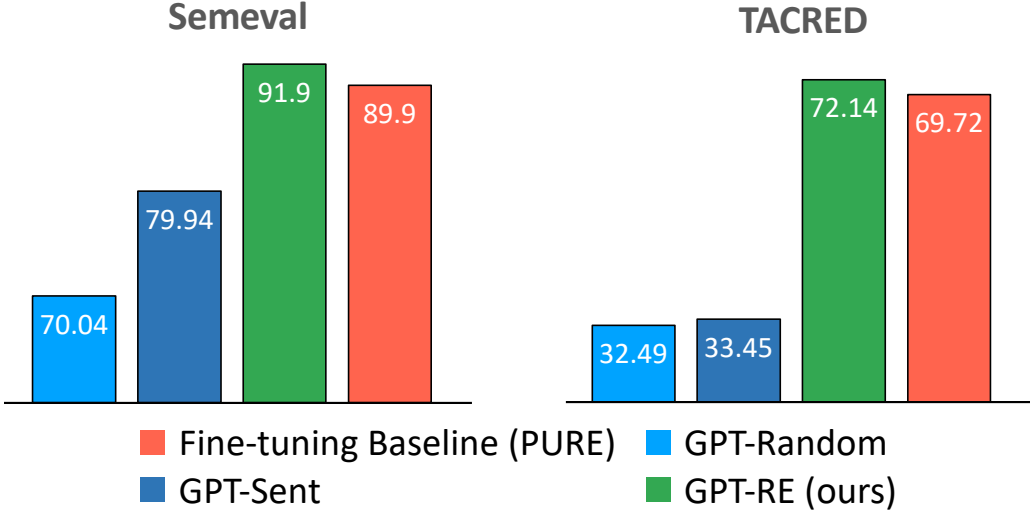


Figure 2.1: **Micro F1 performances on two RE datasets.** Previous GPT baselines (*GPT-Random*: randomly selected demonstrations and *GPT-Sent*: sentence-level demonstration retrieval) largely underperform fine-tuning baseline PURE while our *GPT-RE* substantially outperforms all baselines.

edge retrieval requiring a deep understanding of natural language, which seeks to identify a pre-defined relation between a specific entity pair mentioned in the input sentence or NULL if no relation is found. Given a test input, ICL for RE prompts the input of LLMs with the task instruction, a few demonstrations retrieved from the training data, and the test input itself. Then LLMs generate the corresponding relation. Recent research [74] has sought to apply GPT-3 ICL to biomedical RE, but the results are relatively negative and suggest that GPT-3 ICL still significantly underperforms fine-tuned models.

The reasons that cause the pitfall of GPT-3 ICL in RE are two folds: (1) The low relevance regarding entity and relation in the retrieved demonstrations for ICL. Demonstrations are selected randomly or via k -nearest neighbor (k NN) search based on sentence embedding [98, 74]. Regrettably, k NN-retrieval based on sentence embedding is more concerned with the relevance of the overall sentence semantics and not as much with the specific entities and relations it contains, which leads to low-quality demonstrations. As shown in Figure 2.2, the test input

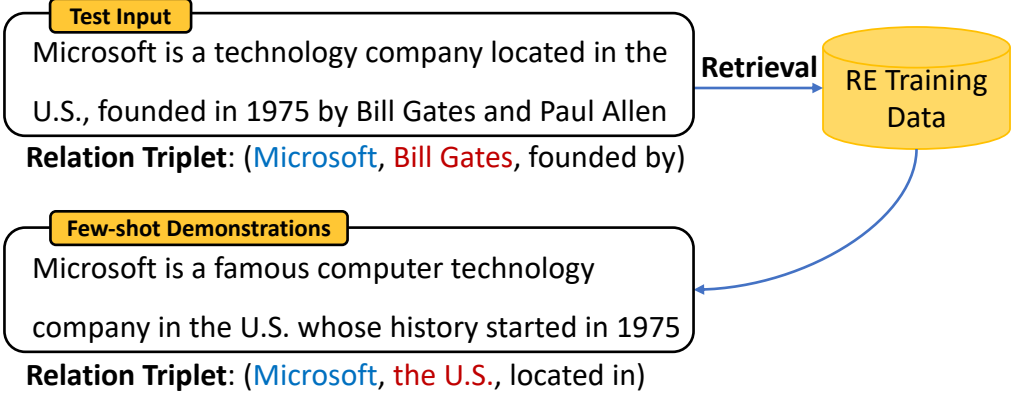


Figure 2.2: Retrieval without considering the task-aware triplet results in noisy demonstrations.

retrieves a semantically similar sentence but is not desired in terms of entities and relations.

(2) The lack of explaining input-label mappings in demonstrations leads to poor ICL effectiveness: A vanilla form of ICL lists all demonstrations as input-label pairs without any explanations. This may mislead LLMs to learn shallow clues from surface words, while a relation can be presented in diverse forms due to language complexity. Especially when ICL has a maximal input length, optimizing the learning efficiency of each single demonstration becomes extremely important.

To this end, we propose GPT-RE for the RE task. GPT-RE employs two strategies to resolve the issues above: (1) **task-aware retrieval** and (2) **gold label-induced reasoning**. For (1) task-aware retrieval, its core is to use representations that deliberately encode and emphasize entity and relation information rather than sentence embedding for k NN search. We achieve this by two different retrieval approaches: (a) entity-prompted sentence embedding; (b) fine-tuned relation representation, which naturally places emphasis on entities and relations. Both methods contain more RE-specific information than sentence semantics, thus effectively addressing the problem of low relevance.

For (2) gold label-induced reasoning, we propose to inject the reasoning logic into the demonstration to provide more evidence to align an input and the label, a strategy akin to the Chain-of-Thought (CoT) research [187, 184, 87]. But different

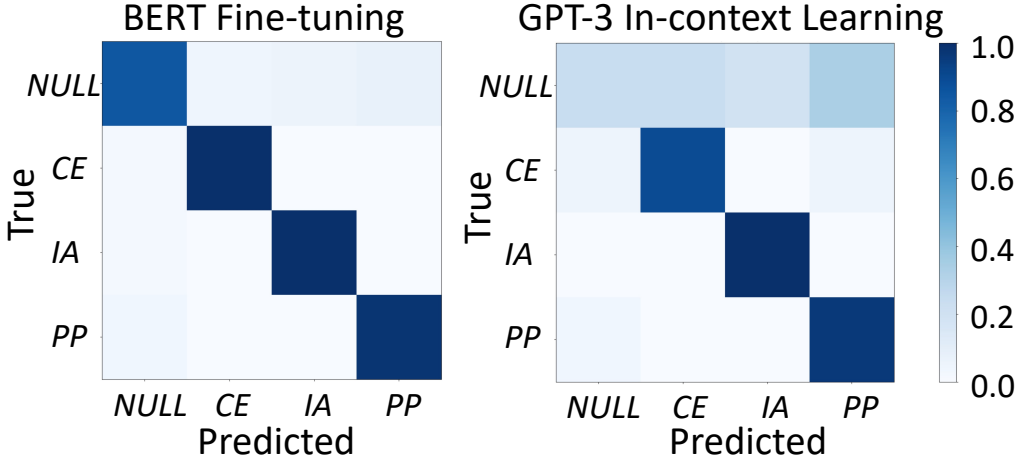


Figure 2.3: Confusion matrix on Semeval dataset with three selected relation labels. The NULL examples are overpredicted to other relations by GPT-3. CE: Cause-Effect, IA: Instrument-Agency, PP: Product-Producer.

from previous work, we allow LLMs to elicit the reasoning process to explain not only why a given sentence should be classified under a particular label but also why a NULL example should not be assigned to any of the pre-defined categories. This process significantly improves the ability of LLMs to align the relations with diverse expression forms.

Recent work reveals another crucial problem named “overpredicting” as shown in Figure 2.3: we observe that LLMs have the strong inclination to wrongly classify NULL examples into other pre-defined labels. A similar phenomenon has also been observed in other tasks such as NER [74, 10]. In this paper, we show that this issue can be alleviated if the representations for retrieval can be supervised with the whole set of NULL in the training data.

We evaluate our proposed method on three popular general domain RE datasets: Semeval 2010 task 8, TACRED and ACE05, and one scientific domain dataset SciERC. We observe that GPT-RE achieves improvements over not only existing GPT-3 baselines, but also fully-supervised baselines. Specifically, GPT-RE achieves SOTA performances on the Semeval and SciERC datasets, and competitive performances on the TACRED and ACE05 datasets.

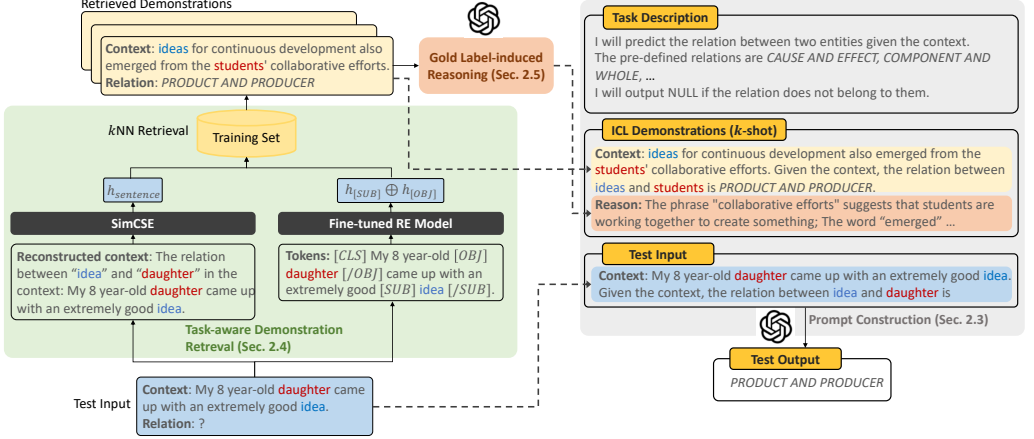


Figure 2.4: **An illustration of GPT-RE.** Given a test input, we first leverage two different task-aware retrieval methods to search for highly relevant demonstrations from the training set, and then incorporate the gold label-induced reasoning for each demonstration. Above contents will then be included in the prompt construction to make the prediction.

2.2 Related Work

In-context Learning Recent work shows that ICL of GPT-3 [17] can perform numerous tasks when provided a few examples in a natural language prompt. Existing work focuses on various aspects to effectively utilize the advantages of GPT-3, from prompt design [134] for proper input to coherence calibration [115] for tackling the diverse generated output. Another research path locates in the demonstration part, including ordered prompts [104] and retrieval-based demonstrations [152, 98, 162].

To the best of our knowledge, there is no previous work exploring the potential of GPT-3 on general domain RE tasks. A recent work attempts to leverage GPT-3 in biomedical information extraction (NER and RE), and reveals issues of ICL that may be detrimental to IE tasks in general. Our work succeeds in overcoming these issues to some extent and confirms the potential of GPT-3 in both general and the scientific domain RE.

Retrieval-based Demonstrations Several studies have demonstrated that dynamically selecting few-shot demonstrations for each test example, instead of utilizing a fixed set, leads to significant improvement in GPT-3 ICL [98, 162, 152]. They also show that nearest neighbor in-context examples yield much better results than the farthest ones. This leads to the significance of better retrieval modules for demonstrations. Existing attempts rely on sentence embedding in retrieval, including the sentence encoders of PLMs such as BERT [37], RoBERTa [217] KATE [98], SimCSE [43], Sentence-BERT [148, 190]. Unlike these sentence embeddings, we propose to fine-tune PLMs on our target RE tasks to produce more task-specific and robust representations for retrieval.

2.3 Methodology: GPT-RE

2.3.1 Task Definition

Let $\mathcal{C}$ denote the input context and $e_{\text{sub}} \in \mathcal{C}$, $e_{\text{obj}} \in \mathcal{C}$ denote the pair of subject and object entity. Given a set of pre-defined relation classes $\mathbb{R}$, relation extraction aims to predict the relation $y \in \mathbb{R}$ between the pair of entities $(e_{\text{sub}}, e_{\text{obj}})$ within the context $\mathcal{C}$, or if there is no pre-defined relation between them, predict $y = \text{NULL}$.

2.3.2 Overview

We will first introduce the prompt construction to formalize RE as a language generation task in Sec. 2.3.3. Then to improve the ICL framework for RE, we will introduce two modules: (1) task-aware demonstration retrieval to select higher-quality demonstrations (Sec. 2.3.4); (2) gold label-induced reasoning to enrich each demonstration with explanations (Sec. 2.3.5). In Figure 6.1, we show the concrete workflow of processing a test input.

2.3.3 Prompt Construction

We construct a prompt for each given test example, which is fed to the GPT-3 model. Each prompt consists of the following components:

Instructions $\mathcal{I}$ We provide a succinct overview of the RE task description and the set of pre-defined classes $\mathbb{R}$. The model is explicitly asked to output the relation, which belongs to the pre-defined classes. Otherwise, the model will output NULL.

ICL Demonstrations $\mathcal{D}$ We first leverage a task-aware retriever to acquire a k -shot demonstration set, then enrich each demonstration (x_i, y_i) with the gold label-induced reasoning r_i to build a new set of (x_i, y_i, r_i) as $\mathcal{D}$.

Test Input x_{test} Similar to the demonstrations, we offer the test input x_{test} , and GPT-3 is expected to generate the corresponding relation y_{test} .

In summary, GPT-RE can be formulated as:

$$p(y_{test} \in \mathbb{R} \cup \{\text{NULL}\} | \mathcal{I}, \mathcal{D}, x_{test}) \quad (2.1)$$

2.3.4 Task-aware Demonstration Retrieval

Since ICL demonstrations closer to the test sample in the embedding space result in more consistent and robust performance [98]. Recent work [74, 98] employs the k NN to retrieve the most similar examples in the training set as the few-shot demonstrations for each test input. As k NN relies on the choice of the embedding space to encode both test input and examples in the training set, they propose to obtain sentence embedding using pre-trained language models, or other improved sentence embedding.

However, using sentence embedding for k NN retrieval has a severe drawback: relation extraction focuses on pair-wise entities, which diverge from the semantic meaning of the entire sentence, leading to an ambiguous retrieval using sentence embedding. In this study, we propose two novel methods to provide more robust representations for better retrieval quality: (1) a naive entity-prompted sentence embedding in Sec. 2.3.4; (2) an advanced fine-tuned relation representation in Sec. 2.3.4.

Entity-Prompted Sentence Embedding

Given the discrepancy between sentence embedding and relation extraction, the original context is insufficient for demonstration retrieval. Considering the importance of entity information in RE, we propose reconstructing the context by incorporating entity pair information. For example, given the context “He has a sister Lisa,” the reconstructed context with the entity prompted will be “The relation between ‘He’ and ‘Lisa’ in the context: He has a sister Lisa.” This approach preserves both the semantic meaning of the sentence and the entity pair-centered information during retrieval. In the paper, we employ the latest robust model SimCSE [43] for computing sentence embedding-based similarity.

Fine-tuned Relation Representation

Compared to prompt entity information into context sentences, a more straightforward solution is to extract the relation representation from a fine-tuned RE model for retrieving demonstrations.

Current BERT-based fine-tuning methods for RE [6, 216, 179] attempts to capture both the context information and the entity information by adding extra marker tokens to highlight the subject and object entities and their types.

Specifically, given an example: “He has a sister Lisa,” the input tokens are “[CLS] [SUB_PER] He [/SUB_PER] has a sister [OBJ_PER] Lisa [/OBJ_PER]. [SEP]” where “PER” is the entity type if provided. Denote the n -th hidden representation of the BERT encoder as $\mathbf{h}_n$. Assuming i and j are the indices of two beginning entity markers [SUB_PER] and [OBJ_PER], we define the relation representation as $\mathbf{Rel} = \mathbf{h}_i \oplus \mathbf{h}_j$ where $\oplus$ stands for concatenation of representations in the first dimension. Subsequently, this representation is fed into a feedforward network for predicting the relation probability $p(y \in \mathbb{R} \cup \{\text{NULL}\} \mid \mathbf{Rel})$.

The entity markers have explicitly encoded subject and object entities and the relation representation $\mathbf{Rel}$ is naturally enriched with the entity information. We believe this approach can potentially compensate for the limitations of GPT-3 in RE. While GPT-3 ICL has a constraint of limited demonstrations, the fine-tuning process is unbundled and can be done on the whole train data. It has two

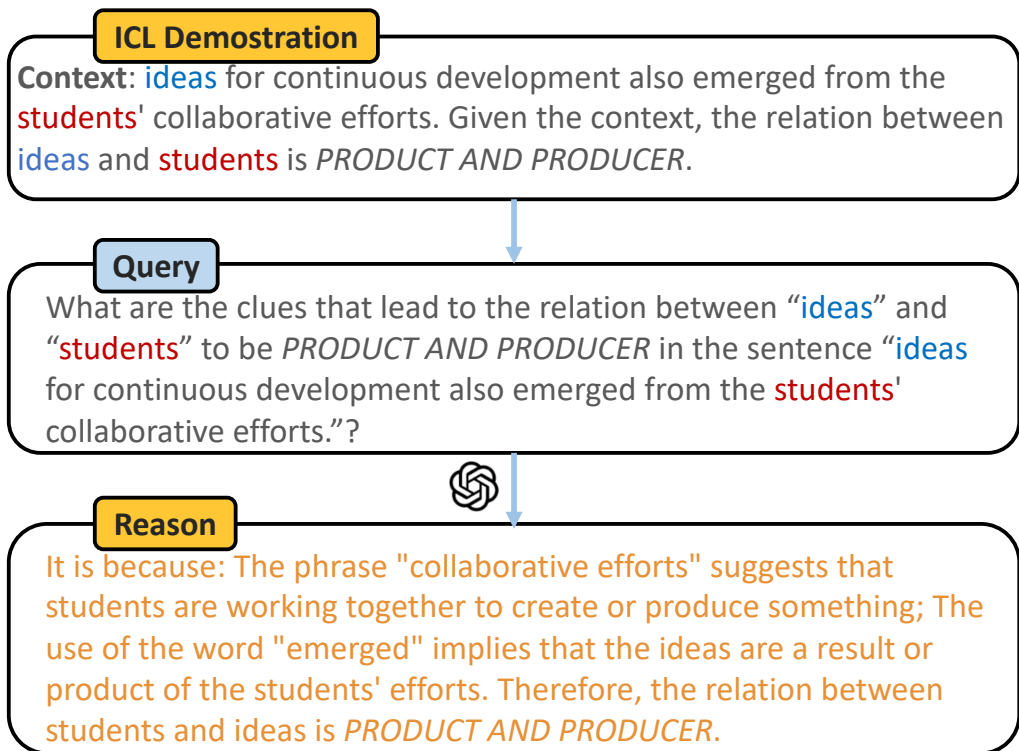


Figure 2.5: An illustration of adding reasoning.

subsequent merits. First, the relation representations are directly fine-tuned to fit the RE task, which could significantly boost the overall retrieval quality. Second, the overpredicting NULL issue will be substantially alleviated because the similar NULL demonstrated can be accurately recognized by the fine-tuned model.

2.3.5 Gold Label-induced Reasoning

Recent CoT work has reported significant progress in the commonsense and numerical reasoning tasks by automatically eliciting the reasoning steps for solving a question. While in the RE task, two entities can possibly hold multiple relations, e.g., “Joe Biden” can be either the president of or lives in “U.S.”. The reasoning generation could be out of focus if it lacks interaction with the gold label.

In this section, we propose to let GPT-3 induce the reasoning logic for each demonstration by the corresponding gold relation label. As shown in Figure 2.5,

Dataset	# Relation	# Train	# Dev	# Test (# Subset)	NULL (%)
Semeval	9	6,507	1,493	2,717 (2,717)	17.40%
TACRED	41	68,124	22,631	15,509 (1,600)	79.40%
SciERC	7	16,872	2,033	4,088 (4,088)	90.16%
ACE05	6	121,368	27,597	24,420 (2,442)	95.60%

Table 2.1: **Statistics of datasets.**

given a selected demonstration, we first generate a query prompt “What are the clues that lead to the relation between [entity1] and [entity2] to be [relation] in the sentence [context]?” based on the demonstration and subsequently ask GPT-3 to generate clues “It is because: ...” on the labeled relation between the pair of entities in the context. Finally, we augment the demonstration by incorporating the generated clues induced by GPT-3.

2.4 Experiment Setup

2.4.1 Datasets

We evaluate on three popular general domain RE datasets and one scientific domain dataset. Due to the cost of running the model in the API with GPT-3, in our main results, we sample a subset (will be released) from the original test set for two datasets: ACE05 and TACRED as shown in Table 2.1.

Semeval 2010 task 8 [53] focuses on semantic relations between pairs of nominals collected from general domain resources.

TACRED [212] is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text.

SciERC [105] collects AI paper abstracts and annotated relations, especially for scientific knowledge graph construction.

Methods	Retriever	Semeval	TACRED	SciERC	ACE05
<i>GPT-3 Baselines (Best k-shot)</i>					
GPT-Random	-	70.04 (30)	32.49 (15)	17.92 (25)	9.04 (25)
GPT-Sent	SimCSE	79.94 (30)	33.45 (15)	20.96 (25)	6.31 (25)
<i>Ours (Best k-shot)</i>					
GPT-RE_SimCSE	SimCSE	81.02 (30)	37.44 (15)	26.46 (25)	8.67 (25)
GPT-RE_SimCSE*	SimCSE	77.49 (15)	31.58 (10)	-	-
+ Reasoning	SimCSE	79.88 (15)	33.18 (10)	-	-
GPT-RE_FT	PURE	<u>91.90</u> (25)	<u>72.14</u> (15)	<u>69.00</u> (30)	68.73 (25)
GPT-RE_FT*	PURE	<u>91.11</u> (15)	<u>70.38</u> (10)	-	-
+ Reasoning	PURE	<u>91.82</u> (15)	<u>70.97</u> (10)	-	-
<i>Fine-tuned RE Baselines</i>					
[32]		91.90	-	-	-
[180]		-	♣76.80	-	-
PURE [216]		89.90	69.72	68.45	70.09

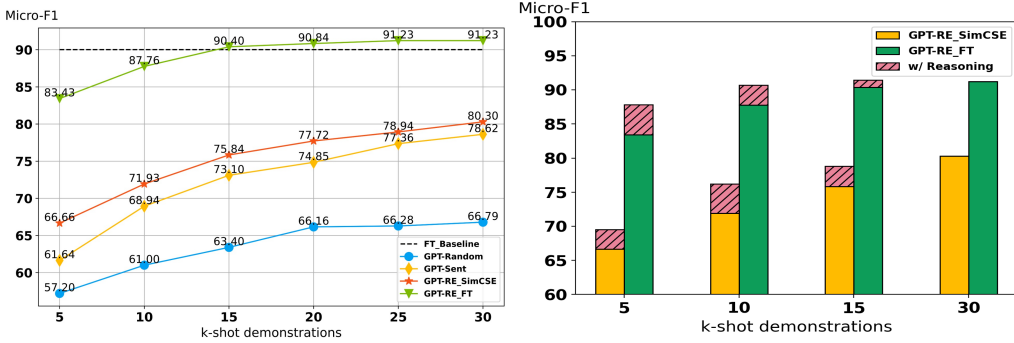
Table 2.2: **Main Results on four RE datasets.** All results are given by Micro-F1. * denotes the same k -shot for the comparison with + Reasoning. Due to the costly GPT-3 expense, we conducted Reasoning experiments on the two relatively smaller datasets Semeval and TACRED. ♣ denotes that this performance is not comparable as it evaluates on the entire test set. The underline denotes the results outperforming the fine-tuning baseline PURE.

ACE05 contains the entity, relation, and event annotations collected from domains including newswire, broadcast, discussion forums, etc.

2.4.2 Baseline Methods

GPT-3 baselines For GPT-3 baselines and our methods, we select “text-davinci-003” with maximal 4,097 input tokens and use the identical prompt construction (Sec. 2.3.3) via OpenAI API. We implement two categories of GPT-3 baselines:

- (1) **GPT-Random** Instead of randomly selecting few-shot demonstrations from the training data for each test input, we add extra constraints to make the label distribution of selected demonstrations more uniform. Our preliminary experiments suggest that this is a stronger baseline than the vanilla random.
- (2) **GPT-Sent** Previous work attempts various sentence embedding in retrieval.



(a) The comparison on retrieval models (b) Reasoning with fewer demonstrations.

Figure 2.6: **Ablation study on the retrieval and reasoning components on Semeval.** We sampled a subset from the test data with 300 examples. We show the ‘w/o reasoning’ results with $k = 30$ for comparison.

In this work, our implementation adopted SimCSE [43], which has been demonstrated to be the state-of-the-art method for sentence similarity tasks.

Fine-tuned RE Models In our experiment, we choose PURE [216], an entity marker-based fine-tuned model mentioned in Sec. 2.3.4 to obtain the representations for retrieval. Meanwhile, PURE performs as a directly comparable baseline. We also compare with corresponding SOTA fine-tuned baselines on Semeval [32] (reformulate RE as the question answering task) and TACRED [180] (extra pre-training to capture RE structure) datasets.

All implementation details are in Appendix A.1.

2.5 Experimental Results

2.5.1 Main Results

We compare our main experiment results with previous methods in Table 5.3. **GPT-RE_SimCSE** denotes our entity-prompted sentence embedding for retrieval and **GPT-RE_FT** denotes our fine-tuned relation representation for retrieval. From the table, we can observe that: (1) both *GPT-RE_SimCSE* and

Micro-F1

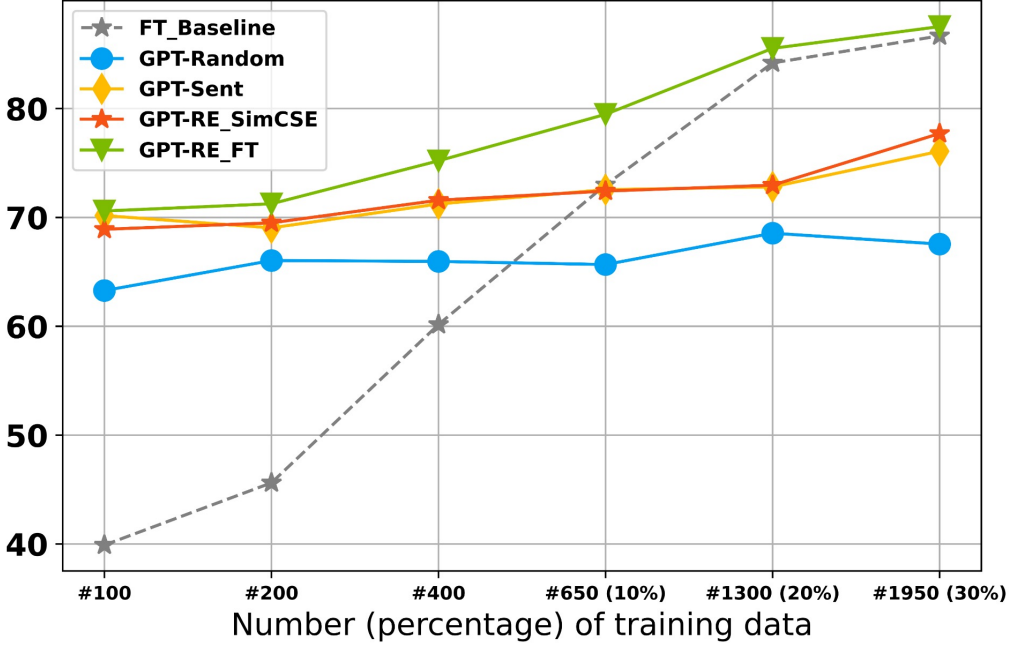


Figure 2.7: **Low-resource Scenario on Semeval.** We limit the percentage of training data for both fine-tuning and retrieval in GPT-RE.

GPT-RE_FT outperform the retrieval-based *GPT-Sent*, indicating that it is necessary to inject the task-specific information into sentence embedding for selecting proper demonstrations; (2) *GPT-RE_FT* succeeds to outperform the fine-tuning baseline PURE on three datasets by +2.00, +2.42, +0.55 Micro-F1. It suggests that GPT-3 has the potential to beat fine-tuning when the retriever has prior task knowledge. *GPT-RE_FT* eventually achieves SOTA results on Semeval and Sci-ERC. (3) reasoning module improves *GPT-RE_SimCSE* by around 2% Micro-F1, indicating that gold label-induced reasoning successfully enriches the knowledge of demonstrations. Meanwhile, the high-quality demonstrations obtained by *GPT-RE_FT* offset the effort of enriching reasoning into demonstrations, which shows relatively trivial improvements. Since reasoning aims at enriching demonstrations, this feature potentially works better with fewer demonstrations, as shown in Section 2.5.3.

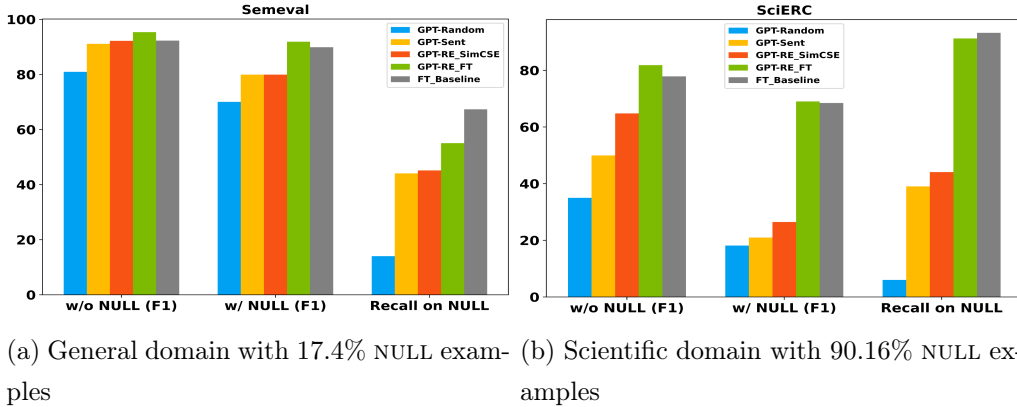


Figure 2.8: **Analysis on the effects of NULL examples.** w/o NULL refers to the classification setting that NULL examples are excluded from the train and test data. w/ NULL refers to the original extraction setting. We use the full test set for the evaluation.

2.5.2 Ablation Study on Task-aware Retrieval

We first implement the ablation experiments of the retrieval component with the setting of increasing k -shot demonstrations (Figure 2.6a). We find that: (1) compared to *GPT-Random*, all the retrieval-based models have higher F1 scores and large gradients of the performance curves. It means that GPT-3 can learn from high-quality demonstrations more effectively; (2) after adding entity information to the SimCSE retrieval, *GPT-RE_SimCSE* achieves better performance throughout all K shots, indicating that task-aware sentence embedding can capture the feature of RE and provide more proper demonstrations; (3) finally, the fine-tuned relation representation retriever *GPT-RE_FT* significantly outperforms all retrieval-based methods and beats the fine-tuning baseline when $k > 15$. Note that even with $k = 5$ demonstrations, *GPT-RE_FT* still works better than *GPT-RE_SimCSE* with $k = 30$ ($80.30 \rightarrow 83.43(+3.13)$), which indicates that the quality of demonstrations shows much more important than the number of demonstrations.

2.5.3 Ablation Study on Reasoning Enhancing

We then check the influence of our proposed reasoning-enhanced demonstration, as shown in Figure 2.6b. Due to the limited amount of input tokens of GPT-3, we have to set the $k \leq 15$ for the tokens of reasoning, leading to a trade-off between adding reasoning and adding more demonstrations. From the result, we find that: (1) with reasoning-enhanced demonstrations, GPT-3 always achieves better scores across all the k -shot settings of both *GPT-RE_SimCSE* and *GPT-RE_FT*, indicating that the reasoning induced from ground truth relation labels can effectively unlock the reasoning ability of GPT-3 and improve the ICL with a deeper understanding of demonstrations. Specifically, for *GPT-RE_FT*, the performance improvement becomes less significant when more demonstrations are provided, which is feasible as with more high-quality demonstrations available, GPT-3 can already learn the internal reasoning behind each demonstration; (2) since the reasoning enhancement works better with fewer demonstrations, we expect this method can be an effective solution to low-shot relation extraction [51, 44, 97], which aims at recognizing novel relations with very few or no examples, and we leave this for future work.

2.5.4 Low-resource Scenario

We conduct the experiment for observing the low-resource performance in the general domain Semeval task. As shown in Figure 2.7, we observe that: (1) all the GPT-3 based results work better than fine-tuning in when the training examples are less than # 650 (10%). It indicates that in the general domain RE, GPT-3 benefits from its abundant prior knowledge to understand the relations; (2) *GPT-RE_SimCSE* starts to show a substantial difference to *GPT-Sent* after the training size surpasses 30%. We believe fewer training candidates could limit the effects of retrieval; (3) *GPT-RE_FT* achieves an upper bound performance in all settings, even when the fine-tuned model shows poor performance with hundreds of training data (from #100 to #400). This emphasizes the impressive effectiveness of fine-tuned relation representations for capturing higher-quality demonstrations. The observation in the low-resource setting is very different from [74]. We assume

the difference could be caused by the domain and NULL proportion of the task.

2.6 Analysis

2.6.1 The Issue of “Overpredicting”

To analyze the influence of NULL class, we compare the effectiveness of each method for alleviating this issue on two datasets: general domain Semeval with 17.4% NULL examples and scientific domain SciERC with 90.16% NULL examples. As shown in Figure 2.8, (1) by comparing the performance on Semeval and SciERC, a larger percentage of NULL examples results in more significant performance drop showing the negative influence of overpredicting NULL examples; (2) by comparing *w/o* NULL and *w/* NULL, our *GPT-RE_FT* shows the most robustness to the influence of NULL examples, indicating that the RE fine-tuned representations in retrieval can release the overpredicting issue of GPT-3 by providing higher-quality demonstrations; (3) however, even with task-aware representations, all GPT-3 methods still underperform the fine-tuning baseline on NULL examples, this is due to the confusing definition of NULL, in many cases, there is a certain relation between entities in the context, but out of the distribution of pre-defined classes. In these cases, GPT-3 tends to overpredict as the relation information may be covered in its prior knowledge. We think this ability of GPT-3 can be useful in more open fields, such as open RE [7] which has no pre-defined relation classes.

2.6.2 Case Study of Demonstration Quality

We select one typical test example to better illustrate the amendment of our task-aware demonstration retrieval. As shown in Figure 5.7, given the NULL Example, we show the most similar demonstration in retrieval based on three methods. The *GPT-Sent* retrieved demonstration focuses on the semantic meaning of “CONTENT AND CONTAINER” which is shared in the test context, but not revealed in the target entity pair. This mismatch confirms the problem of lacking entity information in retrieval. Instead, *GPT-RE_SimCSE* retrieves a much more relevant demonstration that shows the same semantic relation between “catch” and “fish” but still faces a minor mismatch as the gold label is between “catch” and

“scuttle.” Finally, *GPT-RE_FT* demonstration shares a similar structure with the test input regarding the pair of entities, which is the key clue for predicting the relation between entities. This result shows a level-by-level enhancement with more entity information provided in retrieval. We also show some other case examples in Appendix A.2.

2.7 Limitations and Discussions

Despite the overall positive results, GPT-RE still faces two shortcomings: (1) the issue of overpredicting has been significantly alleviated but not completely solved, and the NULL recall still lags behind full-supervised baselines, especially on the datasets containing a large proportion of NULL examples such as ACE05 (“95.60%”); (2) Though the task-aware retriever optimizes the representations of PLMs such as SimCSE and BERT, it is widely considered that LLMs can generate more robust representations than small PLMs. Future work can replace representations generated by smaller PLMs with GPT-3 itself. However, due to the access limitation to the representations of GPT-3, we can nearly confirm this proposal up to now.

2.8 Summary of This Chapter

This work explores the potential of GPT-3 ICL on RE for bridging the performance gap to the fine-tuning baselines via two strategies: (1) task-aware demonstration retrieval emphasizes entity and relation information for improving the accuracy of searching demonstrations; (2) gold label-induced reasoning enriches the reasoning evidence of each demonstration. To the best of our knowledge, GPT-RE is the first GPT-3 ICL research that significantly outperforms the fine-tuning baseline on three datasets and achieves SOTA on Semeval and SciERC. We implement detailed studies to explore how GPT-3 overcomes the difficulties such as NULL example influence.

Relationship to the Dissertation: GPT-RE as the Foundational Scaffolding This chapter establishes ICL as the foundational archetype of inference-time

scaffolding. As the most direct manifestation of emergent abilities, ICL marked the dawn of the LLM era by enabling models to tackle novel tasks without additional training—a fundamental shift from the parameter-heavy paradigms of the past.

The significance of ICL, as explored through GPT-RE, remains a cornerstone of modern NLP for two primary reasons:

(1) *Persistent Necessity:* Even as LLMs continue to scale, ICL remains an essential scaffolding mechanism. Our findings highlight a persistent performance gap where even the most advanced models' zero-shot capabilities often fall short of the fine-tuned SOTA. ICL serves as the requisite bridge to elicit latent domain-specific reasoning.

(2) *Methodological Legacy in Reasoning Construction:* Beyond Information Extraction (IE), the gold label-induced reasoning framework introduced in this chapter has exerted a lasting influence on the field of "learning to reason". By utilizing gold labels to seed structured reasoning paths, this work provided an early and influential methodology for constructing the reasoning traces now widely used to train next-generation models with reinforcement learning methods.

Ultimately, this chapter demonstrates that the essence of LLMs lies in their controllability via context, establishing the first level of the scaffolding hierarchy that evolves toward autonomous agents in subsequent chapters.

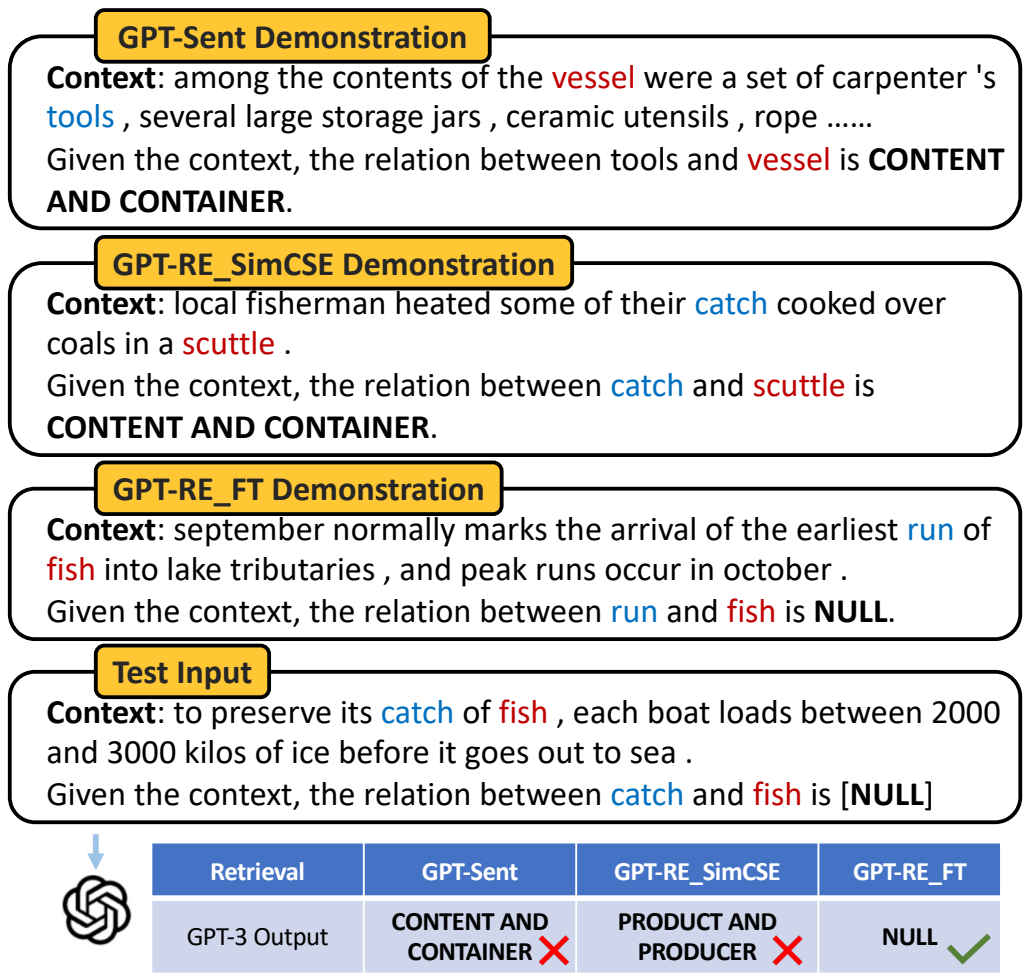


Figure 2.9: A case study of demonstration quality on Semeval. [NULL] is the gold label here.

Chapter 3

GPT-MM: Improving Multimodal In-context Learning with Task-specific Retrieval and Reasoning

3.1 Introduction

While GPT-RE provides a strong foundation for textual ICL, its principles have not yet been validated in multimodal contexts. The rapid progress of MLLMs offers a new opportunity to explore whether the same task-aware retrieval and label-induced reasoning mechanisms can facilitate MM-ICL.

Recent progress has shown that the ICL paradigm can be extended to vision–language and audio–language domains, enabling unified reasoning over heterogeneous inputs. However, despite this progress, the fundamental mechanisms that make ICL successful across modalities are still underexplored. Most existing multimodal prompting approaches focus on modality-specific designs rather than investigating whether a unified ICL formulation can operate consistently across text, vision, and audio.

To address this gap, we extend our textual framework to the multimodal setting

and introduce **GPT-MM**, a unified ICL framework designed to investigate the cross-modal generalizability of the two key components validated in GPT-RE: *task-aware demonstration retrieval* and *label-induced reasoning*. GPT-MM instantiates these components in both vision and audio tasks by leveraging modality-specific encoders while maintaining a consistent ICL prompting structure. Specifically, we implement GPT-MM on two representative multimodal tasks: Visual Question Answering (VQA) and Automatic Speech Recognition (ASR). Through this formulation, GPT-MM aims to examine whether retrieval and reasoning can serve as a unified paradigm for ICL across modalities.

3.2 A Unified Multimodal Extension: GPT-MM

The following sections introduce the multimodal task definitions, describe how retrieval and reasoning are adapted to each modality, and present experimental setups and results.

3.2.1 Task Definition

We consider two multimodal tasks: VQA and ASR.

For the VQA task, let $\mathcal{I}$ denote an image, and q be a natural language question. The goal is to generate the correct answer a from a set of possible answers $\mathbb{A}$:

$$p(a \in \mathbb{A} \mid \mathcal{I}, q)$$

For the ASR task, let $\mathcal{A}$ denote the audio waveform input. The goal is to transcribe the spoken content into text t :

$$p(t \in \text{Text} \mid \mathcal{A})$$

3.2.2 Overview

GPT-MM extends the GPT-RE framework to the multimodal ICL under the same ICL paradigm. We follow two steps: (1) perform task-aware multimodal retrieval to select relevant demonstrations; (2) inject modality-specific reasoning into the

demonstrations. The final prompt is constructed as a concatenation of instructions, demonstrations, and the test input, enabling the MM-ICL to predict the target label in a unified textual format.

3.2.3 Prompt Construction

Instructions $\mathcal{I}$ Each task includes a textual instruction describing the modality and the desired prediction format. For VQA, the instruction specifies how to answer questions based on image content. For ASR, the instruction defines the transcription objective from speech input.

ICL Demonstrations $\mathcal{D}$ We retrieve k relevant examples from the training set using task-aware retrieval (Sec. 3.2.4), and enrich each demonstration with modality-conditioned reasoning (Sec. 3.2.5). Each demonstration consists of an input-output-reasoning triplet: (x_i, y_i, r_i) .

Test Input x_{test} We convert and append the test image (or audio) and the paired question to the prompt. The MLLM is expected to generate the final output y_{test} (answer or transcript).

In summary, GPT-MM infers the output as:

$$p(y_{test} \mid \mathcal{I}, \mathcal{D}, x_{test})$$

3.2.4 Task-aware Multimodal Retrieval

In MM-ICL, selecting high-quality demonstrations is also essential for performance. Traditional retrieval methods rely on modality-specific global features, such as CLIP [138] for vision or Whisper [139] for speech. However, such representations are often task-agnostic and fail to capture the nuanced alignment between input media and task-specific answers.

To address this issue, we propose a unified, task-aware retrieval strategy using the Qwen2.5-Omni [195] model, which is capable of handling both image and audio inputs. Our key insight is to leverage the model’s own prediction behavior to derive meaningful, answer-aware representations.

Specifically, for each training or test example, we first format the input as a multimodal prompt by pairing the media input (image or audio) with its associated question. We then prompt Qwen2.5-Omni to generate a single-word answer. Rather than using external encoders, we extract the last-layer hidden state representation from the language decoder at the token position where the answer word is predicted. This hidden state inherently captures the joint reasoning over both the media and question, as well as the output space.

Formally, for a given multimodal input $x = (\mathcal{M}, q)$ with media $\mathcal{M}$ (image or audio) and question q , we denote the answer token as a , and the final-layer hidden state at that position as $h(x) \in \mathbb{R}^d$. We collect $h(x)$ for all training samples and compute cosine similarity with the test instance's $h(x_{\text{test}})$ to retrieve the top- k most similar examples:

$$\text{sim}(x_i, x_{\text{test}}) = \cos(h(x_i), h(x_{\text{test}}))$$

This retrieval method has several advantages: (1) It captures fine-grained alignment between the media, the question, and the generated answer; (2) It avoids relying on external modality encoders, ensuring consistency with the generation model; (3) Qwen 2.5-Omni is used for demonstration retrieval because it provides a unified embedding space for text, image, and audio inputs. This omni-modal property fits our unified GPT-MM framework, which aims to maintain a consistent retrieval representation across modalities, even though VQA and ASR are evaluated separately.

This embedding extraction process allows us to retrieve demonstrations that are semantically and functionally aligned with the test input, leading to improved ICL performance across modalities.

3.2.5 Label-induced Reasoning for Multimodal

We inject task-specific rationales into each demonstration to improve the model's understanding of input-output mappings using Qwen2.5-Omni.

- **For VQA:** Given a demonstration $(\mathcal{I}, q, a)$, we query the LLM with: *“Why does the image $\mathcal{I}$ lead to the answer ‘ a ’ for the question ‘ q ’?”* The generated

reasoning r (e.g., object recognition, commonsense alignment) is appended to the demo.

- **For ASR:** Given a demonstration $(\mathcal{A}, t)$, we query: *“What features in the audio help transcribe it as ‘t’?”* The model highlights relevant phonemes, speaker clarity, or financial terminology, forming reasoning r to support the transcription.

The final prompt contains:

$$\mathcal{I} + \underbrace{(x_1, y_1, r_1) \dots (x_k, y_k, r_k)}_{\text{Multimodal Demonstrations}} + x_{test}$$

This reasoning-enhanced prompting improves the LLM ’s ability to generalize across multimodal variations and ambiguity.

3.3 Multimodal Experiment Results

3.3.1 Datasets

To evaluate the generalizability of our proposed GPT-MM framework, we conduct experiments on two multimodal tasks from distinct domains: VQA and ASR.

VQA OK-VQA [116] is a knowledge-intensive VQA dataset, where answering questions about images often requires external commonsense or factual knowledge beyond what is directly visible. Each example consists of an image-question-answer triplet. The dataset contains 9,793 training samples and 5,046 testing samples. Due to the cost in API calls, we use a random subset of 1,000 examples. We follow the standard protocol and use accuracy as the evaluation metric.

ASR Earnings-22 [150] is a long-form, real-world ASR dataset consisting of English earnings call recordings and their transcripts. It contains 142 audio recordings totaling over 100 hours of speech, with significant variation in speaker accents, terminology, and background noise. Same as in VQA, we randomly select 1,000 segments in our test and remained segments will be the pool for demonstration

selection. We use word error rate (WER) as the evaluation metric. Each audio segment is paired with a transcription prompt and cast into a single-turn ASR generation task under the ICL setting.

3.3.2 Baselines

GPT-Random For each test example, we randomly select k demonstrations from the training set without considering any modality- or task-specific similarity. This simple strategy serves as a lower bound for ICL performance and highlights the importance of informed retrieval.

GPT-feature For the VQA task, we concatenate CLIP image features to form the sentence representation. For ASR, we use the encoder of Whisper-large-v3 to extract utterance-level embeddings from audio inputs. Demonstration retrieval is based on cosine similarity in the corresponding embedding space. This method reflects a typical multimodal feature-level retrieval baseline that ignores the model’s generative behavior.

GPT-MM (Ours) Our method adopts a task-aware retrieval mechanism that derives representations from the generation model itself. For VQA, we use GPT-4o to generate the answer token given the image and question. We then extract the hidden state corresponding to the predicted answer token as the embedding for retrieval. For ASR, we prompt Whisper-large-v3 with the audio input and extract the hidden state corresponding to the generated transcript tokens. These answer-state representations reflect the model’s internal decision-making process and are used to retrieve demonstrations based on cosine similarity. This approach ensures alignment between the retrieval representation and the model’s actual output behavior.

All retrieval methods use the same model as the inference engine for their respective tasks (GPT-4o for VQA and Whisper-large-v3 for ASR), and share the same number of 3-shot demonstrations. This design enables a controlled comparison that isolates the contribution of retrieval quality to final performance.

3.3.3 Main Results

We report the main results of our multimodal in-context learning experiments in Table 3.1. The performance is evaluated on the OK-VQA dataset (accuracy) and the Earnings-22 dataset (word error rate, WER). All methods use the same backbone model for inference (GPT-4o for VQA, Whisper-large-v3 for ASR), and differ only in their retrieval strategy.

Method	OK-VQA (Accuracy) $\uparrow$	Earnings-22 (WER) $\downarrow$
Zero-shot	51.32	11.44
GPT-Random	51.29	11.39
GPT-feature	51.32	11.37
GPT-MM (Ours)	52.24	11.26
w/o reasoning	52.11	11.31

Table 3.1: Main results on VQA and ASR tasks under 3-shot in-context learning. GPT-MM achieves consistent improvements by using task-aware answer-state embeddings for retrieval. Note that due to the limited length of multimodal inputs, we restrict the number of demonstrations to be 3.

We observe that both baseline methods, *GPT-Random* and *GPT-feature*, fail to improve over the zero-shot performance in both tasks. This is attributed to the low relevance of retrieved demonstrations: Random retrieval is entirely unstructured, and GPT-feature, while more informed, uses embeddings that are detached from the model’s actual generation behavior that focus on the task-related representations.

In contrast, our proposed **GPT-MM** method consistently improves performance across both modalities. By extracting task-specific representations from the Qwen2.5-omni model, the retrieved demonstrations are highly aligned with the decision process of the target model. This alignment leads to more effective conditioning and better in-context generalization. After combining with the label-induced reasoning, the performance further improves to achieve the best results.

These results highlight the importance of task-aware retrieval and reasoning in multimodal ICL. Simply using semantically similar media content or shallow

embeddings is insufficient; the retrieval space must reflect the model’s own inductive biases. GPT-MM bridges this gap by rooting retrieval directly in the model’s output dynamics.

3.4 Summary of This Chapter

This work presents a unified in-context learning (ICL) framework that enhances both textual and multimodal understanding through two complementary mechanisms—task-aware demonstration retrieval and label-induced reasoning. We first validated these mechanisms on textual relation extraction (RE), showing that aligning demonstrations with entity–relation semantics and injecting label-grounded reasoning logic substantially narrows the gap between ICL and fully supervised models. Building on this foundation, we extended the same framework to the multimodal domain, developing GPT-MM, which applies identical retrieval and reasoning principles to visual and audio question answering tasks. Experiments across all modalities confirm that task-specific alignment in both retrieval and reasoning plays a key role in unlocking the potential of ICL under heterogeneous input settings.

Overall, our findings demonstrate that the core mechanisms enabling effective textual ICL generalize naturally across modalities, providing a unified and interpretable foundation for multimodal reasoning. Future work will explore cross-modal transfer of retrieval strategies and reinforcement-based demonstration selection, aiming to establish an end-to-end, self-adaptive ICL paradigm.

Relationship to the Dissertation: Multimodal Expansion and Structural Heuristics This chapter extends the concept of Inference-time Scaffolding from the unimodal textual domain to the complex landscape of Multimodal Large Language Models (MLLMs). As models evolve from text-only to multimodal architectures, In-Context Learning (ICL) naturally transitions into a more challenging paradigm, where the strategic selection and construction of multimodal demonstrations become paramount.

The primary contribution of this chapter, GPT-MM, offers a dual impact on

the overarching scaffolding framework: (1) A Unified Framework for Modality Integration: By introducing special tokens for demonstration selection, we provide a structured and scalable framework that manages the high-dimensional complexity of interleaved image-text data. This serves as an architectural heuristic for how heterogeneous signals can be organized into a coherent inference-time prompt;

(2) Legacy in Data-Centric Reasoning: Beyond inference-time utility, our Logic-Induced Reasoning mechanism has established a long-term influence on the field. It has become a foundational method for generating reasoning traces, which are now essential for constructing high-quality multimodal instruction-tuning datasets.

In summary, this chapter demonstrates that as the complexity of the "context" grows in multimodal scenarios, the necessity for structured scaffolding becomes even more acute. It bridges the gap between basic ICL and the highly specialized, domain-specific systems explored in the subsequent chapters.

Chapter 4

Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain

4.1 Introduction

Recent large language models like GPT-4 bring remarkable improvements in various general domain NLP tasks [17, 126, 174, 30, 143, 58]. However, in specific domains such as the Chinese legal domain, the performance of such general LLMs still lags considerably behind. We show a real example of Chinese LegalQA [22] on the left of Figure 4.1, which requires the model to generate the corresponding legal provision (i.e., the law name and the clause index) and the rationale for the judgment, given a brief case description as the query.

We initialize the research with a preliminary examination of utilizing GPT-4 to address the Chinese LegalQA task, which involves responding with a law clause relevant to a given query case. Figure 4.1 reveals the extremely low performance

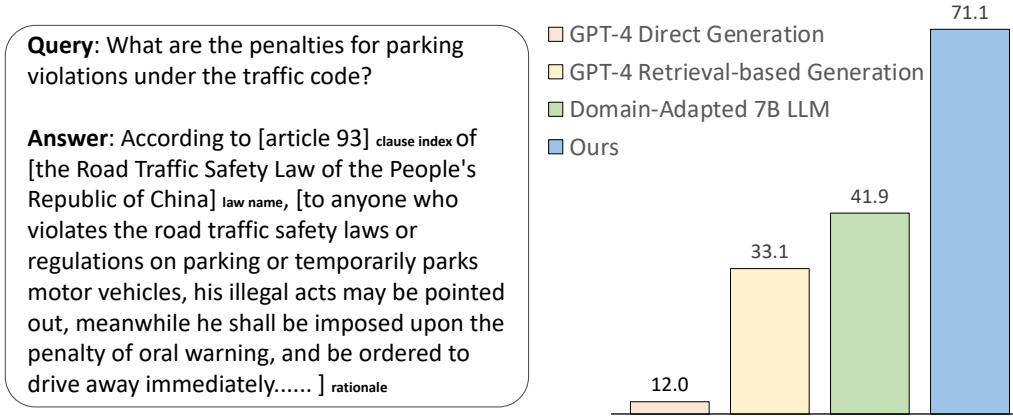


Figure 4.1: **Left:** A real translated example of Chinese LegalQA. The square brackets and subscripts are offered for the purpose of clear demonstration, not actually exist in the ground-truth answer or generation. **Right:** Models' F1 scores on the LegalQA dataset.

(F1 12.0) of directly prompting the query case to ask GPT-4 to generate the corresponding law clause. Though the generated answers are grammatically fluent, they often consist of non-logical content, factual mistakes, and fail to refer to the correct legal provision (also known as "hallucination"). For example, in the left answer of Figure 4.2, the direct generation of GPT-4 seems to retell the case description but fails to point out the corresponding clause. A potential reason is the insufficient Chinese legal domain text during pretraining, leading to a lack of domain knowledge acquisition, and consequently generating hallucinatory content.

For the LLMs with the scale of GPT-4, it's generally not feasible for researchers to do learning-based adaptation. The enormous model size could make the cost of continual learning extremely high, and meanwhile, the access functions are often limited by APIs. Therefore, recent work [94, 205, 163, 108] introduces retrieval-based methods that first use the given query to retrieve relevant evidence candidates from the external domain-specific knowledge base or the internet and then concatenate the query and the evidence candidates into the prompt. GPT-4 could implicitly validate the relevance between the query and the evidence, as well as

the correctness of the evidence, before producing a generation.

Our replicated retrieval-based method improves the LegalQA F1 from the 12.0 points of direct generation to 33.1 as shown in Figure 4.1. It indicates that even though GPT-4 may not generate domain content, it possesses sufficient evidence-assessing capacity to select the correct evidence from candidates. Nevertheless, the retrieval module is limited by the capability of representation mapping from query to evidence and is also influenced by the domain issue, leading to a decline in search quality. GPT-4 still produces hallucinations in responses as the middle answer in Figure 4.2.

On the other hand, with the rapid development of open LLMs led by LLaMA [175], it becomes affordable to continually train an open LLM tailored to your demands on sufficient in-domain texts, resulting in a domain-adapted LLM. We therefore conduct the second examining of continually training Baichuan 7B [5], a Chinese foundation model, on over 50B token Chinese legal data. Its performance (the green bar in Figure 4.1) even surpasses the retrieval-based GPT-4 generation on Chinese LegalQA. Hallucinations caused by the lack of domain knowledge are largely reduced but not completely solved. As shown in the right answer in Figure 4.2, adapted LLM generates generally correct responses but still makes errors in certain words. Although the law name is correct and the rationale part is reasonable, the clause index is a hallucinatory generation which raises the difficulty in anchoring target clauses. We argue that these fails are accordingly caused by the limited capability of a 7B size to memorize the knowledge accurately.

Building upon the observation of the evidence-assessing capability of GPT-4 and the high-quality domain content generated by the domain-adapted 7B model, this paper proposes a novel approach to reformulate GPT-4’s domain content generation to an **adapt-retrieve-revise** process: (1) the **domain-adapted** model generates a draft answer given a query; (2) the **retrieval** module uses the draft answer as input for searching external evidence candidates because the answer is usually more informative and semantically similar to the evidence compared to the query as long as the answer quality is acceptable; (3) GPT-4 assesses retrieved evidence and **revises** the draft answer to generate the final answer.

The rest sections of the paper anchor the Chinese legal domain and compre-

Query: What are the penalties for parking violations under the traffic code?

Answer:

GPT-4's Direct Generation:

In China, the penalties for parking violations may vary depending on the specific area and circumstances. Typically, parking penalties may include fines, demerit points and, in severe cases...

GPT-4's Retrieval-based Generation:

According to [article 37] of [the Road Traffic Safety Law of the People's Republic of China], [where a special driveway is delimited on a road, only prescribed...]

Domain-adapted LLM's Generation:

According to [article 90] of [the Road Traffic Safety Law of the People's Republic of China], [to anyone who violates the road traffic safety laws or regulations on parking or parks...]

Figure 4.2: Examples of hallucinations of various models. Red denotes the content containing hallucinations. The ground-truth answer refers to the left case in Figure 4.1.

hensively validate the effectiveness of our proposal. In Section 4.2, we explain each component of our adapt-retrieve-revise method and elaborate on the implementation details. In Section 4.3 and 4.5, we conduct the experiments and the result analysis on four Chinese legal domain tasks. The experimental results show substantial improvements against the direct generation and the retrieval-based generation baselines. In the final Section 4.7, we elicit the conclusion and future work. To the best of our knowledge, this is the first study to examine the zero-shot performance of LLMs on four Chinese legal benchmarks.

4.2 Methodology

Our adapt-retrieve-revise method consists of three steps. In the first step (Section 4.2.1), we continually train a Chinese pre-trained LLM on the Chinese legal domain corpora to derive a domain-adapted legal LLM and given the query, the legal LLM will generate the draft answer. In the second step (Section 4.2.2), we use a sentence embedding model to produce embeddings for both the draft answer and each paragraph in the corresponding knowledge base, then evidence retrieval will be computed by the similarities between the answer embedding and the paragraph embeddings. In the third step (Section 4.2.3), we concatenate the query, the draft answer, and the retrieved evidence in the prompt for GPT-4 to revise

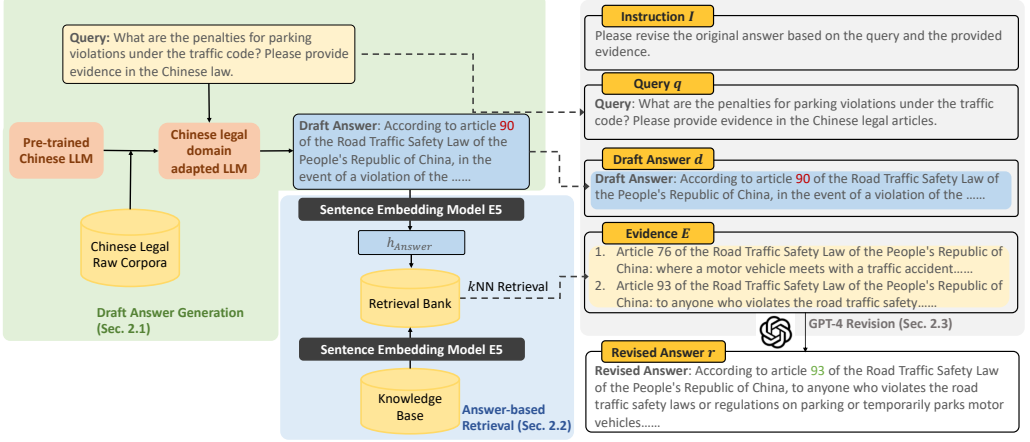


Figure 4.3: **Overview of our proposed method.** The example and prompt are translated from Chinese to English for the demonstration purpose.

and produce the final response. Figure 6.1 shows the overview of our method. In the following sections, we will introduce details of each step.

4.2.1 Draft Answer Generation by the Domain-adapted LLM

This step could be actually flexible to save the effort of training a LLM yourself. For instance, an off-the-shelf Chinese legal domain LLM can be used for generating draft answer. However, most of such LLMs have been fine-tuned on various in-domain evaluation tasks. The potential data leakage could lead to unreliable evaluation, especially in a zero-shot setting. Therefore, in this paper, we adapt an open 7B LLM to the Chinese legal domain by continual learning on the domain data. We collect the training data from the following two open Chinese legal sources:

- **Chinese Law Clauses** (<https://flk.npc.gov.cn/>) form the foundation of the judicial system, containing a wealth of legal terms, provisions, and judicial practices. They are essential for the model to understand and generate relevant content.
- **Chinese Judgments Online** (<https://wenshu.court.gov.cn/>) is the largest online publication platform for legal documents in China. The plat-

form contains judicial documents from courts at all levels, covering various legal fields such as civil, criminal, administrative, and enforcement. Such documents contain knowledge for LLMs to understand the usage of laws in various scenarios.

During the inference, given an input query, we will first prompt the trained 7B legal LLM to generate the draft answer, which will be used in the next step. For the prompt, we add the instruction "Please provide evidence in the Chinese law" at the end of the query to enforce the model to generate related law clauses, as in Figure 6.1.

4.2.2 Answer-based Evidence Retrieval

Since the draft answer of the 7B legal LLM is usually more informative and semantically similar to the evidence than the query. We further use the generated evidence to retrieve ground-truth evidence from the target knowledge base for the purpose of revision since it contains much more information than the query, even though the hallucinations can not be totally reduced. We implement this method with two subsequent steps: knowledge bank construction and retrieval.

Knowledge Bank Construction For the i -th paragraph p_i , we construct the *key-value* pair $(\mathbf{p}_i, p_i)$ where the *key* $\mathbf{p}_i$ is the representation obtained from the sentence embedding model and the *value* p_i denotes the paragraph. The memory $(\mathcal{K}, \mathcal{V}) = \{(\mathbf{p}_i, p_i) | p_i \in \mathcal{KB}\}$ is thus the set of all *key-value* pairs constructed from all the paragraphs in the external knowledge base $\mathcal{KB}$.

Retrieval Given the generated draft answer d , the sentence embedding model outputs its representation $\mathbf{h}_{\text{Answer}}$. We then query the constructed knowledge bank with $\mathbf{h}_{\text{Answer}}$ to retrieve its k nearest neighbors E according to a distance function by L^2 distance.

4.2.3 GPT-4 Revision

To effectively combine the high-quality draft answers generated by the 7B domain adapted model with GPT-4’s powerful evidence-assessing capability, we propose

the following process. As shown in Figure 6.1, the whole prompt consists of the following components: (1) the instruction I to require GPT-4 to revise the draft answer given the query and the evidence candidates; (2) the query q itself; (3) the draft answer d for GPT-4 to revise; (4) and the retrieved evidence candidates E to provide related Chinese legal knowledge for GPT-4. Then, the final revised answer r will be outputted by $GPT4(I, q, d, E)$.

4.3 Experiment Settings

We conducted a series of experiments to compare our adapt-retrieve-revise method to the baselines of direct generation and retrieval-based generation on various Chinese legal benchmarks. We show the model details and the task settings in this section.

4.3.1 Model Settings

Details of training 7B legal LLM: We utilize the general domain Baichuan 7B model¹ for continual learning Chinese legal corpora. In total, we trained 50B tokens of Chinese Law Clauses and Chinese Judgments Online corpora with the input length limit of 16K and the batch size of 256 on 32 A100 GPUs, and the time-consuming is 167 hours. After continual learning, we subsequently supervised fine-tuning our model on 70K instruction examples, including 52K GPT-4 self-instruct Chinese data [133] and 18K legal instructions (See Appendix B.0.1) for the alignment.

Retriever setting: We utilize Multilingual-E5-large [182], a Roberta-based [100] sentence embedding model that achieves robust performance on various tasks. We also compare with other retrieval modules in Appendix B.0.2.

GPT-4 setting: For the utilization of GPT-4, we select “gpt-4-0613” with maximal 8K input tokens and use the original Chinese prompt as shown in Sec 4.2.3 and Figure 6.1 via OpenAI API.

¹<https://huggingface.co/baichuan-inc/Baichuan-7B>

4.3.2 Evaluation of Chinese Legal Tasks

We evaluated our Adapt-Retrieve-Revise method on a diversity of tasks with different knowledge base for retrieval in the zero-shot setting:

- **Law Clause Recommendation (LCR) and Criminal Prediction (CP)** [192] are two tasks using the legal report as the input, and let the model generate the most related law clause and predict the criminal type based on the law clause. For these two tasks, we use the **Chinese law clauses** as the domain knowledge base for retrieval.
- **LegalQA** is a filtered set of EUQALS [22] that, given an input query, the model should generate an answer based on the most related legal clause. The filtering is based on the quality of the questions and we will release the filtered set. We also use the **Chinese law clauses** as the domain knowledge base for retrieval.
- **JEC-QA** [215] is the official test for getting a lawyer’s certificate in China. We chose the single-choice selection questions in our evaluations with the **Legal Textbooks** (<https://github.com/thunlp/jec-qa>) as the knowledge base for retrieval.
- **Similar Case Retrieval** [109] is the task, given a query legal scenario as the input, we aim at selecting similar **Legal Judgement Documents** from the 100 candidates. We conducted this experiment to assess the reliability of our proposed retrieval method in Section 4.5.1.

Due to the cost of GPT-4 API and the human evaluation, we randomly sampled a subset of 250 test examples for each task of LCR, CP, LegalQA, and JEC-QA. Please refer to Table 4.2 for the statistics.

4.3.3 Evaluation Metrics

Since generative models produce diverse formats in the output and the Chinese legal domain has its own features, checking the evaluation metrics in the experiments is crucial. For tasks LCR, CP, and LegalQA, our metric is the *Micro F1*

Generator	Retriever	Revisor	Chinese law clauses			Textbooks	Avg.
			LCR F1 (Rec.)	CP F1 (Rec.)	LegalQA F1 (Rec.)	JEC-QA Acc.	
Direct Generation							
GPT-4	-	-	61.7 (67.6)	70.7 (71.2)	12.0 (14.4)	36.2	45.1
LawGPT-7B [164]	-	-	19.3 (25.3)	33.3 (34.6)	10.3 (16.6)	27.4	22.6
ChatLaw-13B [33]	-	-	25.6 (27.6)	43.6 (49.6)	14.6 (18.3)	31.8	28.9
7B legal LLM	-	-	83.0 (88.4)	82.9 (84.0)	41.9 (48.8)	39.8	61.9
Retrieval-based Generation							
GPT-4	Query-based	-	72.0 (74.4)	74.0 (75.2)	33.1 (36.0)	41.6	55.2
7B legal LLM	Query-based	-	77.3 (87.6)	81.3 (82.4)	47.4 (50.2)	40.8	61.7
Adapt-Retrieve-Revise							
7B legal LLM	Answer-based	7B legal LLM	84.1 (88.4)	82.7 (83.6)	47.3 (49.0)	40.2	63.6
7B legal LLM (ours)	Answer-based	GPT-4	90.6 (96.4)	86.9 (87.8)	71.1 (72.4)	66.2	78.7

Table 4.1: **Zero-shot Results on four Chinese legal datasets.** “Rec.” denotes recall, “Acc.” denotes accuracy, and “Avg.” is computed by the F1 and accuracy scores of all four tasks

Datasets	CP	LCR	LegalQA	JEC-QA
# test	965, 219	965, 219	1,000	13,341

Table 4.2: **Statistics of datasets.**

and the *Recall* of whether the title of the ground-truth law clause is included in the generated answer. This is because, in real-world applications, with the correct title, the contents of the law clause can be easily revised by the rule-based system, indicating that the title is more important than the content.

For the JEC-QA task, we use accuracy as the metric, but controlling the output into an identical format for automatic evaluation is difficult, especially for the 7B LLM that has not been fine-tuned on the JEC-QA task. We select human evaluation to ensure its accuracy by manually comparing the model output and the gold answer provided by the dataset.

For the Similar Case Retrieval task, we chose the widely used *precision@k* and *MAP* as the evaluation metrics.

4.4 Main Experimental Results

We provide the main results as in Table 4.1. As we claimed before, this is the first work to targeting the zero-shot LLM performance on Chinese legal domain tasks. Generally, we compare our adapt-retrieve-revise proposal with baselines of direct generations and retrieval-based generations using the query, showing that our method outperforms all baselines by substantial margins. Our main results also provide some ablation results.

We first observe the effectiveness of domain adaption. Our 7B legal LLM significantly beats GPT-4, and even compared with the retrieval-based generation of GPT-4, the 7B legal LLM still outperforms on three tasks and has competitive results on the JEC-QA task, indicating that our continual learning on Chinese legal raw corpora shows a fast and effective domain adaptation on various legal tasks. However, related work (LawGPT and ChatLaw) fails to benefit largely from the continual training due to the much less data used in training, and their base models are multilingual Llama.

Then, considering the results of GPT-4 and the GPT-4 retrieval-based generation, we find that after providing evidence of related legal knowledge, GPT-4 can improve its responses significantly (+10.1 points). This indicates that the retrieval-based method is a proper way to reduce hallucinations caused by the lack of domain knowledge, and owing to the robust evidence-assessing capacity, GPT-4 can adapt to the Chinese legal domain well with convincing evidence available.

In our final experiment, using the draft answers generated by the 7B legal model for retrieval and revision, the performance significantly exceeded two query-based retrieval baselines by large margins of +17.0 and +23.5 points. It’s worth noting that the improvement here comes from both the enhanced answer-based retrieval quality and the revision setup. In the subsequent ablation study, we will further examine the quantified improvement of the retrieval quality through an additional retrieving task.

An interesting observation is that, by comparing the direct generation of the 7B legal model and the adapt-retrieve-revise method with the revision model as

Retrieval Recall

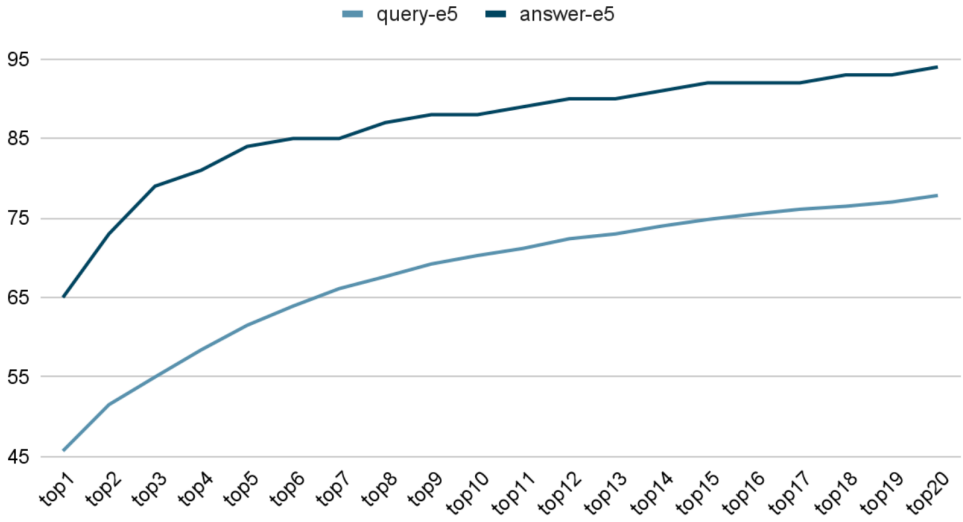


Figure 4.4: Comparison of retrieval recalls on the LegalQA dataset.

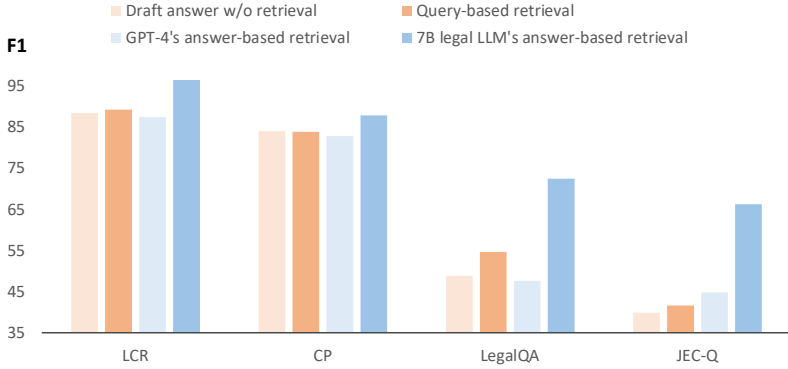


Figure 4.5: We compare performances of the draft answer of 7B legal LLM and our proposed adapt-retrieve-revise model using different contents in retrieval.

the legal 7B model, we find that with retrieved evidence, the revised answers seem to be no obvious difference from the direct generation. This indicates that the 7B legal LLM shows almost zero evidence-assessing capacity.

4.5 Further Analysis

4.5.1 Analysis of Retrieval Methods

The previous section has demonstrated the step-wise effectiveness of our adapt-retrieve-revise proposal. Nevertheless, the retriever, as a key component that directly affects the quality of evidence, how its variations would impact the final performance is a crucial research question to investigate.

Retrieving A Query or Retrieving an Answer?

We believe the answer-based approach is more effective due to two reasons. (1) The query-based retrieval requires a query-to-evidence representation mapping. The answers are usually more semantically similar to the evidence, which avoids the mapping process. (2) A query is often very brief, while an answer containing the legal provision and rationale is more informative. In this sub-section, we analyze the retrieval component, including the apple-to-apple comparisons between the query-based and answer-based performance and the influence of answer quality for the answer-based retrieval.

We ordered the top-similar law clauses in each retrieval and evaluated the recall in top- k , indicating whether the ground-truth law clause appears in the top- k retrieved law clauses. As shown in Figure 4.4, the top-1 retrieved law clause based on the answer competes with the top-5 law clauses based on the query, and the answer-based retrieval beats the query-based retrieval with a large margin for all k . This confirms our first reason that the draft answer contains much more information than the query for retrieval, indicating that LLMs can be intrinsic retrievers.

We further compare the query-based and answer-based retrieval on a public **Similar Case Retrieval** task. This task aims to select similar legal judgments based on the query from the candidates with a query the case brief given. As shown in Table 4.3, we compare two setups: 1) using the original query to retrieve, 2) using the legal 7B LLM to complete a whole legal judgment document given the brief query, and then retrieving. We follow the original task repository for the

Setup	Precision@5	Precision@10	MAP
Query-based	42.1	42.0	47.8
Answer-based	45.2 (+3.1)	42.1 (+0.1)	49.5 (+1.7)

Table 4.3: Results of two retrieval setups on the Similar Case Retrieval dataset.

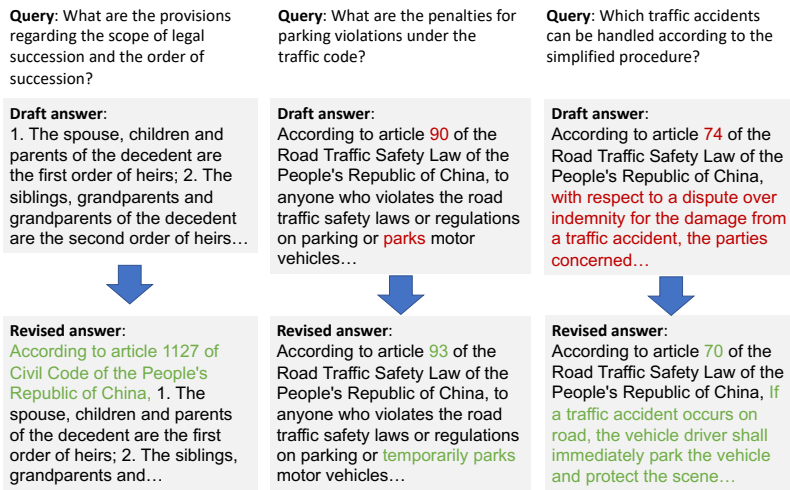


Figure 4.6: Case studies on the GPT-4 revision. The examples are translated from Chinese to English for the demonstration purpose.

other settings². The results show that on each metric, the answer-based retrieval works better, indicating that using the generated answer provides a more robust retrieval.

Does the Quality of Answer Matter for Answer-Based Retrieval?

It's an intuitive thought that the quality of answers will significantly impact the outcome of answer-based retrieval. Therefore, we compare the retrieval using the answers of GPT-4 and the 7B legal LLM. We change the contents in retrieval for our proposed adapt-retrieve-revise method. As shown in Figure 4.5, by comparing

²<https://github.com/myx666/LeCaRD>

query-based and GPT-4’s answer-based retrievals, we find that the answer-based retrieval fails on three datasets (LCR, CP, LegalQA), indicating that the lack of domain knowledge in the GPT-4 responses leads to a more noisy retrieval, which even hurts the performance of the draft answer (LCR, CP, LegalQA). Meanwhile, after domain adapting, our 7B legal LLM provides robust answers in retrieval and leads to the best performances, indicating that the learned Chinese legal domain knowledge improves our answer-based retrieval.

4.5.2 Case Analysis of the Improvements after the GPT-4 Revision

We conclude the improvements by GPT-4 in three aspects as shown in Figure 4.6:

- **Adding law clauses for reference:** Sometimes, the 7B legal LLM only provides a fluent response without following the input instructions to provide the key information of the referred law name and clause index. In this case, the faithfulness of the answer remains unchecked for the users. However, after the revision, each answer is equipped with the referred law clause, which makes it easier to check the accuracy of the responses.
- **Revising hallucinations in the evidence:** even the domain-adapted LLM can provide evidence from its learned legal knowledge; the hallucination remains to some degree, such as the wrong clause index, even the law name and rationale are roughly correct. Since the rationale content is accurate, the answer-based retrieval will search for the correct evidence, and the revision by GPT-4 will solve the hallucination to produce a more robust response.
- **Choosing correct evidence:** In a more significant scenario, even though the 7B legal model’s answers might contain partial hallucinatory content, the retrieval component can still possibly identify correct evidence through the partially correct descriptions in the rationale generation. During the revision stage, GPT-4 could assess the correct evidence, leading to the generation of correct answers.

Retrieval Recall

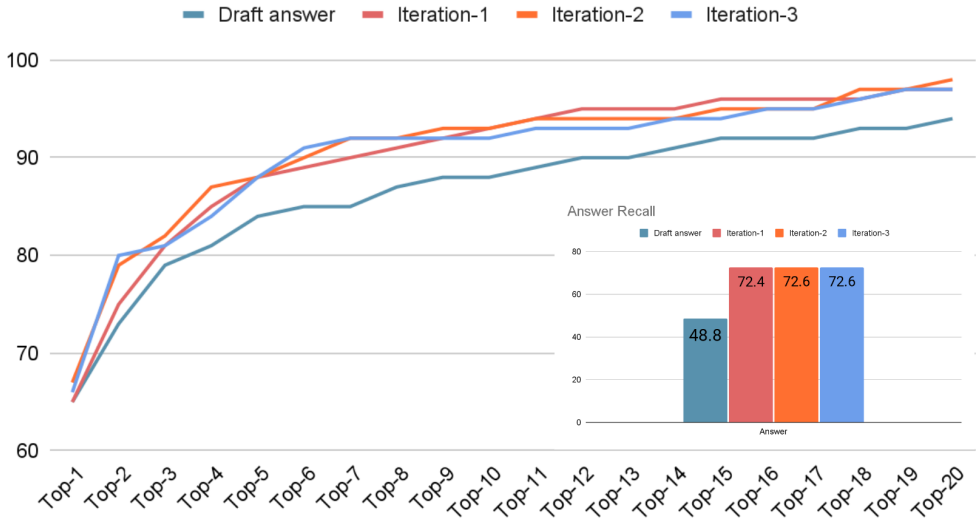


Figure 4.7: Comparison of iterations on LegalQA dataset.

4.5.3 Does the iteration make the generation better?

Since our method provides more accurate responses than the original response from the domain LLM, one question is whether this procedure can be iterated to improve the responses. We can use the revised response to retrieve related evidence and further improve the response. To verify this probability, we iteratively test on the LegalQA dataset. As the result is shown in Figure 4.7, during the iteration, the retrieval recall does not show consistent improvements compared with the first revision, leading to the performance nearly unchanged.

4.6 Related Work

4.6.1 Chinese Legal Domain Tasks

The rapid advancements in LLMs have significantly impacted various domains, including the legal industry. This gives rise to the occurrence of legal datasets, such as the Challenge of AI in Law (CAILf)³, LeCaRD [109], JEC-QA [215] and

³<http://cail.cipsc.org.cn/index.html>

EQUALS [22]. These datasets cover document classification, summarization, question answering, information extraction, similar document retrieval, and other popular NLP tasks in the Chinese legal domain. To the best of our knowledge, this paper is the first work to exam the zero-shot performances on these legal datasets.

4.6.2 Chinese Legal LLMs

As for Chinese legal LLMs, recent work utilizes a paradigm of continual learning in the legal domain, and a substantial number of instruction fine-tuning datasets were constructed to augment the proficiency in rendering legal advice. Particularly, the series of LaWGPT [164] has been developed by leveraging foundational models such as Chinese-LLaMA-7B [34], ChatGLM [39], and Chinese-alpaca-plus-7B [34]. Lawyer LLaMa [65] base on the more advanced Chinese-LLaMa-13B [34], On the other hand, LexiLaw [50], built on the foundation of ChatGLM-6B [39], underwent training through the application of three distinct methods, namely LoRA [61], P-tuning [99], and fine-tuning. Furthermore, Chatlaw [33] received training based on both Ziya-LLaMA-13B-v1 [69] and Anima-33B [106]. DISC-LawLLM [206] adopted legal syllogism prompting strategies to construct supervised fine-tuning datasets and fine-tune LLMs with legal reasoning capability. A primary reason hindering us from utilizing such existing models is that they have often been trained on those publicly legal tasks already. Therefore the zero-shot capabilities can not be truly reflected. We thus continue training the general Baichuan 7B on legal data by ourselves.

4.6.3 Retrieval-augmented Inference

In scenarios where language models are confronted with tasks necessitating an infusion of external knowledge, a retriever mechanism can be used to provide evidence. The Retrieval-Augmented Generation (RAG) [94] system incorporates a BERT-based [36] Document Retrieval Process (DRP) and utilizes BART [93] for answer generation. Analogously, the EMDR2 [205] employs the expectation-maximization algorithm to account for multiple retrieved documents. The Atlas [71] builds upon the EMDR2 framework, and by synergistically training the retriever and reader components, it demonstrates few-shot learning capabilities commensurate with the

540B PaLM [30]. RETRO [12] benefits from retrieval mechanisms on expansive corpora during its pre-training phase and exhibits performance in close alignment with those of GPT-3 [18].

4.7 Summary of This Chapter

In this paper, we reformulate the zero-shot domain content generation of large language models as an adapt-retrieve-revise procedure. This approach combines the merits of efficient continual training of a smaller 7B LLM for domain adaptation, robustly retrieving the supporting evidence from an external knowledge base, and effectively leveraging the evidence-assessing and revision capabilities of GPT-4. Our proposal substantially enhances the zero-shot performance on the Chinese legal tasks.

Relationship to the Dissertation: RAG as the Dynamic Link to an Evolving World While In-Context Learning (ICL) empowers LLMs to perform novel tasks with remarkable proficiency without parameter updates, it encounters a fundamental limitation: the temporal obsolescence of parametric knowledge. In specialized and rapidly evolving domains such as law and medicine, a model’s internal knowledge base inevitably becomes outdated. Given the computational intractability of frequent retraining, RAG emerges as a critical and indispensable inference-time scaffolding.

This chapter, through the Adapt-Retrieve-Revise framework, demonstrates that RAG is more than a simple data retrieval tool; it is a sophisticated mechanism that bridges static models with a dynamic world. By leveraging the LLM’s inherent context utilization and reasoning abilities, RAG enables a high-efficiency, low-cost connection to external, real-time information. This approach mirrors human professional intelligence: even the most proficient experts do not rely solely on memory but utilize reference manuals and external databases as a vital component of their cognitive process.

Consequently, this work establishes RAG as a pivotal layer in the scaffolding hierarchy, transitioning the dissertation’s narrative from static capability elicit-

tion to dynamic knowledge integration. This sets the stage for the final stage of our framework: the evolution into fully autonomous agentic systems that can self-reflect on the validity of retrieved information.

Chapter 5

SpeechIQ: Speech-Agentic Intelligence Quotient Across Cognitive Levels in Voice Understanding by Large Language Models

5.1 Introduction

The rapid rise of voice understanding Large Language Models ($\text{LLM}_{\text{Voice}}$) that can process voice input via the speech portal has ushered in a new paradigm of human-machine interaction [146], where $\text{LLM}_{\text{Voice}}$ process spoken instructions, infer semantic intent, and execute downstream tasks [72, 77, 218, 176, 119, 80, 64, 113]. $\text{LLM}_{\text{Voice}}$ bridge *speech* and *language intelligence* [166], enabling applications from voice assistants to interactive robots. A fundamental prerequisite for deploying robust $\text{LLM}_{\text{Voice}}$ lies in establishing reliable evaluation metrics for voice understanding. These metrics ensure that the system accurately comprehends voice inputs. Automatic speech recognition (ASR) is the central component of exist-

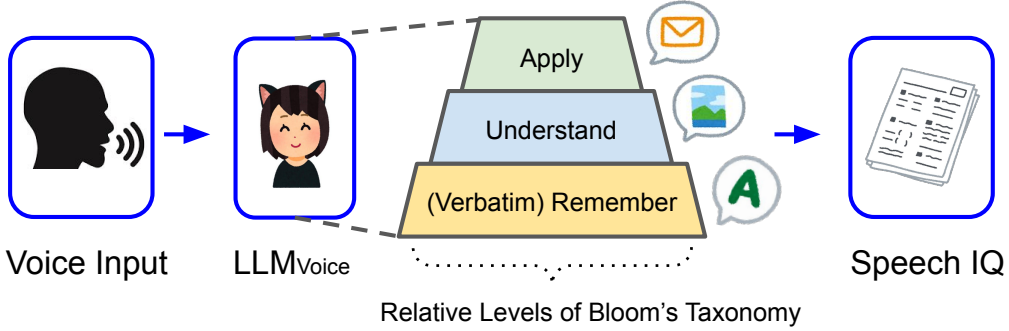


Figure 5.1: An overview of cognitive levels in voice understanding large language model systems related to the first three foundational hierarchies of Bloom’s Taxonomy. We design a corresponding examination to measure SpeechIQ as a detailed pipeline in Figure 5.2.

ing cascaded¹ LLM_{Voice} systems, transcribing voice into text for subsequent LLM processing. Consequently, the research community primarily assesses voice understanding using transcription error metrics [68, 84], with Word Error Rate (WER) standing as the de facto standard.

However, WER primarily measures the lexical recall of ASR, its transcribed quality does **not** fully capture the semantic understanding [52, 26, 169] and task completion [177] capabilities of LLMs via their speech portal. For instance, WER treats all lexical errors equally, ignoring how transcription inaccuracies propagate to higher-level comprehension (*i.e.*, semantic understanding) and executions (*i.e.*, downstream tasks).

Meanwhile, a growing trend in LLM_{Voice} research is the shift toward end-to-end multi-modal models [151, 141, 114, 63, 35, 27], which bypass explicit ASR transcription and directly map audio inputs to task-specific outputs (e.g., answers, actions). While this paradigm simplifies pipelines and reduces error cascades, its end-to-end voice service obviates the need for generating an intermediate transcript, rendering traditional WER-based evaluation obsolete [89, 171]. This creates a critical methodological gap: without a reliable and potentially unified metric,

¹We refer to two-pass based neural ASR-LM systems [59, 153, 66], which have been widely adopted in the current voice assistants, including Siri, Alexa, and Google Assistant.

comparing voice understanding across architectural choices, such as the modularity of “*cascaded neural systems*” and the efficiency of “*end-to-end multimodal models*,” remains insufficiently characterized.

The above two limitations mirror a fundamental challenge in assessing artificial intelligence: as Figure 5.1, human-like intelligence is inherently hierarchical. Drawing on Bloom’s Taxonomy [11, 9], a cognitive framework where skills ascend from basic remembering to advanced creating, we posit that $\text{LLM}_{\text{Voice}}$ should be evaluated through analogous levels. In this framework, WER represents the lowest level by measuring verbatim recall while neglecting higher-order abilities such as semantic understanding and task-solving. For instance, two transcripts with identical WER may differ drastically in meaning [168], leading to divergent $\text{LLM}_{\text{Voice}}$ responses or failed instructions.

To address this gap, we propose a hierarchical IQ test for $\text{LLM}_{\text{Voice}}$ that evaluates intelligence across three fundamental levels (refer to limitations) aligned with Bloom’s Cognitive Taxonomy: (1) Remember in verbatim: Lexical accuracy is quantified using WER; (2) Understand: Semantic consistency is assessed by comparing LLM responses to ASR outputs and ground-truth transcripts. LLMs are prompted to infer the background context (*e.g.*, domain, speaker intent) and generate a summary of the speech, then measure hidden states similarity between ASR-derived and ground truth-derived responses; (3) Apply: Task-solving capability is tested via multi-choice question-answering (QA) pairs constructed from ground-truth transcripts. $\text{LLM}_{\text{Voice}}$ answers questions based on speech inputs, with accuracy reflecting real-world utility. Then, we draw inspiration from Raven’s Progressive Matrices [145, 75] (*i.e.*, a human IQ test) that aggregates performance across various dimensions into a single score.

Beyond overall evaluation, our SpeechIQ framework yields two key insights:

- Cascade systems (ASR+LLM) outperform end-to-end models under similar scaling, revealing modality interference in joint speech-text training.
- Cross-model QA consistency exposes ground-truth errors. By isolating questions most $\text{LLM}_{\text{Voice}}$ cannot solve, we create an “unanswerable” set that helps detect hallucinations [41, 85] and refine benchmark annotations.

Our work advances $\text{LLM}_{\text{Voice}}$ evaluation by merging cognitive principles with practical metrics, revealing hidden challenges in multimodal inference (*e.g.*, hallucination inheritance), and offering tools to build more robust systems. Code and data will be publicly available.

5.2 Related Work

5.2.1 ASR Evaluation

ASR systems have long been evaluated using lexical similarity metrics such as WER, which quantifies the Levenshtein distance [124] between ASR outputs and reference transcripts. Metrics like character error rate [111] (CER), sentence error rate [76] (SER), and translation error rate (TER) extend this framework to the sentence or translation level but share WER ’ s core limitation: treating all errors equally, regardless of semantic impact. While effective for benchmarking transcription fidelity, these metrics do not fully capture how errors propagate to downstream tasks.

Recent work mitigates WER ’ s semantic insensitivity by incorporating sentence embeddings into evaluation. Sentence similarity-based metrics [81, 82] and BERTScore [210] leverage pre-trained language models to compute semantic correspondence between ASR hypotheses and ground-truth references. Others propose task-specific metrics, such as Medconcept WER [3] which arranges more weights on *medical entities* in WER computation or severity score [188] that involves sentiment similarity in the evaluation. However, these methods are either confined to specific tasks or require **curated labeled data**, limiting their generalizability to open-domain $\text{LLM}_{\text{Voice}}$.

Efforts to develop hybrid metrics that combine error rate and semantic similarity, such as H_eval [157], and Sema [156], further enrich the evaluation landscape. Concurrently, machine translation inspired metrics like BLEU Score [132] and COMET [147], have started to be adapted for ASR to assess fluency and pragmatic adequacy. However, these approaches still emphasize text-to-text alignment between ASR outputs and references, rather than from a view of influencing the responses and actions of downstream LLMs, motivating the proposed SpeechIQ as

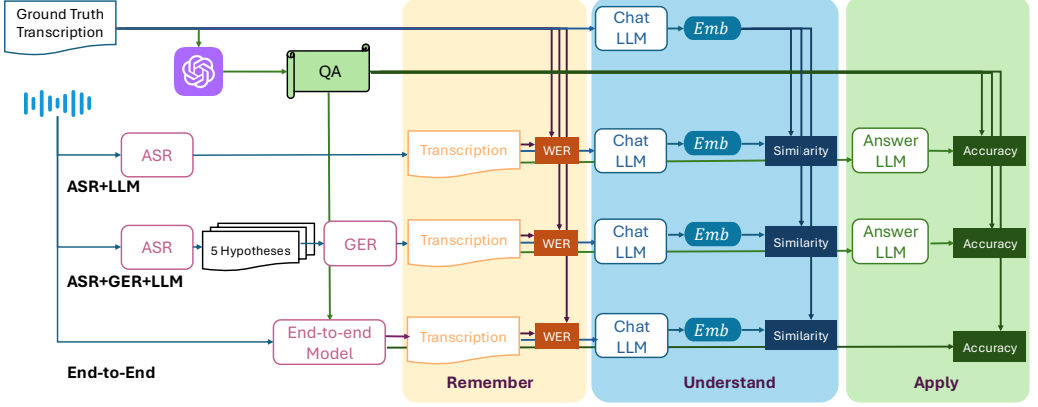


Figure 5.2: **Overview of Speech IQ (SIQ) test.** We compared cognition-inspired categories of $\text{LLM}_{\text{Voice}}$ represented in three rows, with the right columns denoting the SIQ sub-tests. The truth reference will be used in the sub-tests.

a multi-dimensional metric.

5.2.2 $\text{LLM}_{\text{Voice}}$ Understanding Systems

In general, speech-based LLMs encompass both the understanding of voice signals and the generation [4, 13, 202] of vocal or general audio outputs. In this work, we focus exclusively on **voice understanding tasks** with text outputs as one first step to examine voice intelligence [166] or its potential voice world model [49, 118] via LLM backbone, which we denote as $\text{LLM}_{\text{Voice}}$. $\text{LLM}_{\text{Voice}}$ understanding systems are designed to interact [161, 102] with users and execute task-oriented instructions via speech-based inputs. This form of “voice-to-text” architecture can be broadly categorized into three main types, all of which achieve state-of-the-art ASR quality on public leaderboards [139, 23, 135]:

(1) **Cascaded ASR + LLM** This approach follows a traditional pipeline-based structure, where an ASR model [186, 48, 139] transcribes the spoken input, and the 1-best transcription is passed to textual LLM as one operation system [191, 38] for direct response generation.

(2) Cascaded ASR Hypotheses + Generation Error Correction (GER)

+ **LLM** In this approach, the ASR model first generates multiple possible transcriptions (*i.e.*, hypotheses) for a given speech input, where “texts” are then treated as input features for a many-to-one GER post-editing module [199, 178, 60]. The GER module refines these hypotheses to determine the most accurate transcription, which is subsequently processed by the LLM.

(3) End-to-End Multi-Modal Models These models are inherently designed to support both speech and text modalities [151, 63, 107, 208]. Unlike cascaded approaches, they process audio inputs directly and generate textual outputs **without** requiring intermediate ASR transcriptions. This paradigm leverages a multimodal learning framework (*i.e.*, a speech embedding encoder [141] or tokenizer [14, 211]) to achieve end-to-end speech understanding and reasoning.

In summary, we carefully selected and examined aforementioned three ASR-LLM systems. System 1 serves as the most cost-effective solution [54] for voice interactions with LLMs. System 2 is designed to decode texts that offer richer information [21, 201] for textual applications. System 3 highlights a low-latency, multimodal embedding injection approach for large language models [129, 172]. Next, we introduce our proposed method, which corresponds to Bloom’s cognitive taxonomy.

5.3 SpeechIQ: Speech Intelligence Quotient

We will subsequently introduce each level of the test, from the remember level to both the understand level and apply level. At the end of this section, we will discuss the computation of the final SpeechIQ. An overview of our workflow is in Figure 5.2.

5.3.1 Remember: WER Metric

The remember level assesses the ability to reproduce spoken content accurately. WER serves as a natural metric since it measures the verbatim discrepancy between ASR output and the ground truth at a granular level, with the advantage

that it captures even the smallest differences, making it highly sensitive to transcription errors.

5.3.2 Understand: Similarity Metric

The understand level assesses whether the ASR-generated transcript effectively conveys the intended meaning of the original speech. This is particularly important when ASR output serves as input to the LLM, and a minor transcription difference could lead to significant semantic variations to LLMs. For instance, two transcriptions with identical WER scores can have vastly different meanings:

- Ground truth: "I feel pain in the lower back."
- ASR 1: "I feel **like** pain in the **_** back."
- ASR 2: "I feel **painting** in the **world** back."

Although both have the same WER = 29%, they convey different meanings and lead to drastically different LLM responses.

To quantify the impact of ASR errors on semantic comprehension, we measure the deviation in LLM responses caused by ASR transcription errors. Specifically, inspired by recent work [73, 101] showing that instructing LLMs to generate *one word* is a simple and straightforward approach to representing semantic meanings, we evaluate LLM-generated responses via two key questions: (1) *b*: The background scenario of the speech [] in *one word* is. This helps resolve ambiguities and measure whether the ASR content preserves core contextual meaning; (2) *s*: The summary of this speech [] in *one word* is. This assesses the extent to which ASR errors affect the LLM's high-level understanding. Instead of directly comparing LLM-generated words, we use the last layer of hidden states of LLM for generating the next token as embeddings and compute their cosine similarity with the embeddings generated from the ground truth transcription:

$$\text{Sim}_b = \cos(\mathcal{M}_b(\text{ASR}), \mathcal{M}_b(\text{Ground})) \quad (5.1)$$

$$\text{Sim}_s = \cos(\mathcal{M}_s(\text{ASR}), \mathcal{M}_s(\text{Ground})) \quad (5.2)$$

where $\mathcal{M}_-$ denotes the LLM’s hidden states. We then select the **lower** similarity as the final Sim score since we intend to capture the semantic gap between the ASR and the ground truth. This embedding-based approach enables a robust evaluation of semantic preservation in ASR outputs, beyond simple word-matching metrics.

5.3.3 Apply: QA Accuracy

The Apply level evaluates a $\text{LLM}_{\text{Voice}}$ ’s ability to leverage transcribed information for solving downstream problems, reflecting its real-world utility in task-oriented scenarios. We simulate this utility by constructing multi-choice QA, which is a typical listening test for our human language learners. Specifically, for each speech example, we construct 3 questions based on the ground-truth transcription along with 5 choices per question. (including 1 option as “None of the above”)

During QA generation, we leverage GPT-4o and prompt it to focus on either the core concept or information details in the transcription. (Appendix 5.5) To mitigate potential errors in generated QA, we also employ GPT-4o itself and Gemini-1.5-flash to answer these questions using the ground-truth transcription. Questions, where GPT-4o or Gemini-1.5-flash fails to produce correct answers, are discarded and regenerated. Then during evaluation, cascaded systems answer questions based on ASR transcriptions, while end-to-end systems directly process speech inputs to generate responses without intermediate ASR steps. Each $\text{LLM}_{\text{Voice}}$ answers 5 times per question and use the majority vote as the final answer [183]. QA accuracy is then measured via an exact match and higher accuracy indicates a stronger capability to contextualize and apply spoken content to real-world tasks.

Moreover, we realize that this QA test can also help recognize potential annotation errors in the benchmarks efficiently, since QA generated by annotation errors will be unanswerable to $\text{LLM}_{\text{Voice}}$, we use this feature to filter out annotation errors in our main experiments and further utilize them to detect hallucinations in Section 5.6.5.

5.3.4 Final SIQ Score

Inspired by standard Raven’s progressive matrices [75], a popular human IQ mechanism, we have three steps to compute SIQ: (1) harder samples have higher scores; (2) global standardization making scores among levels computable; (3) dynamic weight among levels in computing final SIQ. For each LLM_{Voice} model j and each speech sample i , we first compute raw scores for the three dimensions as $X_{\text{WER},i,j}^{\text{dim}}$ (taking the negative in subsequent computation), $X_{\text{Sim},i,j}^{\text{dim}}$, and $X_{\text{Acc},i,j}^{\text{dim}}$.

Speech Sample Discrimination Weights To account for variations in sample difficulty, we introduce “discrimination weights” based on inter-model score variance. For each speech sample i , we compute the variance across models:

$$V_i^{\text{dim}} = \text{Var}([X_{\text{Raw},i,j}^{\text{dim}} \ \forall j]) \quad (5.3)$$

where ”dim” refers to 3 levels. Using these variance values, we compute weighted scores:

$$X_j^{\text{dim}} = \frac{\sum_{i=1}^N X_{\text{Raw},i,j}^{\text{dim}} \cdot V_i^{\text{dim}}}{\sum_{i=1}^N V_i^{\text{dim}}} \quad (5.4)$$

where N denotes the number of models. This ensures that speech samples with greater discrimination power have a larger influence.

Global Standardization For each dimension, we perform standardization to normalize scores across models:

$$Z_j^{\text{dim}} = \frac{X_j^{\text{dim}} - \mu^{\text{dim}}}{\sigma^{\text{dim}}} \quad (5.5)$$

where μ_{dim} and σ_{dim} are the mean and standard deviation of all model scores for that dimension.

Dynamic Weight Computation To ensure a balanced contribution of each evaluation dimension, we assign dynamic weights based on variance. Since higher variance may indicate greater instability, we use the “inverse variance”:

$$w^{\text{dim}} = \frac{1}{\sigma_{\text{raw}}^{\text{dim}} + \epsilon}, w_f^{\text{dim}} = \frac{w^{\text{dim}}}{\sum w^{\text{dim}}} \quad (5.6)$$

where $\sigma_{\text{raw}}^{\text{dim}}$ is the standard deviation computed by all raw scores, and ϵ is a small constant to prevent division by zero. The weights are then normalized on the summation of three dimensions.

Final IQ Score Computation The final intelligence score for each model j is computed as a weighted sum of the standardized scores:

$$Score_j = \sum_{\text{dim}} w_f^{\text{dim}} \cdot Z_j^{\text{dim}} \quad (5.7)$$

Finally, the score is converted into an IQ-like scale:

$$SIQ_j = 100 + 15 \cdot Score_j. \quad (5.8)$$

As one additional study, we also perform normalization based on model scale (*i.e.*, analogous to age factors in human IQ). However, due to the involvement of multiple variables in neural scaling laws [79], we report the non-normalized SIQ and provide preference rankings from a human study to validate its correlation.

5.4 Experiment Setup

5.5 QA Generation

Since we intend to follow the English test for human learners, our prompt emphasize our intention to instruct GPT-4o is in Figure 5.3. We then add another option “None of the above” to each question. Besides, to mitigate potential hallucinations in question generation, we employ a two-step validation process: (1) GPT-4o generates three candidate questions per speech example, and (2) GPT-4o itself and Gemini-1.5-flash attempt to answer these questions using the ground-truth transcription. Questions where GPT-4o or Gemini-1.5-flash fails to produce correct answers are discarded and regenerated, ensuring robust alignment between audio content and QA pairs

"A listening test is to answer the questions based on the audio context, and the questions reflect the core information in the audio context. Below is the audio context, please generate 3 different questions (depending on the length of the context), each should be a 4-option question including the correct answer.

"The format will be:
 Question: What is $1 + 1$?
 (A) 1
 (B) 2
 (C) 3
 (D) 4
 Correct Answer: [[B]]
 Note that to avoid the random guess of the correct answer, other 3 options should be similar and confusing in listening."

f"Audio context: {text}"

Figure 5.3: Prompt for QA generation

5.5.1 Datasets

To comprehensively evaluate $\text{LLM}_{\text{Voice}}$ intelligence across varied domains and real-world challenges, we carefully select two datasets earning22 [150] and voxpopuli [181] from the popular OpenASR Leaderboard, while an additional dataset Med-ASR-EN² is chosen from the medical domain, featuring diverse accents and environments. This ensures that our evaluation tests $\text{LLM}_{\text{Voice}}$ performance in different real-world scenarios, including domain-specific speech and challenging acoustic conditions. Considering the high computational cost of leveraging GPT-4o APIs for QA evaluation and running end-to-end $\text{LLM}_{\text{Voice}}$ (e.g., Gemini), we extract a subset constructed by the longest audios from each dataset shown in Table 5.1.

5.5.2 Models

We conduct experiments across three major $\text{LLM}_{\text{Voice}}$ architectures as follows:

²<https://huggingface.co/datasets/jarvisx17/Medical-ASR-EN>

Dataset	# Subset	Domain	WER
Earning22	200	Financial meetings	12.05
Voxpopuli	200	European Parliament	7.48
Medasr	400	Hospital Accented Patients	7.7

Table 5.1: **Statistics of datasets.** WER results are on whole dataset via Whisper-large-v2.

ASR+LLM We compare the following widely used ASR models: Whisper-Large-v2, Whisper-Large-v3 [139], Canary [135], and ESPnet (owsm_ctc_v3.1_1B) [186]

ASR+GER+LLM We select GPT-4o [129] as the GER module and use Whisper-large-v2 to generate 5 hypotheses for each audio based on beam search.

End-to-end Multimodal Models For speech-to-text multimodal models, we compare Salmonn [170], Qwen2-audio [31], desta2 [103] and the latest Qwen-2.5-Omni[195], we also involve any-to-any multimodal models with speech-to-text capabilities: AnyGPT [207], Baichuan-omni-1.5 [96], and Gemini-1.5 [172]. We specifically denote Baichuan-omni-1.5 and Gemini-1.5 as “end-to-end-large” since they are either scaled up in training data or model size.

In the understanding level test, we select LLaMA-3.1-8B-Instruct to generate hidden states as responses since related work [73, 101] has proved the strong embedding capability of LLaMA-based models.

Meanwhile, for the LLM in both ASR + LLM and ASR + GER + LLM to answer QA, we use Qwen2-7b [197] for fair comparisons in size. Preliminary experiments to validate Qwen2-7b and other experiment details are in Section 5.5.3.

5.5.3 Setup Details

We keep the default settings (e.g., temperature) for all models referring their official repositories or huggingface pages.

ASR	Answering LLM	Accuracy
Whisper-large-v2	GPT-4o	93.3
Whisper-large-v2	Qwen2-7B	93.1
Whisper-large-v3	GPT-4o	92.3
Whisper-large-v3	Qwen2-7b	92.3
Canary	GPT-4o	93.7
Canary	Qwen2-7B	93.4

Table 5.2: **Results of same ASR models with different LLMs in answering QA.**

Validation of Qwen2-7B

To ensure the robustness of Qwen2-7B in cascaded approaches, we compared with GPT-4o for several ASR models. As shown in Table 5.2, Qwen2-7B shows a very closer performance with GPT-4o in QA, as well as a similar model size with end-to-end models for fair comparisons.

Prompts for end-to-end models

Considering that a model-specific prompt is not realistic due to the closed training data for some end-to-end models, we use a unified prompt for instructing ASR transcription as in Figure 5.4

Prompts for cascaded models in QA

Our prompt for instructing cascaded LLM_{VoiceS} in QA as in Figure 5.5.

ASR for end-to-end:

"The transcription of the audio is"

QA for end-to-end:

"Given the audio, what is the correct answer to the question?"

Note that if it is confusing to select the correct answer based on the context or the information is lost in the context, please select `[[E]]`.

f"{question}"

Correct Answer: "

Figure 5.4: Prompt for end-to-end models

QA for cascaded LLM:

"Here is an English test, given the question, please select the correct answer based on the context.

Note that if it is confusing to select the correct answer based on the context or the information is lost in the context, please select `[[E]]`."

f"Context: {context}"

f"{question}"

"Correct Answer:"

Figure 5.5: Prompt for QA generation

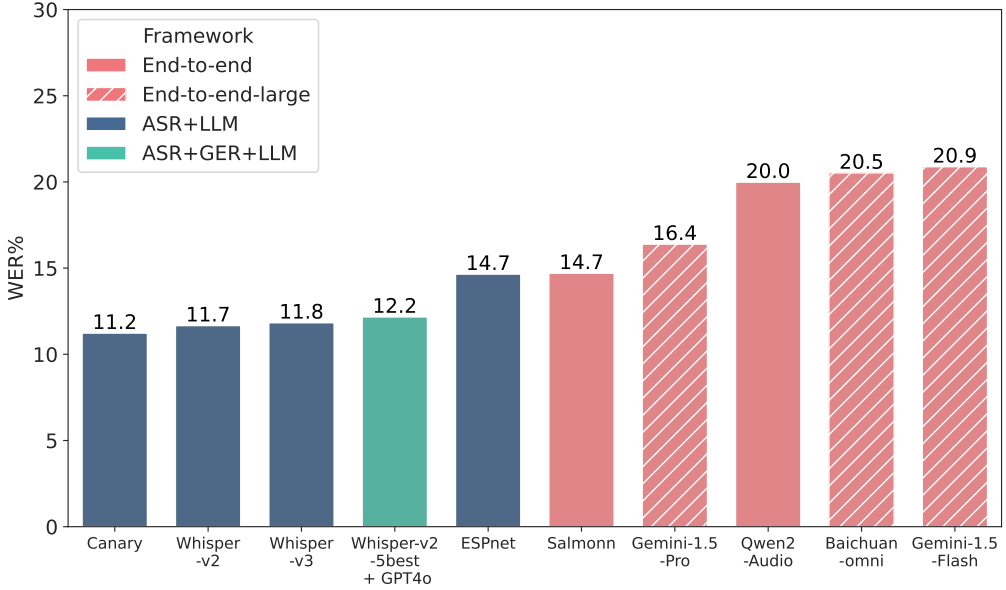


Figure 5.6: **Raw remember scores.** All cascaded models outperform end-to-end models including Gemini-1.5.

5.6 Experimental Results

In this section, we will first show the results of each level, and then the final SIQ scores. Considering the limited space, for each level we show the average score on 3 datasets without Desta2 and AnyGPT since they show a relatively low performance, the complete results are in Appendix.

5.6.1 Remember-Level Results

Figure 5.6 shows the raw WER scores of each model. Among all tested models: (1) Canary achieves the lowest WER, demonstrating the best remembering capability; (2) GER even negatively affects ASR if we focus on the increasing WER; (3) More broadly, all ASR models significantly outperform end-to-end LLM_{Voice} in terms of WER, and even large-scale end-to-end models, such as Gemini-1.5, perform worse than ASR models, reinforcing the robustness of traditional dedicated ASR models.

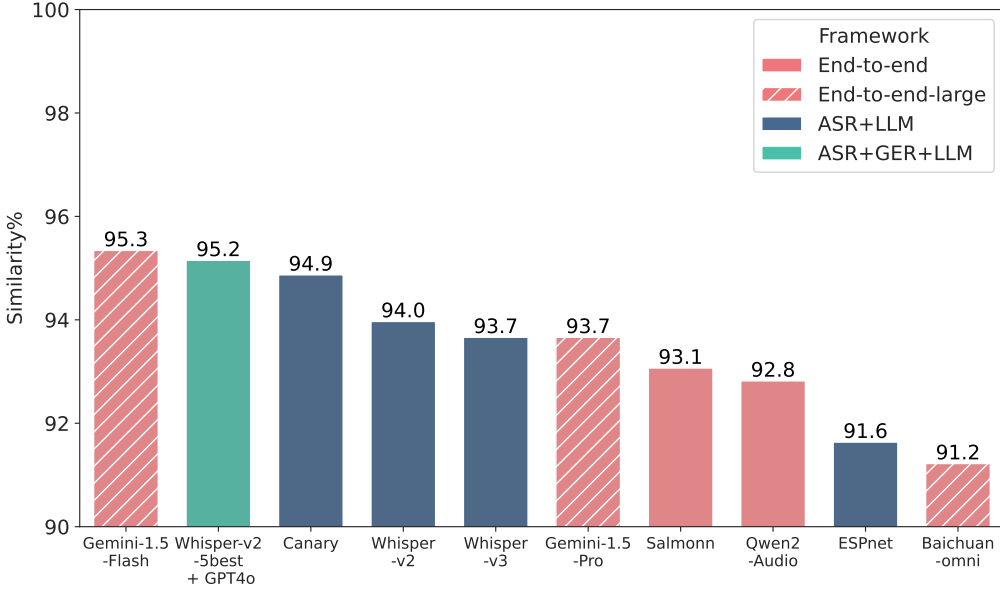


Figure 5.7: **Raw understand scores.** Gemini-1.5-flash shows the best semantic capturing.

5.6.2 Apply-Level Results

As in Figure 5.8: (1) While end-to-end models with similar sizes as cascaded models make much poorer accuracies, all scaling-up end-to-end models achieve superior performance in capturing apply level knowledge, which is consistent with the scaling law; (2) GER not only improves semantic retention but also enhances downstream reasoning capabilities in the apply level.

5.6.3 SIQ Scores

Table 5.3 shows the final SIQ scores and we can observe: (1) A key observation is that the rankings based on WER do not hold for overall LLM_{Voice}’ intelligence: while single ASR models outperform ASR+GER and end-to-end approaches on WER, they fail to maintain their lead in higher-level intelligence evaluations; (2) End-to-end LLM_{Voice} systems underperform cascaded models of the same scale, but when model size increases as Gemini-1.5, they achieve competitive SIQ with cascaded approaches; (3) GER provides stable SIQ score improvements

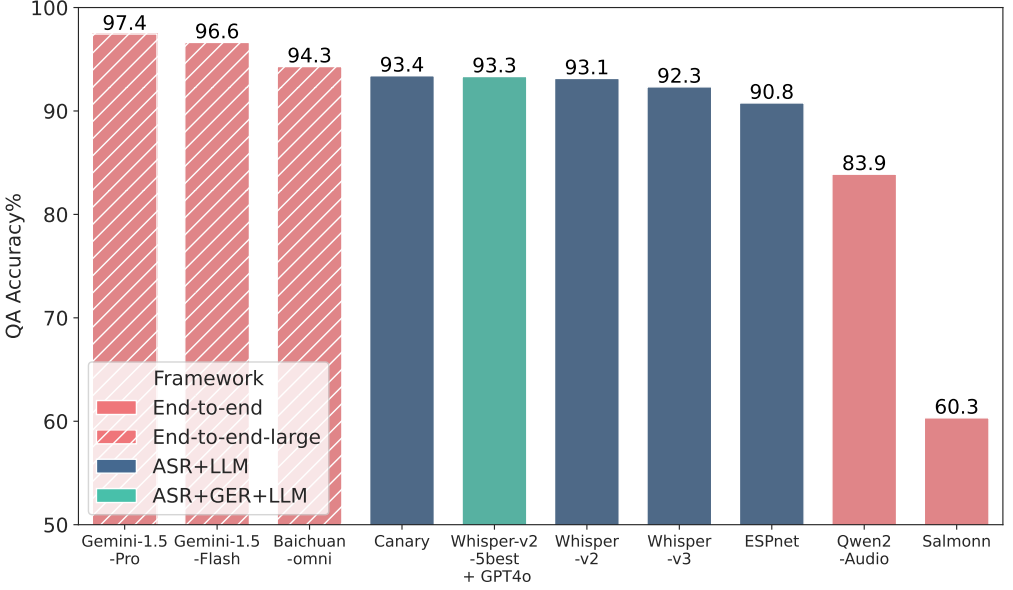


Figure 5.8: **Raw apply scores.** End-to-end-large models perform best while smaller ones perform worst.

for ASR+GER+LLM models and achieve the best SIQ.

Human Evaluation on Comparing with Existing Metrics To validate that SIQ better reflects the voice understanding capabilities of LLM_{Voice} models compared to existing metrics, we conduct a human evaluation on nine examples involving ten randomly selected anonymous models to avoid human bias on models. For each example, we invite ten human experts, including both native speakers and non-native speech researchers, to rank the ten transcriptions produced by each model based on the ground-truth transcription. We then aggregate these rankings to derive a final rank for each model and compare the results with the rankings obtained from existing metrics on the same nine examples. For SIQ, we compare two variants: (1) SIQ_{rm}: compute on 9 examples but removing weight standardization due to the limited sample size; (2) SIQ_{all}: the SIQ in Table 5.3. Then we compute Spearman’s Rank Correlation Coefficient ρ with human rankings. In Table 5.4, SIQ consistently outperforms existing metrics in capturing human

Relative Bloom 's Taxonomy Levels	Remember $\uparrow$	Understand $\uparrow$	Apply $\uparrow$	SIQ $\uparrow$
<i>ASR+LLM Approaches</i>				
Whisper _{v2-1.5B} + Qwen2 _{7B}	0.554	0.499	0.481	107.43
Whisper _{v3-1.5B} + Qwen2 _{7B}	0.553	0.433	0.432	106.49
Canary _{1B} + Qwen2 _{7B}	0.559	0.566	0.504	107.78
OWSM-CTC _{v3.1-1B} + Qwen2 _{7B}	0.534	0.151	0.353	103.05
<i>ASR+GER+LLM Approach</i>				
Whisper _{v2-1.5B} + GPT-4o + Qwen2 _{7B}	0.543	0.632	0.487	108.64
<i>Multi-Modal End-to-End Approaches</i>				
Qwen2-Audio _{7B} w/ 1.5B Whisper	-0.187	0.366	0.011	103.88
Qwen2.5-Omni _{7B} w/ 1.5B Whisper	0.472	0.410	0.509	105.74
Salmonn _{13B} w/ 1.5B Whisper	0.508	0.381	-1.146	101.03
Desta2 _{8B} w/ 1.5B Whisper	-2.575	-1.604	-0.233	79.69
AnyGPT _{7B}	0.314	-2.718	-2.893	60.02
Baichuan-omni-1.5 _{7B}	0.448	0.184	0.546	104.02
Gemini-1.5-flash	-1.885	0.641	0.673	107.85
Gemini-1.5-pro	0.492	0.409	0.710	107.08

Table 5.3: **Main Results.** The results of remember and understand levels are from single experiment, while apply level include the majority results of 5-time generations.

preferences, particularly SIQ_{all}, demonstrating its effectiveness in assessing voice understanding capabilities.

5.6.4 Cascaded vs End-to-end

Our SIQ scores provide strong empirical evidence that cascaded systems achieve significantly higher intelligence scores compared to end-to-end models of similar size. However, a crucial factor influencing this result is the LLM used in QA. Cascaded systems can typically utilize the latest LLMs (e.g., Qwen2-7b, GPT-4o) for QA-based reasoning, whereas end-to-end models are often constrained by their foundation LLMs that jointly handle speech tokens and language understanding. To eliminate this confounding factor, we conducted an ablation experiment comparing cascaded and end-to-end systems built on the same base model.

As shown in Table 5.5, the cascaded approach consistently outperforms the end-to-end model in 3 levels. Specifically, Qwen2-audio and Salmonn are largely

LLM _{Voice}	Human	WER	SemDist	LLM-S	BLEU	SIQ _{rm}	SIQ _{all}
Model _A	3	2	4	4	3	4	2
Model _B	8	10	8	8	6	8	10
Model _C	7	7	7	6	8	7	7
Model _D	1	4	2	5	4	3	1
Model _E	4	6	5	7	5	6	4
Model _F	6	5	6	2	7	5	6
Model _G	2	3	1	3	2	2	3
Model _H	5	1	3	1	1	1	5
Model _I	9	8	10	10	9	10	8
Model _J	10	9	9	9	10	9	9
correlation ρ ($\uparrow$)	-	0.770	0.721	0.624	0.806	0.830	0.952
p -value ($\downarrow$)	-	0.009	0.005	0.019	0.054	0.003	0.00002

Table 5.4: Human Voted Rankings & Existing Metrics.

worse than their cascaded variants at the apply level. We hypothesize that the multi-modal training may sacrifice the original reasoning capability of foundation models if not well designed, which is crucial for understanding the limitations of joint training for multi-modal intelligence.

Additionally, we observe that Desta-2 exhibits unexpectedly high QA accuracy while its performances on the WER and similarity are pretty poor. Interestingly, the base model of Desta-2, LLaMA3-8B-instruct, also achieves a higher-than-expected performance than cascaded experiments. We implemented a closer case study (Appendix 5.7) which reveals that Desta-2 may stem from LLaMA3-8B-instruct showing a strong tendency to hallucinate: (1) Guess the answer even not been recognized in its ASR results; (2) Change the options in the question. This indicates that SIQ can not be represented by the apply level alone, which may suffer from the hallucination issue in question answering.

Moreover, manually verifying hallucinated QA evaluation is prohibitively expensive. Inspired by the “annotation error” introduced in Section 5.3.3, in the following section, we introduce an unanswerable set to detect hallucinations in LLM_{Voice}.

Model	WER ↓	Similarity ↑	Accuracy ↑	Hallucination ↓
Whisper _{v2-1.5B} + Qwen _{7B}	11.8	93.7	91.7	2 (12.0%)
Qwen2-Audio _{7B}	20.0	92.8	83.9	4 (24.0%)
Whisper _{v2-1.5B} + Llama _{38B-ins}	11.7	94.0	95.7	7 (41.2%)
Desta _{28B} w/ 1.5B Whisper	76.55	81.6	83.9	12 (70.6%)
Whisper _{v2-1.5B} + Vicuna _{13B-1.1}	11.7	94.0	87.3	1 (6.8%)
Salmonn _{13B} w/ 1.5B Whisper	14.70	93.1	60.3	4 (24.0%)

Table 5.5: **Cascaded v.s. End-to-end including hallucination detection.** Cascaded models perform better correspondingly. Desta2 shows unexpectedly high QA accuracy and hallucinations along with its foundation Llama_{38B-ins}.

5.6.5 Unanswerable Set: Detecting Hallucination in LLM_{Voice}

One major advantage of QA-based evaluation is that it allows us to efficiently identify potential annotation errors. If a given QA pair consistently fails across most LLM_{Voice}, it is likely that the failure is due to incorrect or ambiguous annotations resulting in unanswerable questions rather than genuine model errors.

To leverage this property, we filter out questions that the majority (a threshold of "half") of LLM_{Voice} fail to answer correctly and manually check the corresponding speech and annotations. This significantly reduces human efforts overhead (only 26 out of 800 samples require human review). After verification, we confirm that 17 questions are truly unanswerable based on the available speech content, which forms our unanswerable Set for detecting hallucinations.

We measure both cascaded and end-to-end LLM_{Voice} in Table 5.5. For each unanswerable question, the hallucination will be counted when the answer to the unanswerable question is not "(E) None of the above,". Our results indicate that LLaMA3-8B-Instruct exhibits a significantly higher hallucination ratio than other LLMs, and this problem gets even worse in its end-to-end variant, which explains the higher-than-expected performance in apply level test. This inspires us that the hallucination of foundation LLMs may be inherited by its multi-modal variances, emphasizing the importance of foundation model selection and reducing hallucination in multi-modal training.

Model	Remember ↓	Understand ↑	Apply ↑
Qwen2-Audio _{7B} (end2end)	20.00	92.8	83.9
Qwen2-Audio _{7B} (cascaded)	20.00	92.8	84.2
Gemini-1.5-flash (end2end)	20.9	95.3	96.6
Gemini-1.5-pro (cascaded)	20.9	95.3	95.1

Table 5.6: **Take ASR from end-to-end models**

5.6.6 End-to-end Models on the Cascaded Category

We then checked the performance of end-to-end models (Qwen2-audio and Gemini-1.5-flash) in the cascaded category by replacing the ASR model by the end-to-end model, and then used the LLM in the cascaded category to answer questions, noting that this variant only affects the apply level evaluation. Interestingly, in figure 5.6 we find that for the smaller model Qwen2-audio, cascaded version improves the apply level while for larger model Gemini, end-to-end version wins. This may indicate that higher intelligence may result in better end-to-end apply level performance even if sometimes perfect transcription is not generated. We further extend this assumption by involving a human SIQ test, we manually select 3 hard examples and invite 10 human individuals to report the average scores as in Table 5.7. It seems that humans are still good at apply level even with the worst WER performance in transcription, which is consistent with our assumption that higher intelligence may largely benefit higher-order cognitive levels. This may inspire future work on how to improve the intelligence level of LLM_{Voice}.

5.7 Case Study

We manually check the outputs of Dsta2 and LLama3-8B-Instruct to investigate the reason behind unexpectedly high QA accuracy. As shown in Figure A.1a A.1b, in 5-time generations, the hallucination of Llama3-8B-Instruct lies in either giving a correct answer based on the error ASR or directly changing the options in QA, and interestingly, this hallucination is kept by its end-to-end variant Dsta2.

"ASR": "We discussed a couple of times competitive landscape, but on your new markets like Nordics or Dash, who do you see usually in those markets and do you see any fast-growing acquirers in those markets moving from online to in-stores and trying to compete on the SMBs or is it not already the case?" "question": "In which new markets is there a discussion about competitive landscape?" "options": "(A) Nordics and Spanish\n(B) Nordics and Dutch\n(C) Dutch and French\n(D) Dutch and Italian " "correct answer": "B", "whisper-large-v2_meta-llama/Meta-Llama-3-8B-Instruct_answer": <pre>["Based on the context, the correct answer is:\n\n(B) Nordics", "Based on the context, I would select:\n\n(A) Nordics and Dash\n\nThe question mentions '\nNordics or Dash\n' as the new markets where the competitive landscape is being discussed.", "Based on the context, I would select:\n\n(A) Nordics and Dash", "Based on the context, the correct answer is:\n\n(A) Nordics and Dash", "Based on the context, the correct answer is:\n\n(A) Nordics and Dash"]</pre>	"question": "In which new markets is there a discussion about competitive landscape?" "options": "(A) Nordics and Spanish\n(B) Nordics and Dutch\n(C) Dutch and French\n(D) Dutch and Italian " "correct answer": "B", "desta2_destas2_answer": <pre>["The correct answer is (B) Nordics and Dash.\n\nThe audio transcript mentions '\nNordics\n' as one of the new markets, and there is no mention of Dash as a country or region. Dash is likely referring to Dashlane, a password manager company.", "Based on the audio, the correct answer is (B) Nordics and Dash.", "Based on the audio, the correct answer is (B) Nordics and Dash. The speaker mentions '\nnew markets like Nordics or Dash\n' in the question.", "The correct answer is (B) Nordics and Dash.\n\nThe audio snippet mentions '\nNordics\n' as one of the new markets, and '\nDash\n' is likely a typo or a mistake, as it's not a common market name.", "Based on the audio, the correct answer is (B) Nordics and Dash."]</pre>
--	--

Figure 5.9: Left: Llama3-8B-Instruct, Right: Desta2. **Red** the errors in answering, we find that Llama-8B-Instruct based on the error ASR even hallucinate to change the option in QA, while Desta2 though answer correctly, still inherit this hallucination.

"ASR": " and the mixed effect has to do with the fact that the net, not Germany per se, but net was more exposed and still more exposed in terms of mix of sectors to the high impact sectors in terms of volumes." "question": "Which entity is mentioned as being more exposed in terms of mix of sectors?" "options": "(A) Germany \n(B) Nexi \n(C) France \n(D) Italy " "correct answer": "B", "whisper-large-v2_meta-llama/Meta-Llama-3-8B-Instruct_answer": <pre>["Based on the context, I would select:\n\n(B) Net", "Based on the context, the correct answer is:\n\n(B) Net", "Based on the context, I would select:\n\n(B) Net", "Based on the context, the correct answer is:\n\n(B) Net", "Based on the context, I would select:\n\n(B) Net"]</pre>	"question": "Which entity is mentioned as being more exposed in terms of mix of sectors?" "options": "(A) Germany \n(B) Nexi \n(C) France \n(D) Italy " "correct answer": "B", "desta2_destas2_answer": <pre>["Based on the given audio, the correct answer is:\n\n(B) Net", "Based on the audio transcript, the correct answer is:\n\n(B) Net\n\nThe speaker mentions that '\nthe net, not Germany per se, but net was more exposed in terms of mix of sectors to the IE impact sectors in terms of volumes.\n' This indicates that the entity being referred to is '\nnet\n', which is likely a financial institution or a company, and not Germany or any other country.", "Based on the audio, the correct answer is (B) Nexi.", "Based on the audio, the correct answer is:\n\n(B) Net", "Based on the audio transcript, the correct answer is (B) Net."]</pre>
--	--

Figure 5.10: Left: Llama3-8B-Instruct, Right: Desta2. **Red** the errors in answering, we find that Llama-8B-Instruct gives the correct answer but based on the error ASR and also hallucinate to change the option in QA, while Desta2 keeps the same hallucination.

Model	Remember ↓	Apply ↑
Whisper _{v2-1.5B} + Qwen _{7B}	23.71	88.89
Whisper _{v2-1.5B} + Qwen _{7B}	17.26	88.89
Canary _{1B} + Qwen _{7B}	24.80	88.89
Whisper _{v2-1.5B} + GPT-4o + Qwen _{27B}	21.13	88.89
Salmonn _{13B} w/ 1.5B Whisper	31.55	44.44
Qwen2-Audio _{7B}	25.50	88.89
Gemini-1.5-flash	19.05	66.67
Gemini-1.5-pro	19.05	77.78
Human	53.67	99.27

Table 5.7: **Human test on 3 examples.**

5.8 Limitations and Discussions

In our attempt to provide an examination of Speech IQ, our study limits that warrant discussion.

Data and Evaluation Limitations Although this work represents the first systematic investigation of SIQ assessment with demonstrated effectiveness, our current evaluation is based on moderately sized test sets. In future work, we plan to extend our validation framework across diverse domains and languages, and we will employ rigorous quality assurance techniques to filter annotation errors in widely used benchmarks. Moreover, while using publicly available datasets ensures reproducibility, it also introduces potential risks of data leakage. To mitigate this, we are developing closed datasets with controlled knowledge cutoffs for subsequent research phases.

Scaling and Quantitative Analysis Our results indicate that SIQ is sensitive to model scaling effects, including variations in architecture size and training data volume. However, our current analysis does not quantitatively characterize these

relationships, such as a limitation reminiscent of the challenges in interpreting human IQ with age-group normalization. As discussed in Appendix C.2, our next SIQ iteration will incorporate scaling law normalization protocols to decouple intrinsic voice understanding capabilities from artifacts induced by parametric scaling, enabling a more nuanced analysis of model performance.

Ethical and Societal Considerations The introduction of SIQ raises important ethical and societal questions. Since IQ classification in humans can lead to social discrimination, we are concerned that analogous issues might arise in AI, potentially resulting in biased treatment of systems with lower SIQ scores. We are committed to addressing these concerns proactively and developing safeguards to prevent such discrimination.

Upper 3 Levels in Bloom Taxonomy Our study focuses only on the bottom three layers of the Bloom Taxonomy, leaving the upper three layers unexplored. While this approach provides a solid foundation for understanding the core aspects of the problem, it does not capture the full hierarchical structure. The upper layers, which may involve more advanced levels of *analysis*, *evaluation*, and *creation*, remain an open avenue for our future research incorporating audio generation, physical simulation, and acoustic event reasoning models [88]. Expanding the analysis to all 6 layers could provide a more comprehensive understanding of its broader implications toward a form of “audio intelligence” for voice assistants.

5.9 Summary of This Chapter

Existing metrics fail to comprehensively evaluate $\text{LLM}_{\text{Voice}}$ ’s capability in semantic understanding or task-solving. Therefore, we build a three-level SIQ test following Bloom’s taxonomy, each level owns unique features and represents different intelligence levels. Our study involves various $\text{LLM}_{\text{Voice}}$ frameworks, and the SIQ test serves as a unified metric for benchmarking different frameworks. We then discuss whether current end-to-end $\text{LLM}_{\text{Voice}}$ s outperform cascaded methods with the same size, indicating the potential insights in training multi-modal models. Moreover, benefiting from the QA test, we need much less human effort to

filter out annotation errors in existing benchmarks and build an unanswerable set, which can further help evaluate the hallucination of $\text{LLM}_{\text{Voice}}$ which is also crucial for multi-modal training.

Relationship to the Dissertation: Evaluating Scaling Inference in Agentic Systems As the marginal returns of model parameter scaling begin to diminish, the frontier of AI research has shifted toward scaling inference. While inference-time computation can be scaled through repeated sampling within a single model, this dissertation explores a more sophisticated trajectory: agentic systems. In these systems, multiple specialized models interact iteratively to achieve a "1+1 > 2" synergistic effect, as demonstrated by the collaborative architecture of speech and text models in this chapter.

However, as we transition toward this agentic scaffolding, traditional evaluation paradigms have become a bottleneck. Existing benchmarks are predominantly human-facing, failing to account for how a model's output impacts downstream agents or whether the communication between machines should strictly mimic human understanding. This chapter, through the introduction of SpeechIQ, initiates a fundamental rethinking of evaluation: if agents can communicate more efficiently through non-human-centric protocols, then "agent-centric utility" must supersede traditional metrics.

By establishing this new benchmark, we provide the necessary infrastructure for the next generation of agentic scaffolding. It enables the development of new reward mechanisms optimized for system-level utility rather than mere surface-level mimicry of human speech. This chapter marks the dissertation's transition from designing individual reasoning layers to architecting and auditing the intelligence of collectives.

Chapter 6

Speech-Hands: A Self-Reflection Voice Agentic Approach to Speech Recognition and Audio Reasoning with Omni Perception

6.1 Introduction

Omni-modal models [193, 128, 195, 96] that jointly process audio and text have unified a range of audio understanding tasks, including automatic speech recognition (ASR), temporal sound event reasoning, and knowledge-heavy question answering. However, human perception is not naturally perfect at understanding acoustic patterns across different resolutions at soundscape [19]. For example, while professional speech interpreters could produce superior ASR transcriptions, this specialized ability does not guarantee a comparable aptitude for understanding animal sounds or complex music [42].

Inspired by developmental psychology [160], we draw a parallel to the human capacity for self-reflection, where children mature from a purely egocentric viewpoint to a stage of *self-reflective perspective-taking*, which serves the critical ability to “step outside” one’s own thoughts, evaluate one’s beliefs against others’, and im-

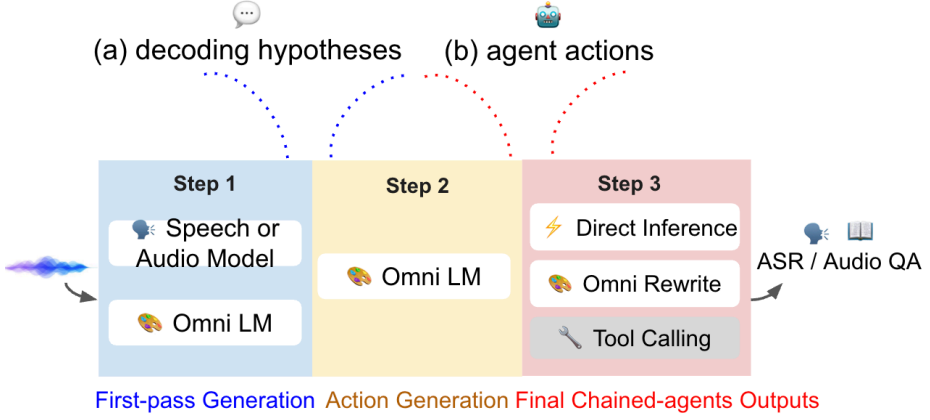


Figure 6.1: *Speech-Hands* acts as a dynamic orchestrator that predicts a special action token to govern its cognitive strategy.

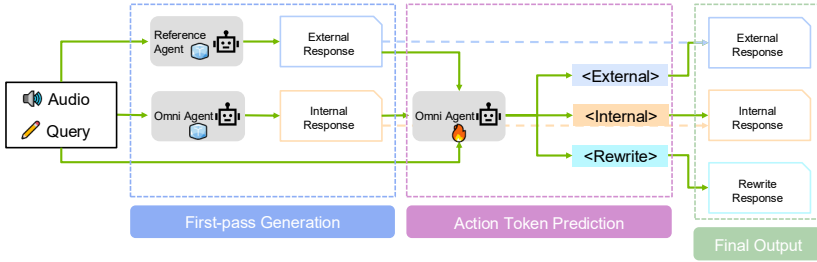


Figure 6.2: Overview of our proposed Self-Reflection Multimodal GER framework. A special token is generated at the beginning to decide whether to use audio perception (i.e., transcription hypotheses or caption) or not.

portantly, to recognize the boundaries of one’s own knowledge. In contrast, current models often operate egocentrically, implicitly trusting their internal perception without the capacity to critically assess its reliability or seek external assistance when necessary. We aim to instill a form of computational self-reflection [125] into an omni-modal agent, designing a collaborative framework that explicitly reasons about when to trust its own perception, when to defer to an expert, and even when to utilize tools.

We frame these voice understanding models not just as passive predictors, but as agentic that have access to multiple internal and external information sources,

and must decide how to best use them. For such an agent, a central decision-making challenge arises [91]: should it rely on its own auditory perception, or consult external suggestions, such as ASR alternatives or other perceptual consultants? Prior work, such as ASR and large language model (LLM) cascaded approach of Generative Error Correction (GER) [200], as shown in Figure 6.4, sidesteps this agentic dilemma entirely. By operating only on text hypotheses without access to the original audio, these methods are fundamentally non-agentic; they cannot weigh internal perception against external advice because they have no internal perception to begin with. Our preliminary experiments reveal that naively combining modalities often degrades performance, as the model struggles to resolve conflicts between its own perception and flawed external suggestions [78]. Without a mechanism to decide which source to trust, the model is easily confused.

To address this, we introduce *Speech-Hands*, a learnable framework that instantiates self-reflection as a core control primitive. As illustrated in Figure 6.1, our agent operates as a dynamic orchestrator. It begins by aggregating multi-source decoding hypotheses (a) from the first-pass generation. Instead of blindly fusing inputs, the agent critically evaluates them to predict an explicit agent action (b) during the action generation phase. This control token effectively dictates the model’s cognitive strategy: triggering fast direct inference when confidence is high (selecting whether its internal perception or external perceptions), engaging in omni rewrite over available evidence, or initiating tool calling when special utilities are required. In this work, we will focus on the actions of direct inference and omni rewrite, leaving the tool calling action in future work. This approach unifies transcription and reasoning under a single, controllable framework that knows when to trust, when to rethink, and when to ask for help.

6.1.1 Preliminary: The Surprising Failure of Multimodal Correction with Omni-LM

A natural hypothesis is that providing an omni model with both audio and text hypotheses during supervised fine-tuning (SFT) should enhance GER performance. We tested this assumption by fine-tuning Qwen2.5-Omni [194] to correct N-best hypotheses ($N = 5$) from Whisper-v2-large [140] on OpenASR datasets.

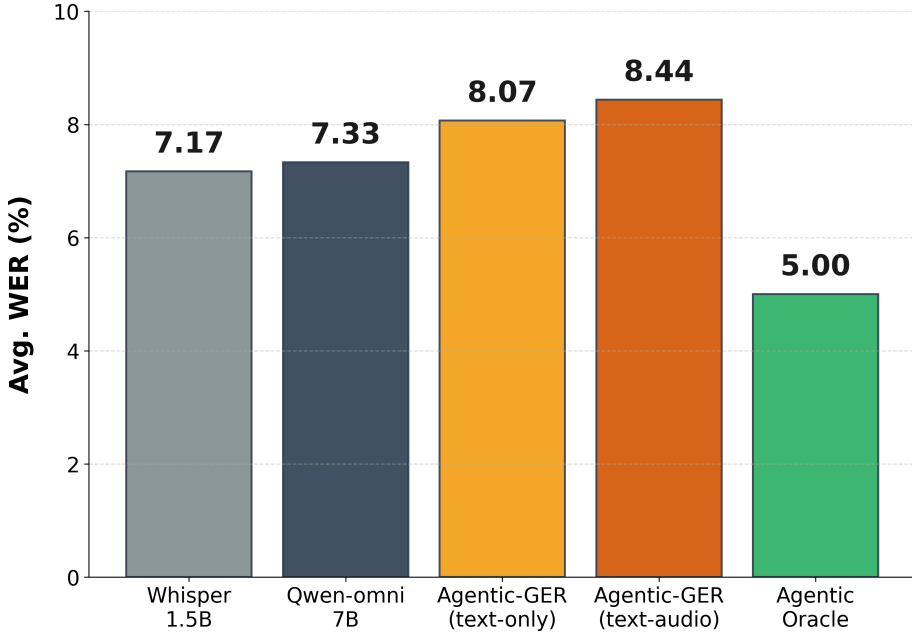


Figure 6.3: Preliminary results on the cascaded agentic Qwen-omni baseline for generative error correction (GER) with supervised fine-tuning show that both text-only and text-audio GER degrade ASR performance, where the best ASR and LLM combination achieves a low agentic output oracle of 5% WER.

Figure 6.3 results, however, are surprisingly negative. As shown in our preliminary study (Table 6.3), this naive SFT approach consistently degrades performance across seven ASR benchmarks, yielding a higher Word Error Rate (WER) than either of the baseline models alone.

To verify that this degradation was not due to suboptimal prompting, we extensively tested varying instructions during SFT, which aims to emphasize internal audio perception, external transcripts, or a balanced fusion. However, as detailed in Section 6.1.2, all prompting setups failed to recover performance (e.g., WER increased to 8.52%–9.05% compared to baselines). Furthermore, our zero-shot analysis reveals that the base model lacks intrinsic arbitration capabilities: its decisions are highly sensitive to prompt wording rather than ground truth, often collapsing into trivial heuristics (see Section 6.1.3).

These findings demonstrate a fundamental flaw in the naive omni-LM fusion

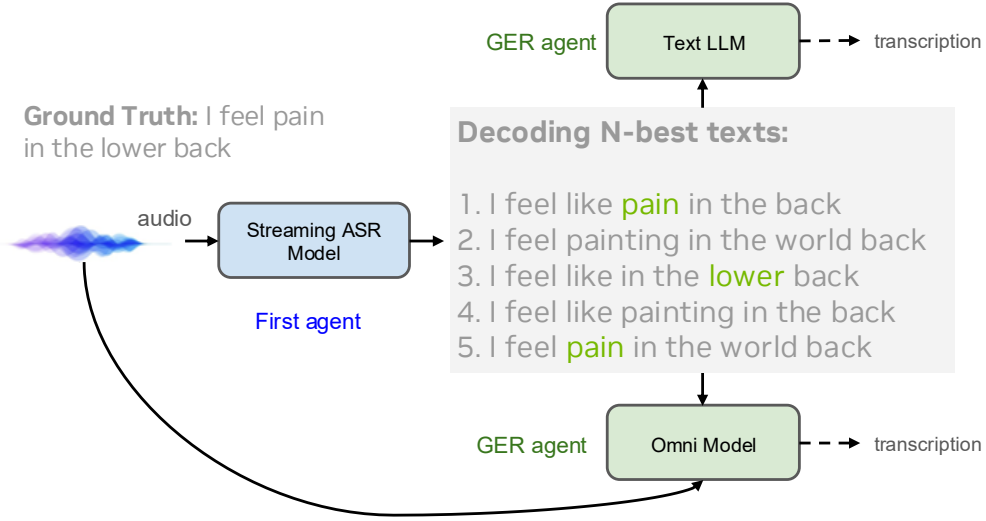


Figure 6.4: Text-base GER uses only ASR hypotheses. This setup fails to correct deletion or hallucination if all hypotheses are wrong. Multimodal GER include the original audio as grounding to improve error correction.

approach: without a mechanism to adjudicate between its own perception and potentially flawed external advice, the omni-model is easily confused, often amplifying hallucinations or overcorrections as shown in the ASR failure case (Appendix D.1 and Section 6.7.3). This provides strong motivation for a more principled mechanism that allows the model to learn when and how to incorporate external information.

6.1.2 Prompt Ablations in Preliminary SFT

We conduct these experiments during early-stage exploration: for inputs we investigated both (a) audio + external whisper 5-best and (b) audio + internal 1-best + external whisper 5-best, together with four prompting strategies instructing the model to emphasize internal hypotheses, external hypotheses, audio grounding, or a balanced fusion. The average WER results on OpenASR are shown in Table 6.1.

AVG. WER	(a)	(b)
Emphasize internal	-	8.63
Emphasize external	8.58	8.67
Emphasize audio	9.02	9.05
Balanced	8.44	8.52

Table 6.1: Prompt ablation results.

6.1.3 Zero-shot Qwen in Preliminary

We test three different zero-shot prompting strategies as shown in Table 6.2. We find that zero-shot decisions are highly prompt-sensitive: the model often collapses into trivial heuristics, while the balanced prompt produces unstable and inconsistent behavior across samples. The corresponding 2×2 confusion matrices show large off-diagonal mass for all zero-shot prompts, confirming that zero-shot Qwen does not perform genuine self-reflection.

Table 6.2: Confusion matrix of zero-shot decisions under different prompting strategies. The results show that the model’s arbitration is highly sensitive to prompt wording rather than the ground truth correctness (Oracle), often collapsing into trivial heuristics.

Ground Truth	Internal-biased Prompt		External-biased Prompt		Balanced Prompt	
	Pred. Int	Pred. Ext	Pred. Int	Pred. Ext	Pred. Int	Pred. Ext
Oracle Internal	0.83	0.17	0.35	0.65	0.68	0.32
Oracle External	0.71	0.29	0.14	0.86	0.34	0.66

6.2 Related Work

6.2.1 Voice Retrieval and Agentic Framework

Voice retrieval has become a useful component in augmenting audio tasks such as captioning [86, 213], audio-to-text generation [67], dialogue system [25], and music generation [47]. Recently, voice retrieval modules have been incorporated

into multimodal agents [204, 209, 185], providing access to external memory or specialized tools across modalities.

Yet, even in these agentic frameworks, retrieval is often treated as an auxiliary enhancer, rather than as a distinct source of information. When the retrieved content diverges from the internal prediction of the model, current systems lack principled mechanisms for arbitration.

6.2.2 Omni-Modality and Self-Reflection in Multimodal Models

By weaving together text, vision, and audio into a single fabric of understanding, these systems begin to approach the fluidity of human perception [194, 196, 46, 1]. Modality-specific reflection methods [62] suggest that introspection within each sensory channel can partially bridge these gaps, aligning representations with quiet precision.

Recent works also introduce explicit self-reflection [149, 112] into multimodal reasoning [28, 40]. The self-reflected video reasoner [165] iteratively critiques its own visual understanding to reinforce policy stability, while other efforts call for reflective checks against overconfidence and modality neglect [203]. Still, these mechanisms operate *after* perception, which treats reflection as a corrective mirror once fusion has already occurred.

Rather than reflecting on the output, our Speech-Hands framework reflects on the act of perception itself. It learns an action mechanism that decides, in real time, whether to trust its own “ears” or the “words” of others. We aim to discover self-reflection from a post-hoc repair strategy into a preemptive act of perceptual discernment toward an early glimmer of meta-cognition within multimodal understanding.

6.3 Methodology

We present **Speech-Hands**, a learnable omni-agentic framework for audio understanding and reasoning (Figure 6.2). It enables a multimodal language model to explicitly choose a special token: trusting its own internal perception or defer to external hypotheses, enables efficient **Fast Inference**, while rewriting a new

response engages the **Omni Rewrite** process for deeper reasoning. This token is generated during inference and guides the downstream generation process, allowing interpretable and agentic decision-making.

6.3.1 Task Formulation

We formulate a unified self-reflection agent that generalizes across both speech recognition and audio question answering. Given an input audio A , an optional query Q , the agent first generates its own response H_{omni} , and then combined with an external response H_{ext} provided by an external model.

Next, rather than fusing these sources implicitly as normal GER researches, Speech-Hands introduces a learned policy to explicitly choose among them. Our agent model first emits a special action token from the set $\{\langle\text{internal}\rangle, \langle\text{external}\rangle, \langle\text{rewrite}\rangle\}$ to indicate whether to trust itself or rely on external hypothesis or even whether rewriting a new response after rethinking the audio task and all input resources for the final answer (GER). This self-reflection decision is made based on the full context $(A, Q, H_{\text{omni}}, H_{\text{ext}})$, and the selected action conditions the final generation.

To supervise the self-reflection mechanism, we construct ground-truth action token labels on the whole training dataset by comparing the performance of internal, external, and GER predictions. We leverage detailed strategies for action token construction in ASR and audio QA.

6.3.2 Action Token Construction for ASR

In the ASR setting, we use WER as a pointer to decide our actions. For each audio example, we first prompt the omni model to generate the transcript T_{int} and in parallel leverage an ASR model to predict the external T_{ext} , we then let omni model to generate again based on both the audio and the external T_{ext} to acquire the GER prediction T_{ger} . Next, we compute the WER between the ground-truth transcript T_{gt} and each of the three candidates. If T_{int} is the same as T_{gt} ($WER = 0$) or has the lowest WER, the label is assigned as $\langle\text{internal}\rangle$, this is to encourage the model to trust itself when it can successfully solve the problem, otherwise $\langle\text{external}\rangle$ if T_{ext} performs best, or $\langle\text{rewrite}\rangle$ if T_{ger} performs best.

6.3.3 Action Token Construction for Audio QA

Unlike the ASR setting where token selection is based on fine-grained WER scores, Audio QA presents a discrete supervision signal: each prediction is either correct or incorrect. For each instance consisting of an audio segment A , a question Q , and answer choices $\mathcal{C}$, we first prompt the omni model to produce an internal prediction c_{int} based on (A, Q) . In parallel, we obtain an external prediction c_{ext} from an audio reasoning model.

We then compare both predictions against the ground-truth answer c^* . If $c_{\text{int}} = c^*$, we assign the label `<internal>`, encouraging self-reliance when the model performs well. If c_{int} fails but $c_{\text{ext}} = c^*$, we assign `<external>` to delegate control. Otherwise, when both predictions are incorrect, the label is set to `<rewrite>`, signaling a need to re-evaluate the question with all available context.

However, this binary decision process introduces instability during training. External predictions can be stochastic. Therefore, repeated sampling may yield different answers, especially under high uncertainty or directly change another external model can also lead to a different accuracy. This stochasticity makes the decision boundary between `<external>` and `<rewrite>` inherently less robust.

To mitigate this, we adopt a multiple decoding-based strategy: for each example, we sample the external model five times and collect their predicted choices. If the majority of predictions match the ground-truth answer, we assign `<external>`; otherwise, we assign `<rewrite>`. This approach stabilizes supervision by reducing the variance in external outputs and yields more reliable action labels.

6.3.4 Prompt Formatting and Training

Subsequently, each training instance is formatted as a single target string consisting of the decision token followed by the final target transcript or answer, e.g. `<rewrite>` + ground-truth transcription. This unified string allows the model to learn not only how to generate the task output but also how to decide the action before generation. We adopt an instruction-style prompt to guide the model to make decisions as shown in Appendix 6.5. During training, we optimize a single cross-entropy loss over the concatenated target sequence, which jointly supervises

Dataset	AMI	Tedlium	Gigspeech	Spgispeech	VoxPopuli	Libri-clean	Libri-other	avg. WER ↓
ASR model or Omni-LLM								
Whisper-v2-large	16.88	4.32	11.45	3.94	7.57	2.91	5.15	7.17
Canary-1b-v2	19.80	4.78	11.66	3.08	6.35	1.73	3.17	7.22
Parakeet-tdt-0.6b-v3	12.69	4.90	12.24	3.16	6.48	1.89	3.37	6.68
Qwen2.5_omni	19.77	5.17	11.26	4.58	6.59	2.09	3.85	7.33
Phi-4-MM	11.69	2.90	9.78	3.13	5.93	1.68	3.83	6.14
Gemini-2-Flash	21.58	3.01	10.71	3.82	7.89	2.49	5.84	8.56
GPT-4o-voice	57.76	5.79	13.64	5.66	10.83	3.48	7.97	15.76
Qwen2.5-Omni: Cascaded								
GER: ⇒ Whisper-v2-large	23.44	6.15	12.15	3.94	7.53	2.97	4.89	8.44
GER: ⇒ canary	24.58	6.38	12.43	4.02	7.72	3.05	5.01	8.74
GER: ⇒ parakeet	22.91	6.09	12.10	3.98	7.49	2.92	4.84	8.33
Qwen2.5-Omni: Parallel								
Speech-Hands ⇒ whisper-v2	15.03	4.45	12.37	3.01	6.49	1.86	3.46	6.67
Speech-Hands ⇒ canary	15.29	4.21	10.87	2.17	5.96	1.61	3.07	6.17
Speech-Hands ⇒ parakeet	11.20	4.37	11.10	2.26	6.02	1.67	3.18	5.69

Table 6.3: WER (%) results across 7 datasets, with the average WER shown in the rightmost column. Speech-Hands training significantly outperforms both baseline systems (ASR model, Qwen) and prior cascaded GER setups.

the action token and the subsequent prediction. Concretely, the model first predicts the action token and then continues decoding the target transcript or answer; both parts contribute to the same loss. This end-to-end objective encourages the model to internalize the mapping from multimodal evidence to action choice and to generate the corresponding output under that decision.

6.3.5 Agentic Inference via Action Tokens

At inference time, the model receives the same multimodal inputs as during training. The model performs a two-stage decoding process: it first emits an action token: `<internal>`, `<external>`, or `<rewrite>`. The decoding process is subject to decide which information source to prioritize, and then generates the final output accordingly. This explicit agentic self-reflection mechanism offers interpretability and control over how the model balances internal perception and external knowledge. It enables direct analysis (*e.g.*, F1 score) of when the model relies on its own understanding, consultants from external systems, or synthesizes a new response. The unified prediction format ensures that the model not only learns what to gen-

erate, but also which to trust across both speech recognition and audio reasoning tasks.

6.4 Experimental Setup

6.4.1 Datasets

Speech Recognition. We use seven representative datasets covering a range of domains, styles, and noise conditions from OpenASR leaderboard: AMI [20] (meeting speech), Tedlium [55] (TED talks), GigaSpeech [24] (large-scale podcasts and YouTube-style speech), SPGISpeech [130] (long-form read speech), VoxPopuli [181] (English subset of multilingual political recordings), and LibriSpeech [131] (clean and noisy audiobook speech). For baseline fine-tuning and prompt-based GER, we use all available training sets. For our proposed method, unless otherwise specified (**w/ FULL Datasets**), we train on at most 20,000 examples per dataset, this is due to the limitation of heavy computation requirements when doing inference on the whole training set for internal, external and GER (token distribution is discussed in 6.7.1).

Dataset Details for ASR To ensure a fair and balanced evaluation across diverse speech corpora, we uniformly sampled up to 20,000 audio-question pairs from each dataset for training as in Table 6.4. All datasets were aligned to a consistent prompt-question-answer format to support unified multi-dataset training.

Dataset	AMI	Tedlium	gigaspeech	spgispeech	Voxpopuli	Libri	Libri-clean	Libri-other
Sampling #	Subset in Speech-Hands Training							
Train	20,000	20,000	9,389	20,000	20,000	20,000	20,000	20,000

Table 6.4: Number of training samples used from each dataset. For GigaSpeech, only 9,389 valid samples met our filtering criteria.

Audio Reasoning. We evaluate on the multi-domain audio question-answering benchmark [198] (MD-Audio), which consists of multiple-choice questions grounded in real audio clips. This benchmark includes three complementary

subsets that probe different reasoning capabilities. While previous audio reasoning benchmarks such as MMAU [154] and MMAR [110] provide only MMLU-style test sets, MD-Audio releases both training and development sets, enabling evaluation of the proposed trainable agentic framework, as shown in Table D.1. Detailed dataset descriptions can be found in Appendix 6.4.1. For each sample, we construct inputs as described in Section 6.3, including the original audio, a first-pass internal prediction from Omni model, one external hypothesis and GER result (for ASR).

6.5 Prompt Template

Below is the prompt template for ASR:

You are an omni-agent for speech understanding with access to three inputs:

- (1) The original audio;
- (2) Five transcription hypotheses from another ASR system (external);
- (3) Your own first-pass transcription (internal).

Your task is to:

- First decide whether your internal transcription is reliable.
- If yes, output <internal> and your transcription.
- If the external system is more reliable, output <external> and use one of its hypotheses.
- Otherwise, output <rewrite> and generate a new answer using both sources and the audio.

Also, here is the prompt for AudioQA, we explicitly prompt the model to output <external> only when the internal prediction is wrong while external perception is correct.

You are an audio understanding model with access to three inputs:

- (1) The original audio;
- (2) One answer candidate generated by another model (external);

Model / Setting	Bio-acoustic	Soundscape	Complex QA	avg. Acc. $\uparrow$
Audio LM or Omni Model				
Gemini-2-Flash	42.03	46.34	59.89	56.61
Qwen2.5-Omni	47.32	56.32	59.89	57.87
AudioFlamingo 3 (AF3)	71.88	57.31	81.26	74.49
Qwen2.5-Omni Baselines				
+ SFT with official training data	78.13	34.65	76.61	63.13
+ GRPO with official training data	78.09	39.43	79.12	65.54
+ GRPO with external audio data [95]	62.32	72.10	82.15	75.10
GER: $\Rightarrow$ AF3 (cascaded agentic)	76.29	52.02	77.48	68.93
Qwen2.5-Omni: Speech-Hands				
$\Rightarrow$ (parallel agentic): SFT with official training data	67.86	58.29	83.34	75.75
$\Rightarrow$ (parallel agentic) + majority sampling	81.25	59.4	85.7	77.37

Table 6.5: AudioQA and acoustic content reasoning accuracy (%) across knowledge-intensive bioacoustic QA [158], multi-sound-object soundscapes, and MMAU-style [154] complex audio QA tasks.

(3) Your own prediction (internal).

Your task is to decide which of the following strategies to apply:

- If your internal prediction is correct and acceptable, output `<internal>` and repeat your answer.
- If the external candidate is correct while your internal prediction is incorrect, output `<external>` and use the external answer.
- If all given answers are incorrect, output `<rewrite>` and re-answer the question correctly based only on the original audio.

Return the selected token (`<internal>/<external>/<rewrite>`) followed by your final answer.

6.6 Results

6.6.1 Training Details

All experiments are conducted using the Qwen2.5-Omni model. We extend its tokenizer to include three special action tokens, `<internal>`, `<external>` and

`<rewrite>`, used during both training and inference. Models are trained using the standard supervised fine-tuning (SFT) objective with a cross-entropy loss. We train for 5 epochs with fp16. The batch size is set to 64, and the learning rate is initialized at $1e-4$ with cosine decay. All experiments adopt greedy decoding.

6.6.2 Baselines

External ASR models: We include three high-performing supervised speech recognition models as external references: Whisper-v2-large [139], Canary-1B-v2, and Parakeet-TDT-0.6B-v3 [159]. These systems operate in a closed transcription setting and provide non-generative references.

External audio QA model: For multi-choice audio QA, we include Audio Flamingo 3 [45], a speech-language model with audio frontend capabilities serving as a strong baseline.

We also include the latest model in comparison, e.g., Phi-4-MM [120], Gemini-2-Flash, GPT-4o-voice in ASR evaluations.

6.6.3 ASR Results

Table 6.3 presents WER results across seven diverse datasets, and we find that: (1) Speech-Hands outperforms all baselines. Furthermore, our approach with strong ASR models (canary and parakeet) achieves the lowest WER, even with only 20k training examples, outperforming both ASR models or Omni-LLMs; (2) The prompt GER over Whisper lags significantly behind token-based methods. This underscores the importance of explicit control via action tokens rather than relying solely on natural-language prompts; (3) While pre-trained models like Whisper and Qwen perform well on curated datasets such as LibriSpeech-clean, their performance degrades significantly on conversational benchmarks like AMI. Notably, despite Qwen’s relatively weaker base ASR performance, our framework enhances its generalization to the extent that it surpasses stronger baselines such as Phi-4-MM on the average performance, demonstrating the stability and transferability of Speech-Hands on both clean and noisy datasets.

Dataset	AMI	Tedlium	Gigaspeech	SPGIspeech	Voxpopuli	Libri-clean	Libri-other
Training Distribution							
<internal>	67.95%	86.48%	87.76%	96.25%	93.73%	98.96%	98.96%
<external>	31.01%	11.18%	11.8%	3.64%	6.09%	0.96%	0.96%
<rewrite>	1.04%	2.34%	0.44%	1.21%	0.18%	0.1%	0.1%
Test Distribution							
<internal>	70.28%	83.57%	85.94%	95.42%	92.68%	98.92%	98.75%
<external>	26.91%	15.12%	12.41%	3.81%	6.47%	0.98%	1.08%
<rewrite>	2.27%	1.31%	1.65%	0.77%	0.85%	0.1%	0.17%
<internal> on Test							
Precision	0.85	0.63	0.77	0.89	0.85	0.88	0.83
Recall	0.78	0.72	0.90	0.94	0.90	0.99	0.99
F1	0.81	0.67	0.83	0.91	0.87	0.94	0.9
<external> on Test							
Precision	0.88	0.96	0.83	0.82	0.71	0.81	0.72
Recall	0.89	0.81	0.78	0.76	0.60	0.73	0.75
F1	0.89	0.88	0.80	0.79	0.65	0.77	0.74
<rewrite> on Test							
Precision	0.62	0.33	0.32	0.52	0.50	0.0	0.0
Recall	0.24	0.21	0.05	0.36	0.21	0.0	0.0
F1	0.39	0.28	0.08	0.43	0.33	0.0	0.0

Table 6.6: Training distribution and test-time F1 scores for Speech-Hands’ action tokens.

6.6.4 AudioQA Results

Table 6.5 reports accuracy on each sub-task and overall average accuracy: (1) Our final setup (Speech-Hands + majority sampling) achieves the highest average accuracy (77.37%), outperforming all baselines and pre-trained models. It performs particularly well on Complex QA (85.70%) and Bioacoustics QA (81.25%), indicating its ability to handle both abstract reasoning and fine-grained audio patterns; (2) Standard supervised fine-tuning (SFT) and prompt-based GER exhibit mixed results. SFT achieves good accuracy on Bioacoustics (78.13%) but fails on Soundscapes (34.65%). Prompt-based GER also fails on both Soundscape and Complex QA compared with flamingo 3 baseline. These results highlight the robustness of our agentic framework in diverse audio reasoning settings.

6.7 Additional Analysis

6.7.1 Accuracy of Action Token Prediction

We analyze the model’s ability to correctly emit the three action tokens of `<internal>`, `<external>`, and `<rewrite>`, in the ASR setting under the Action Tokens + Whisper configuration. These tokens interpret whether the model is making correct decisions to guide its final prediction.

Table 6.6 shows both the training distribution and the test-time precision, recall, and F1 scores. The distribution highlights a strong internal bias: across all datasets, `<internal>` dominates, exceeding 95% in Libri-clean, Libri-other, spgispeech, and Voxpopuli. In contrast, `<external>` is sparsely supervised, where below 1% in Librispeech and `<rewrite>` is extremely rare everywhere, often less than 2%. This imbalance poses a natural challenge for learning reliable action.

Despite this skew, the model demonstrates robust performance for the two tokens. On test data, `<internal>` predictions achieve F1 scores above 0.8 on most datasets (0.91 on spgispeech, 0.94 on Libri-clean, 0.90 on Libri-other), indicating that the model can reliably recognize when its own decoding is sufficient. Even for the much rarer `<external>` token, the model attains high F1 scores, showing strong generalization of deferring to external hypotheses despite limited supervision.

The `<rewrite>` token proves to be the most challenging, with F1 scores below 0.4 in all but one dataset and zero in Librispeech, where positive training examples are extremely rare. A closer examination reveals that precision consistently exceeds recall, indicating that when the model does emit `<rewrite>`, its decision is generally correct but under-triggered in the omni model. This suggests a cautious yet reasonably reliable rewrite detector, whose coverage could be further improved through targeted data augmentation.

Overall, these results validate the effectiveness of the proposed agentic action: even under heavy class imbalance, the model learns to accurately identify when to trust its own predictions versus when to consult external information. The main bottleneck remains the `<rewrite>` case, suggesting that richer sampling or augmentation strategies may be needed to stabilize this decision in future work.

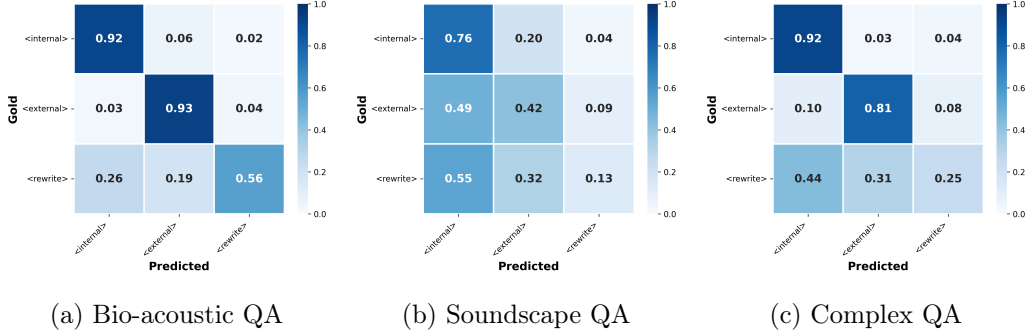


Figure 6.5: Confusion matrices of Speech-Hands’ agentic action execution for audio QA and reasoning three subsets on (a) bio-acoustic QA, (b) temporal and sound event QA, and (c) complex audio information QA.

6.7.2 Confusion Analysis In AudioQA

Subset	<in>/<ex>/<re>
Bio-acoustic QA	106/75/43
Soundscape QA	343/185/81
Complex QA	978/555/100

Table 6.7: Oracle Statistics of the three DCASE2025 AudioQA test sets used in our experiments.

We analyze the confusion matrix over the three subsets under the w/ multiple sampling setup (the oracle token distribution is shown in Table 6.7). As shown in Figure 6.5, the confusion between <internal> and <external> remains relatively low, suggesting that the model can effectively distinguish between them, especially in the Complex QA subset. However, the Soundscape subset shows slightly increased overlap, possibly due to the difficulty of Soundscape compared with other parts, also their original performances (Qwen 56.32 v.s. Flamingo 57.31) are quite closed, leading to much smaller sets of <external>. Besides, the highest confusion still lies with the <rewrite> class. As summarized in Table D.1, the number of training tokens labeled as <rewrite> is significantly smaller across all subsets. Such sparsity likely limits the model’s tendency to generalize the rewriting be-

havior during generation, but when it generates `<rewrite>`, the accuracy is quite high and robust as shown in the Figure. These findings highlight that even with imbalanced actions in training, the current F1 scores can still reliably improve the final performance, emphasizing the effectiveness of action tokens.

AudioQA Inference-Time Case Study

Q. Based on the audio, which natural phenomenon could be occurring?

- A. Earthquake B. Thunderstorm
C. Forest fire D. Snowstorm

Internal pred: B. Thunderstorm

External pred: B. Thunderstorm

Pred: `<rewrite>` C. Forest fire (✓)

ASR Inference-Time Case Study

Internal: you in the way marguerite but how

External: you ll in the way marguerite but how

Rewrite: you are in the way marguerite but how

Pred: `<internal>` you in the way marguerite but how (✓)

6.7.3 Category Analysis and Case Study

To provide systematic insight beyond individual cases, during our preliminary, we categorize three primary failure modes observed in ASR when Speech-Hands is absent. These categories further motivate our design choice of action tokens: (1) **External-induced Misguidance:** the internal omni-model is correct, but the model is misled by an error-heavy external hypothesis. This is the most damaging failure mode and directly motivates the action-token arbitration design; (2) **Over-correction:** LLMs tend to “complete” disfluent or partially heard speech, introducing hallucinated insertions or semantic expansions; (3) **Undercorrection in Dual-Failure:** when both internal and external predictions are wrong, the

omni model often selects one erroneous hypothesis instead of generating a rewrite, due to the lack of an explicit rewrite mechanism. These observations align with the high-precision behavior of the `<rewrite>` token and explain the improvements: the model selectively activates `<rewrite>` only when necessary.

We present illustrative examples (more in Appendix D.3) to show how the model makes selective use of internal, external, and rewrite hypotheses across modalities. In the AUDIOQA CASE), both the internal and external models predict “B. Thunderstorm”, likely influenced by surface acoustic features such as low-frequency rumbling. The rewrite path, however, generates “C. Forest fire”, which aligns with the ground truth, demonstrating the strong influence of `<rewrite>` that even two models have the same first-pass prediction.

In contrast, the ASR CASE reveals a different decision dynamic. Although the rewrite produces a more fluent variant, the model opts to retain the baseline hypothesis, judging the original phrase as sufficiently accurate. This indicates the model’s ability to avoid overcorrection issue in prior GER researches when input ambiguity is low. These examples underscore the benefits of explicit action tokens: the model can either rely on internal or external model predictions or revise them when necessary, yielding both flexibility and robustness in audio tasks.

6.8 Limitations and Discussions

Despite promising results, our study presents several limitations that offer avenues for future exploration.

Token imbalance and rewrite sparsity. Our training data exhibits an inherent imbalance across action tokens (`<internal>`, `<external>`, `<rewrite>`). While both `<internal>` and `<external>` achieve high F1 scores, `<rewrite>` remains under-trained on many datasets. This sparsity partly reflects that certain audio QA datasets rarely require rewriting but this contextual information sparsity also reveals a modeling challenge. Future work may explore more principled strategies for balancing token distribution or adaptively reshaping decision boundaries, especially under varying persona settings or task configurations.

Limited ASR training subset. Our current ASR experiments are trained on a restricted subset of data. While the model already achieves strong performance, it likely underutilizes the available signal. Scaling up training with larger ASR datasets or augmenting with synthetic audio variants may unlock further gains.

No exploration of transfer or multi-external setups. We do not yet study transfer capabilities. For example, training with one external ASR model and testing with another. Moreover, our current system only accepts a single external prediction. Extending the framework to handle multiple external models, each represented by distinct decision tokens, could significantly improve robustness and enable broader deployment across diverse real-world pipeline.

6.9 Summary of This Chapter

In this work, we proposed a learnable voice-agentic framework Speech-Hands for teaching omni models when to trust itself versus when to consult external audio perception. By casting the problem with explicit `<internal>`, `<external>`, and `<rewrite>` action tokens, our experimental results across AudioQA and ASR benchmarks demonstrate strong performance improvements beyond strong baselines, especially when direct finetuning and GER training fail, Speech-Hands can still robustly generate the best prediction.

This framework also benefits the interpretability in analysis, the model achieves high F1 scores for both `<internal>` and `<external>` tokens, even under imbalanced training conditions. While the `<rewrite>` token is rarer, its precision notably exceeds recall, indicating that the model can accurately identify necessary rewrites when it does trigger them. Overall, our method offers an effective framework to inject explicit actions into agent decision, toward reliable audio intelligence.

Relationship to the Dissertation: The Climax of Autonomous Agentic Scaffolding This final chapter, centered on Speech-Hands, represents the ultimate evolution of the Inference-time Scaffolding hierarchy. While previous chapters established the foundations of context utilization and dynamic retrieval,

this work addresses the most critical challenge in agentic systems: the autonomous integration and reconciliation of heterogeneous model outputs.

The core contribution of this chapter lies in its exploration of how individual agents adapt within a collaborative collective. In an agentic scaffold, a model must not only process inputs from other agents but also strategically choose between external context and its own internal knowledge. Failure to manage this choice leads to error integration, where cascaded inaccuracies degrade the entire system's performance. By introducing special action tokens, we provide a minimalist yet highly efficient optimization for this coordination, enabling agents to autonomously navigate trust and resolve perceptual conflicts during the reasoning process.

Ultimately, this chapter fulfills the dissertation's primary vision: defining Intelligence as a System Property. It demonstrates that through self-managed scaffolding and recursive reflection, a group of coordinated agents can achieve a level of synergistic scaling that far exceeds the capabilities of any monolithic architecture. This marks the transition from models that "solve tasks" to systems that "manage their own intelligence".

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This dissertation investigated how to improve the reliability and generality of LLMs *without* parameter updates, by treating inference as a first-class design space. Across text and multimodal settings, this dissertation demonstrates that many practical failures of LLMs are not solely due to a deficiency in their pre-trained knowledge, but rather due to how that knowledge is elicited through inference-time execution. Under this lens, the dissertation presented a series of inference-time scaffolding strategies centered around two recurring behaviors: (i) effective context utilization and (ii) reasoning under constraints. These behaviors enable latent capabilities to be expressed more robustly through carefully designed inference processes, rather than through costly retraining.

Chapters 2 and 3 defined the foundational archetype of this scaffolding through ICL. Although ICL is often effective on broad, loosely constrained tasks, it remains less reliable for classic structured prediction problems that historically benefited from BERT-era fine-tuning. Focusing on relation extraction, I showed that the main bottlenecks of vanilla ICL are not merely the number of demonstrations, but their semantic alignment with the query instance and the absence of explicit, label-grounded logic. To close this gap at inference time, I proposed GPT-RE, which augments ICL with task-aware demonstration retrieval and gold-label-induced reasoning. By constructing demonstrations that are both relevant and logically in-

formative, GPT-RE substantially narrows the performance gap between ICL and fully supervised baselines while improving interpretability through label-aligned rationales.

Building on the insights from structured text, the dissertation then generalized the same inference-time principles to multimodal tasks. I presented GPT-MM as a unified in-context learning framework for text, vision, and audio, where the core idea remains unchanged: the system should retrieve task-relevant demonstrations and enrich them with label-induced reasoning so that inference becomes both more accurate and more interpretable. The resulting framework consistently improves few-shot performance over strong prompting baselines and demonstrates that task-aware retrieval and label-grounded reasoning can serve as modality-agnostic mechanisms. Taken together, GPT-RE and GPT-MM support a unified view that “learning at inference time” becomes reliable when demonstrations are treated as structured, task-aware evidence rather than as arbitrary examples. More importantly, the **gold-label-induced reasoning** framework introduced in these chapters has provided a methodological legacy for constructing the high-quality reasoning traces now widely used in model alignment and training.

Chapter 4 established RAG as the **dynamic link between frozen models and an evolving world**, and focused on inference-time domain adaptation under distribution shift, with an emphasis on faithfulness in high-stakes settings. In domain-specific applications such as law, hallucinations and ungrounded claims are particularly problematic, and naive retrieval augmentation can remain brittle when retrieved evidence is incomplete, noisy, or inconsistently applied. I introduced Adapt-Retrieve-Revise, which reformulates domain adaptation as an iterative inference procedure. Instead of generating a single answer from a fixed context, the framework produces a draft, retrieves evidence based on the draft, and revises the answer using the retrieved provisions. This iterative structure provides a practical mechanism to ground generation in verifiable sources and to improve reliability through explicit correction at inference time, while keeping the base model parameters unchanged.

Chapters 5 and 6 **marked the transition from individual reasoning layers to the architecture of collective intelligence**, and I addressed multi-

modal multi-agent pipelines, where intermediate artifacts (e.g., ASR transcripts) are increasingly consumed by other agents/LLMs rather than humans. This shift challenges conventional evaluation protocols and motivates an agent-centric notion of understanding. I proposed SpeechIQ, a cognition-inspired evaluation pipeline that measures voice-model capabilities across cognitive levels and provides a consistent way to compare systems across architectural choices, including cascaded ASR→LLM pipelines and end-to-end speech LLMs. By moving beyond text-only transcription metrics, SpeechIQ offers a principled benchmark to quantify whether intermediate outputs support downstream understanding and decision-making in agent pipelines.

Finally, motivated by the observation that multimodal hypotheses are often noisy and can derail downstream reasoning, I presented Speech-Hands, a self-reflection voice-agentic framework that learns when to trust internal inference versus when to consult external audio perception. By casting self-reflection as a learnable decision primitive and by unifying perception and decision-making within an agentic inference procedure, Speech-Hands improves robustness on speech recognition as well as on audio reasoning tasks. This completes the dissertation’s overarching narrative: inference-time scaffolding is not only about injecting more context, but also about structuring decision processes so that agents can remain reliable under uncertainty.

In summary, the dissertation establishes inference-time design as a practical and general pathway toward more robust LLM systems. Across structured text understanding, multimodal ICL, domain adaptation, and voice agent pipelines, the results collectively show that retrieval constructs useful context, reasoning enforces task-consistent logic, and agentic self-reflection prevents failures under noisy hypotheses. These components form a coherent methodology for advancing reliable and unified understanding across modalities without parameter updates.

7.2 Future Work

While this dissertation demonstrates the effectiveness of inference-time scaffolding, it also exposes several open directions that can further strengthen the reliability

and generality of LLM-based systems.

The first direction is **inference-time training** as an extension of inference-time design. The methods studied here deliberately avoid parameter updates, which provides strong practicality and deployment stability, but also limits how much a system can adapt to persistent shifts. An attractive next step is to combine the inference-time scaffold with lightweight, controlled adaptation at test time, where updates are small, reversible, and guided by reliability signals. For example, agent-centric evaluation outcomes can be used as training signals to adapt components of the system (e.g., retrieval policies, confidence calibration modules, or compact adapters) only when the system detects systematic failures. The key research question is how to introduce inference-time updates that improve robustness without sacrificing safety, reproducibility, and cross-domain generalization.

The second major trajectory for future research involves the evolution of inference-time design into **autonomous agentic systems**. Building on the insight that many modern AI agents operate by utilizing the same underlying model with varying prompts, this dissertation posits that agentic behavior is essentially a dynamic and sophisticated form of inference-time scaffolding. In this paradigm, the role of LLMs matures from passive generators into proactive cognitive hubs that autonomously orchestrate their own execution strategies. Future research will explore the development of self-managed scaffolding, where agents dynamically decide how to structure their inference process, such as choosing between different reasoning paths or context utilization strategies—based on real-time task demands and environmental feedback. This transition will effectively bridge the gap between human-designed prompts and truly autonomous machine intelligence.

Ultimately, these directions point toward a transformative vision of Artificial General Intelligence (AGI). Rather than a single monolithic model, the future of AGI likely resides in a dynamic agentic collective that a decentralized group of specialized agents coordinated through sophisticated inference-time scaffolding. In this paradigm, intelligence is no longer an isolated property of model parameters, but an emergent system-level property of the interactions, verifications, and reflections among these agents. Inference-time scaffolding serves as the fundamental architecture for this agentic future, providing the essential protocols for

self-managed learning and cross-modal coordination. By evolving from simple textual reasoners into an autonomous agentic group capable of native multimodal reasoning, we pave the way for a resilient and reflective intelligence that can truly navigate and act within the boundless uncertainties of the real world.

Bibliography

- [1] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.
- [2] S. Adavanne, A. Politis, and T. Virtanen. A multi-room reverberant dataset for sound event localization and detection, 2019.
- [3] A. Adediji, S. Joshi, and B. Doohan. The sound of healthcare: Improving medical transcription asr accuracy with large language models, 2024.
- [4] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
- [5] Baichuan-inc. Baichuan-7b. <https://github.com/baichuan-inc/Baichuan-7B>, 2023.
- [6] L. Baldini Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.
- [7] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In J. D. Moore, S. Teufel, J. Allan, and S. Furui, editors,

- Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio, June 2008. Association for Computational Linguistics.
- [8] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [9] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.
- [10] T. Blevins, H. Gonen, and L. Zettlemoyer. Prompting language models for linguistic structure. *CoRR*, abs/2211.07830, 2022.
- [11] B. Bloom and D. Krathwohl. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Number v. 1 in *Taxonomy of Educational Objectives: The Classification of Educational Goals*. D. McKay, 1969.
- [12] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022.
- [13] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al. Audioldm: a

- language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.
- [14] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.
- [15] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.
- [16] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- [17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Informa-*

- tion Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.*
- [19] A. Calcus. Development of auditory scene analysis: a mini-review. *Frontiers in Human Neuroscience*, 18:1352247, Mar 12 2024.
- [20] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.
- [21] D. Chan, A. Myers, S. Vijayanarasimhan, D. Ross, and J. Canny. IC3: Image captioning by committee consensus. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, Singapore, Dec. 2023. Association for Computational Linguistics.
- [22] A. Chen, F. Yao, X. Zhao, Y. Zhang, C. Sun, Y. Liu, and W. Shen. EQUALS: A real-world dataset for legal question answering via reading chinese laws. In M. Grabmair, F. Andrade, and P. Novais, editors, *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023*, pages 71–80. ACM, 2023.
- [23] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E. S. Chng. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [24] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, interspeech_2021. ISCA, Aug. 2021.
- [25] Y. Chen, S. Ji, H. Wang, Z. Wang, S. Chen, J. He, J. Xu, and Z. Zhao. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models, 2025.

- [26] Y.-P. Chen, R. Price, and S. Bangalore. Spoken language understanding without speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE, 2018.
- [27] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada, J. Li, S. Ghosh, J. Balam, and B. Ginsburg. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE, 2024.
- [28] K. Cheng, Y. Li, F. Xu, J. Zhang, H. Zhou, and Y. Liu. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*, 2024.
- [29] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.
- [30] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus,

- D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.
- [31] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-audio technical report, 2024.
- [32] A. D. N. Cohen, S. Rosenman, and Y. Goldberg. Relation extraction as two-way span-prediction. *CoRR*, abs/2010.04829, 2020.
- [33] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases, 2023.
- [34] Y. Cui, Z. Yang, and X. Yao. Efficient and effective text encoding for chinese llama and alpaca, 2023.
- [35] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- [36] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

- 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [38] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik. Leveraging large language models for exploiting asr uncertainty. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12231–12235. IEEE, 2024.
 - [39] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. GLM: General language model pretraining with autoregressive blank infilling. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May 2022. Association for Computational Linguistics.
 - [40] Y. Fang, B. Cheng, J. Peng, X. Li, Y. Xi, C. Zhang, and G. Zhong. Fewer hallucinations, more verification: A three-stage llm-based framework for asr error correction. *arXiv preprint arXiv:2505.24347*, 2025.
 - [41] R. Frieske and B. E. Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models, 2024.
 - [42] B. Galantucci, C. A. Fowler, and M. T. Turvey. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3):361–377, jun 2006. Erratum in: *Psychon Bull Rev.* 2006 Aug;13(4):742.
 - [43] T. Gao, X. Yao, and D. Chen. SimCSE: Simple contrastive learning of sentence embeddings. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
 - [44] X. Geng, X. Chen, K. Q. Zhu, L. Shen, and Y. Zhao. MICK: A meta-learning framework for few-shot relation classification with small training data. In M. d’Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, editors, *CIKM ’20: The 29th ACM International Conference on Information*

- and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 415–424. ACM, 2020.
- [45] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S. gil Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025.
- [46] A. Goel, K. Sapra, M. Le, R. Valle, A. Tao, and B. Catanzaro. Omcat: Omni context aware transformer. *arXiv preprint arXiv:2410.12109*, 2024.
- [47] R. Gonzales and F. Rudzicz. A retrieval augmented approach for text-to-music generation. In A. Kruspe, S. Oramas, E. V. Epure, M. Sordo, B. Weck, S. Doh, M. Won, I. Manco, and G. Meseguer-Brocal, editors, *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 31–36, Oakland, USA, Nov. 2024. Association for Computational Linguistics.
- [48] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, 2020.
- [49] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.
- [50] T. Hai. Lexilaw. <https://github.com/CSHaitao/LexiLaw>, 2023.
- [51] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.
- [52] X. He, L. Deng, and A. Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635. IEEE, 2011.

- [53] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In K. Erk and C. Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [54] H. Hermansky and J. Junqua. Optimization of perceptually-based asr front-end (automatic speech recognition). In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 219–222 vol.1, 1988.
- [55] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing, 2018.
- [56] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [57] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022.
- [58] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022.
- [59] B. Hoffmeister, A. Rastrow, and B. Liu. Language model speech endpointing, Dec. 29 2016. US Patent App. 14/753,811.
- [60] T. Hori, M. Kocour, A. Haider, E. McDermott, and X. Zhuang. Delayed

fusion: Integrating large language models into first-pass decoding in end-to-end speech recognition. *arXiv preprint arXiv:2501.09258*, 2025.

- [61] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [62] R. Hu, D. Qiu, S. Wei, J. Zhang, Y. Wang, S. Liu, and J. Sang. Investigating and enhancing vision-audio capability in omnimodal large language models. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7452–7463, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [63] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran, L. Liu, and F. Wei. WavLLM: Towards robust and adaptive speech large language model. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
- [64] C.-y. Huang, K.-H. Lu, S.-H. Wang, C.-Y. Hsiao, C.-Y. Kuan, H. Wu, S. Arora, K.-W. Chang, J. Shi, Y. Peng, et al. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE, 2024.
- [65] Q. Huang, M. Tao, Z. An, C. Zhang, C. Jiang, Z. Chen, Z. Wu, and Y. Feng. Lawyer llama technical report, 2023.
- [66] R. Huang, O. Abdel-hamid, X. Li, and G. Evermann. Class lm and word mapping for contextual biasing in end-to-end asr. In *Interspeech 2020*, pages 4348–4351, 2020.

- [67] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, 2023.
- [68] M. J. Hunt. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336, 1990.
- [69] IDEA-CCNL. Fengshenbang-lm. <https://github.com/pengxiao-song/LaWGPT>, 2023.
- [70] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. Unsupervised dense information retrieval with contrastive learning, 2022.
- [71] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022.
- [72] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A dynamic language model for speech recognition. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.
- [73] T. Jiang, S. Huang, Z. Luan, D. Wang, and F. Zhuang. Scaling sentence embeddings with large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
- [74] B. Jimenez Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

- [75] John and J. Raven. *Raven Progressive Matrices*, pages 223–237. Springer US, Boston, MA, 2003.
- [76] A. Juffs and M. Harrington. Garden path sentences and error data in second language sentence processing. *Language learning*, 46(2):283–323, 1996.
- [77] D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchaman, and N. Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 189–192. IEEE, 1995.
- [78] M. Kaiser, D. Senkowski, and J. Keil. Mediofrontal theta-band oscillations reflect top-down influence in the ventriloquist illusion. *Human Brain Mapping*, 42(2):452–466, feb 2021. Epub 2020 Oct 14.
- [79] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.
- [80] T. Kawahara. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*, pages 65–75. Springer, 2019.
- [81] S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. In *Interspeech 2021*, pages 1977–1981, 2021.
- [82] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. Seltzer. Evaluating user perception of speech recognition system quality with semantic distance metric. In *Interspeech 2022*, pages 3978–3982, 2022.
- [83] Y. Kim. Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

- [84] D. Klakow and J. Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002.
- [85] A. Koenecke, A. S. G. Choi, K. Mei, H. Schellmann, and M. Sloane. Careless whisper: Speech-to-text hallucination harms. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.
- [86] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval, 2020.
- [87] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916, 2022.
- [88] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25125–25148. PMLR, 21–27 Jul 2024.
- [89] H.-K. J. Kuo, Z. Tüske, S. Thomas, Y. Huang, K. Audhkhasi, B. Kingsbury, G. Kurata, Z. Kons, R. Hoory, and L. Lastras. End-to-end spoken language understanding without full transcripts. In *Interspeech 2020*, pages 906–910, 2020.
- [90] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.
- [91] C. Lebiere and J. R. Anderson. Cognitive Constraints on Decision Making under Uncertainty. *Frontiers in Psychology*, 2:305, nov 2011.
- [92] L. Lee. Book reviews: Foundations of statistical natural language processing. *Computational Linguistics*, 26(2), 2000.

- [93] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.
- [94] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [95] G. Li, J. Liu, H. Dinkel, Y. Niu, X. Sun, T. Wang, J. Zhang, and J. Luan. Miaqa submission for dcase 2025 challenge task 5: A reinforcement learning driven audio question answering method. Technical report, DCASE2025 Challenge, June 2025.
- [96] Y. Li, J. Liu, T. Zhang, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, L. Ming, G. Dong, D. Pan, C. Li, et al. Baichuan-omni-1.5 technical report, 2025.
- [97] F. Liu, H. Lin, X. Han, B. Cao, and L. Sun. Pre-training to match for unified low-shot relation extraction. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [98] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for GPT-3? In E. Agirre, M. Apidianaki, and I. Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learn-*

- ing Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.
- [99] X. Liu, K. Ji, Y. Fu, Z. Du, Z. Yang, and J. Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602, 2021.
 - [100] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pre-training approach, 2019.
 - [101] Z. Liu, S. Kim, and O. Kalinli. Evaluating speech recognition performance towards large language model based voice assistants. In *Interspeech 2024*, pages 4099–4103, 2024.
 - [102] G. López, L. Quesada, and L. A. Guerrero. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pages 241–250. Springer, 2018.
 - [103] K.-H. Lu, Z. Chen, S.-W. Fu, C.-H. H. Yang, J. Balam, B. Ginsburg, Y.-C. F. Wang, and H. yi Lee. Desta2: Developing instruction-following speech language model without speech instruction-tuning data, 2025.
 - [104] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.
 - [105] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Pro-*

- ceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.
- [106] lyogavin. anima. <https://github.com/lyogavin/Anima>, 2023.
- [107] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.
- [108] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan. Query rewriting for retrieval-augmented large language models, 2023.
- [109] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma. Lecard: A legal case retrieval dataset for chinese law system. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2342–2348. ACM, 2021.
- [110] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025.
- [111] I. S. MacKenzie and R. W. Soukoreff. A character-level error analysis technique for evaluating text entry methods. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 243–246, 2002.
- [112] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- [113] A. Mahmood, J. Wang, B. Yao, D. Wang, and C.-M. Huang. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies*, 195:103406, Jan. 2025.

- [114] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe. VoxTlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE, 2024.
- [115] N. Malkin, Z. Wang, and N. Jojic. Coherence boosting: When your pre-trained language model is not paying enough attention. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [116] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019.
- [117] C. Marinoni, R. F. Gramaccioni, C. Chen, A. Uncini, and D. Comminiello. Overview of the l3das23 challenge on audio-visual extended reality, 2024.
- [118] Y. Matsuo, Y. LeCun, M. Sahani, D. Precup, D. Silver, M. Sugiyama, E. Uchibe, and J. Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.
- [119] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539, 2014.
- [120] Microsoft, :, A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, D. Chen, D. Chen, J. Chen, W. Chen, Y.-C. Chen, Y. ling Chen, Q. Dai, X. Dai, R. Fan, M. Gao, M. Gao, A. Garg, A. Goswami, J. Hao, A. Hendy, Y. Hu, X. Jin, M. Khademi, D. Kim, Y. J. Kim, G. Lee, J. Li, Y. Li, C. Liang, X. Lin, Z. Lin, M. Liu, Y. Liu, G. Lopez, C. Luo, P. Madan, V. Mazalov, A. Mitra, A. Mousavi, A. Nguyen, J. Pan, D. Perez-Becker, J. Platin, T. Portet,

- K. Qiu, B. Ren, L. Ren, S. Roy, N. Shang, Y. Shen, S. Singhal, S. Som, X. Song, T. Sych, P. Vaddamanu, S. Wang, Y. Wang, Z. Wang, H. Wu, H. Xu, W. Xu, Y. Yang, Z. Yang, D. Yu, I. Zabir, J. Zhang, L. L. Zhang, Y. Zhang, and X. Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025.
- [121] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.
- [122] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *CoRR*, abs/2202.12837, 2022.
- [123] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022.
- [124] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, Mar. 2001.
- [125] T. O. Nelson. Metamemory: A Theoretical Framework and New Findings. In G. H. Bower, editor, *Psychology of Learning and Motivation*, volume 26, pages 125–173. Academic Press, 1990.
- [126] OpenAI. Gpt-4 technical report, 2023.
- [127] OpenAI. Gpt-embeddings. <https://platform.openai.com/docs/guides/embeddings/what-are-embeddings>, 2023.
- [128] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen,

R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O’Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wain-

- wright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report, 2024.
- [129] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, and the others. Gpt-4o system card, 2024.
- [130] P. K. O’Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman, B. Ginsburg, S. Watanabe, and G. Kucsko. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *Inter-speech 2021*, pages 1434–1438, 2021.
- [131] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.
- [132] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [133] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4, 2023.
- [134] E. Perez, D. Kiela, and K. Cho. True few-shot learning with language models. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070, 2021.

- [135] K. C. Puvvada, P. Żelasko, H. Huang, O. Hrinchuk, N. R. Koluguri, K. Dhawan, S. Majumdar, E. Rastorgueva, Z. Chen, V. Lavrukhin, J. Balam, and B. Ginsburg. Less is more: Accurate speech recognition & translation without web-scale data. In *Interspeech 2024*, pages 3964–3968, 2024.
- [136] Y. Qiu, H. Li, Y. Qu, Y. Chen, Q. She, J. Liu, H. Wu, and H. Wang. DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [137] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [138] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sasstry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [139] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [140] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [141] S. Radhakrishnan, C.-H. Yang, S. Khan, R. Kumar, N. Kiani, D. Gomez-Cabrero, and J. Tegnér. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016, 2023.
- [142] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. v. d. Driessche, L. A. Hendricks,

- M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. d. M. d’Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. d. L. Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training gopher, 2021.
- [143] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d’Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022.
- [144] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

- [145] J. C. Raven and J. H. Court. *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press Oxford, 1998.
- [146] R. Reddy. Foundations and grand challenges of artificial intelligence: Aaai presidential address. *AI magazine*, 9(4):9–9, 1988.
- [147] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, Nov. 2020. Association for Computational Linguistics.
- [148] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [149] M. Renze and E. Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.
- [150] M. D. Rio, P. Ha, Q. McNamara, C. Miller, and S. Chandra. Earnings-22: A practical benchmark for accents in the wild, 2022.
- [151] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.
- [152] O. Rubin, J. Herzig, and J. Berant. Learning to retrieve prompts for in-context learning. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics.

- [153] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Vissontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.-C. Chiu. Two-pass end-to-end speech recognition. In *Interspeech 2019*, pages 2773–2777, 2019.
- [154] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.
- [155] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024.
- [156] Z. Sasindran, H. Yelchuri, and T. V. Prabhakar. Semascore: A new evaluation metric for automatic speech recognition tasks. In *Interspeech 2024*, interspeech_2024, page 4558–4562. ISCA, Sept. 2024.
- [157] Z. Sasindran, H. Yelchuri, T. V. Prabhakar, and S. Rao. H_eval: A new hybrid evaluation metric for automatic speech recognition tasks, 2023.
- [158] L. Sayigh, M. A. Daher, J. Allen, H. Gordon, K. Joyce, C. Stuhlmann, and P. Tyack. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics*, volume 27, page 040013. Acoustical Society of America, 2016.
- [159] M. Sekoyan, N. R. Koluguri, N. Tadevosyan, P. Zelasko, T. Bartley, N. Karpov, J. Balam, and B. Ginsburg. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast, 2025.
- [160] R. L. Selman and D. F. Byrne. A structural-developmental analysis of levels of role taking in middle childhood. *Child development*, pages 803–806, 1974.
- [161] P. Shah, D. Hakkani-Tür, G. Tür, A. Rastogi, A. Bapna, N. Nayak, and L. Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.

- [162] R. Shin, C. Lin, S. Thomson, C. Chen, S. Roy, E. A. Platanios, A. Pauls, D. Klein, J. Eisner, and B. Van Durme. Constrained language models yield few-shot semantic parsers. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
- [163] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation, 2021.
- [164] P. Song. Lawgpt. <https://github.com/IDEA-CCNL/Fengshenbang-LM>, 2021.
- [165] Z. Song, X. Wang, Z. Qian, H. Chen, L. Huang, H. Xue, and W. Zhu. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*.
- [166] R. Sparks, L. Ganschow, and A. Thomas. Role of intelligence tests in speech/language referrals. *Perceptual and motor skills*, 83(1):195–204, 1996.
- [167] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks, 2014.
- [168] P. Szymański, L. Augustyniak, M. Morzy, A. Szymczak, K. Surdyk, and P. Żelasko. Why aren ’ t we ner yet? artifacts of asr errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761, 2023.
- [169] P. Szymański, P. Żelasko, M. Morzy, A. Szymczak, M. Żyła-Hoppe, J. Banaszczyk, L. Augustyniak, J. Mizgajski, and Y. Carmiel. Wer we are and wer we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, 2020.

- [170] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024.
- [171] Z. Tüske, G. Saon, and B. Kingsbury. On the limit of english conversational speech recognition. In *Interspeech 2021*, pages 2062–2066, 2021.
- [172] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.
- [173] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soroaker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. Lamda: Language models for dialog applications, 2022.
- [174] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soroaker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil,

- B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. Lamda: Language models for dialog applications, 2022.
- [175] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.
- [176] G. Tur, D. Hakkani-Tür, and R. E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, 2005.
- [177] D. Van Strien, K. Beelen, M. C. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza. Assessing the impact of ocr quality on downstream nlp tasks. 2020.
- [178] L. Velikovich, C. Li, D. Caseiro, S. Kumar, P. Rondon, K. Joshi, and X. Velez. Spelling correction through rewriting of non-autoregressive asr lattices. *arXiv preprint arXiv:2409.16469*, 2024.
- [179] Z. Wan, Q. Liu, Z. Mao, F. Cheng, S. Kurohashi, and J. Li. Rescue implicit and long-tail cases: Nearest neighbor relation extraction. *CoRR*, abs/2210.11800, 2022.
- [180] C. Wang, X. Liu, Z. Chen, H. Hong, J. Tang, and D. Song. DeepStruct: Pretraining of language models for structure prediction. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [181] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

- 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, Aug. 2021. Association for Computational Linguistics.
- [182] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022.
- [183] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.
- [184] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *CoRR*, abs/2203.11171, 2022.
- [185] Z. Wang, C.-K. Tang, and Y.-W. Tai. Audio-agent: Leveraging llms for audio generation, editing and composition, 2025.
- [186] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. Espnet: End-to-end speech processing toolkit, 2018.
- [187] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.
- [188] R. Whetten and C. Kennington. Evaluating and improving automatic speech recognition using severity. In D. Demner-fushman, S. Ananiadou, and K. Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 79–91, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [189] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and

- A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [190] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.
- [191] J. Wu, X. Fan, B.-R. Lu, X. Jiang, N. Mesgarani, M. Hasegawa-Johnson, and M. Ostendorf. Just asr+ llm? a study on speech large language models’ ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE, 2024.
- [192] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478, 2018.
- [193] Z. Xie and C. Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024.
- [194] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- [195] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin. Qwen2.5-omni technical report, 2025.
- [196] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi,

- T. He, X. Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.
- [197] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report, 2024.
- [198] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, O. Nieto, R. Duraiswami, D. Manocha, G. Kim, J. Du, R. Valle, and B. Catanzaro. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge, 2025.
- [199] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, page 1–8. IEEE, Dec. 2023.
- [200] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.
- [201] C.-H. H. Yang, T. Park, Y. Gong, Y. Li, Z. Chen, Y.-T. Lin, C. Chen, Y. Hu, K. Dhawan, P. Żelasko, et al. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 371–378. IEEE, 2024.
- [202] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, H. Guo, X. Chang, J. Shi, S. Zhao, J. Bian, Z. Zhao, X. Wu, and H. M. Meng. UniAudio: Towards

- universal audio generation with large language models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56422–56447. PMLR, 21–27 Jul 2024.
- [203] Q. Yang, S. Yao, W. Chen, S. Fu, D. Bai, J. Zhao, B. Sun, B. Yin, X. Wei, and J. Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.
- [204] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023.
- [205] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [206] S. Yue, W. Chen, S. Wang, B. Li, C. Shen, S. Liu, Y. Zhou, Y. Xiao, S. Yun, X. Huang, and Z. Wei. Disc-lawllm: Fine-tuning large language models for intelligent legal services, 2023.
- [207] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y. Jiang, and X. Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling, 2024.
- [208] D. Zhang, X. Zhang, J. Zhan, S. Li, Y. Zhou, and X. Qiu. Speechgpt-gen: Scaling chain-of-information speech generation, 2024.
- [209] L. Zhang, T. Zhao, H. Ying, Y. Ma, and K. Lee. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer, 2024.
- [210] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

- [211] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu. Speeche tokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [212] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. Position-aware attention and supervised data improve slot filling. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
- [213] J. Zhao, G. Haffari, and E. Shareghi. Generating synthetic speech from SpokenVocab for speech translation. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1975–1981, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [214] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021.
- [215] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press, 2020.
- [216] Z. Zhong and D. Chen. A frustratingly easy approach for entity and relation extraction. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 50–61, Online, June 2021. Association for Computational Linguistics.

- [217] L. Zhuang, L. Wayne, S. Ya, and Z. Jun. A robustly optimized BERT pre-training approach with post-training. In S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che, S. He, and G. Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, Aug. 2021. Chinese Information Processing Society of China.
- [218] V. W. Zue and J. R. Glass. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180, 2000.

Appendix A

Additional Experimental Details of GPT-RE

A.1 Hyperparameters

A.1.1 GPT-3 Hyperparameters

We use the GPT-3 API during the experiments and set the hyperparameters as in Table A.1. Since the “Temperature” is set to be 0.0, denoting the stable output of GPT-3, we report the result of the single run for all experiments. Due to the input length limitation of GPT-3 and the various average lengths of contexts from each dataset, we set different search ranges for the number of demonstrations of each dataset as shown in Table A.2.

A.1.2 Fine-tuning Baseline PURE

We follow their single-sentence setup to keep consistency among datasets as SemEval and TACRED are both sentence-level RE datasets. For the PLMs, we also follow PURE by using *scibert-scivocab-uncased* [8] as the base encoder for SciERC and *bert-base-uncased* [37] for the remaining three general domain datasets. We follow hyperparameters in their paper. We used 2 NVIDIA RTX3090 for training.

Hyperparameter	In Experiment
Engine	text-davinci-003
Temperature	0.0
Max_tokens	256
Top_p	1
Frequency_penalty	0.0
Presence_penalty	0.0
Best_of	1
Logprob	1

Table A.1: GPT-3 Hyperparamters.

Dataset	Lower bound	Upper bound
Semeval	5	30
TACRED	5	15
SciERC	5	30
ACE05	5	25

Table A.2: Search range for each dataset.

A.1.3 Sentence Embedding Methods

(author?) [74] uses the [CLS] of RoBERTa-large as the representation in retrieval, (author?) [98] fine-tunes RoBERTa-large on two natural language inference (NLI) datasets: SNLI [15] and MultiNLI [189] to enhance the quality of sentence embedding. For the sentence embedding method SimCSE in our experiment, we utilize the version: `sup-simcse-bert-base-uncased`.

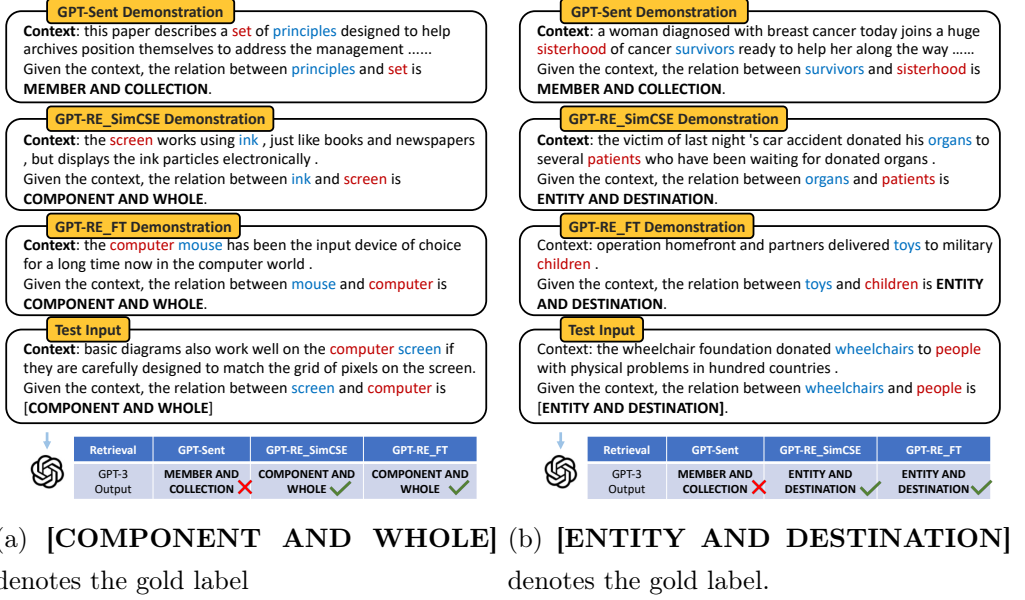


Figure A.1: More cases.

A.2 Case Study

To verify the effectiveness of our task-aware demonstration retrieval, we provide more cases.

For Figure A.1a, *GPT-Sent* retrieves a demonstration that shares the same semantic meaning of “design” with the test input. However, the entity pair is irrelevant to the concept “design” resulting in a noisy demonstration. Instead, *GPT-RE_SimCSE* retrieves a more relative demonstration with closer pair of entities sharing the same relation label. Furthermore, *GPT-RE_FT* retrieves the demonstration containing both the closing entity pair and the same linguistic structure between entities. This case emphasizes level-by-level improvement using our proposed methods. Figure A.1b shows a similar phenomenon.

Appendix B

Additional Experimental Details of Adapt-Retrieve-Revise

B.0.1 Legal Instruction Tuning

We build our legal instruction dataset by human experts. Due to privacy concerns, we are not allowed to disclose the annotated instruction data. However, we will release the instruction annotation guideline along with our 7B legal LLM. We show a template with a toy example below.

- **Due to the Article x in the law y :** [the corresponding content in the law]
- **Considering the fact that** [the fact]
- **The judgment is** [the conclusion]

A toy example could be:

- **Due to the article 91 of the Road Traffic Safety Law of the People’s Republic of China:** [Whoever drives a motor vehicle after drinking alcohol shall be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 1 month but not more than 3 months, and be imposed upon a fine of not less than 200 Yuan but not more than 500 Yuan as well; whoever drives a motor vehicle when he is drunk shall be restricted

by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 3 months but not more than 6 months, and be imposed upon a fine of not less than 500 Yuan but not more than 2000 Yuan as well. Whoever drives a commercial operating motor vehicle after drinking alcohol shall be imposed upon the penalty of temporary seizure of his motor vehicle driving license for 3 months, and be imposed upon a fine of 500 Yuan as well; whoever drives a commercial operating motor vehicle when he is drunk shall be restricted by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for 6 months, be imposed upon a fine of 2000 Yuan as well. Where anyone is penalized for twice or more within one year due to his drunken driving as prescribed in the preceding two paragraphs, his motor vehicle driving license shall be canceled, and he shall not drive a commercial operating motor vehicle within 5 years.]

- **Considering the fact that** [the man was riding a motorbike when drunk.]
- **The judgment is** [to be restricted by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 3 months but not more than 6 months, and be imposed upon a fine of not less than 500 Yuan but not more than 2000 Yuan as well.]

B.0.2 Retrieval Modules

we leveraged multilingual E5-large, which is the SOTA family of text embeddings, which has been reported to outperform BM25, Contriever [70], and GPT embeddings [127]. Since the improvements of our method are consistent and substantial (+33.3% vs vanilla, +15.4%/23.9% vs retrieval baselines), we believe these gaps have shown sufficient robustness in our proposal.

However, we agree that including more established retrieval modules can enhance the robustness of our findings. Therefore, we added extra experiments and compared them with the current SOTA Chinese retrieval module CoROM following [136] on the LegalQA dataset, the same setting as in Section 4.5.1. Table B.1 shows the comparison results. We find that: (1) multilingual e5-large has a competitive performance with CoROM on query-based retrieval and vastly outperforms CoROM on answer-based retrieval; (2) for both modules, the answer-based retrieval primarily improves the retrieval quality than the query-based setting.

Retriever	Retrieval	Top-1	Top-5	Top-10
Multilingual E5-large	Query-based	45.8	61.3	70.5
Multilingual E5-large	Answer-based	65.3	84.5	88.5
CoROM	Query-based	47.5	60.8	71.5
CoROM	Answer-based	58.8	72.5	80.5

Table B.1: **Results of different retrieval seteps.**

Appendix C

Additional Experimental Details of SpeechIQ

C.1 Complete Results

C.1.1 WER Results

As in Figure C.1 C.2 C.3, the WER performance on each dataset show a consistent tendency as the average score that cascaded $\text{LLM}_{\text{Voice}}$ overallly show better performance than end-to-end $\text{LLM}_{\text{VoiceS}}$, and GER seems to negatively influence the ASR. We also notice that AnyGPT and Desta2 show much lower performance than other models, with closer check into their outputs, we find that sometimes they can not recognize the voice and provide meaningless responses, while their foundation ASR models can succeed in recognizing. This may indicate the potential problem in their modality alignment training.

C.1.2 Similarity Results

As in Figure C.4 C.5 C.6, two architectures become competitive on each dataset, especially for end-to-end-large $\text{LLM}_{\text{VoiceS}}$ show strongest performance. Meanwhile, GER transfers to provide positive influence on ASR results by capturing the semantic meaning of the voice inputs.

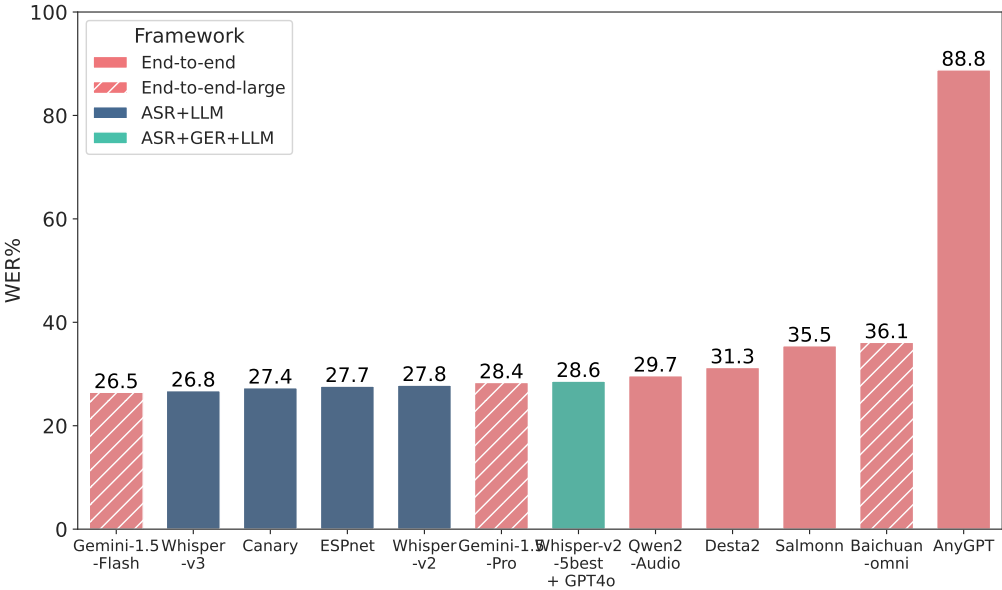


Figure C.1: WER on earning22

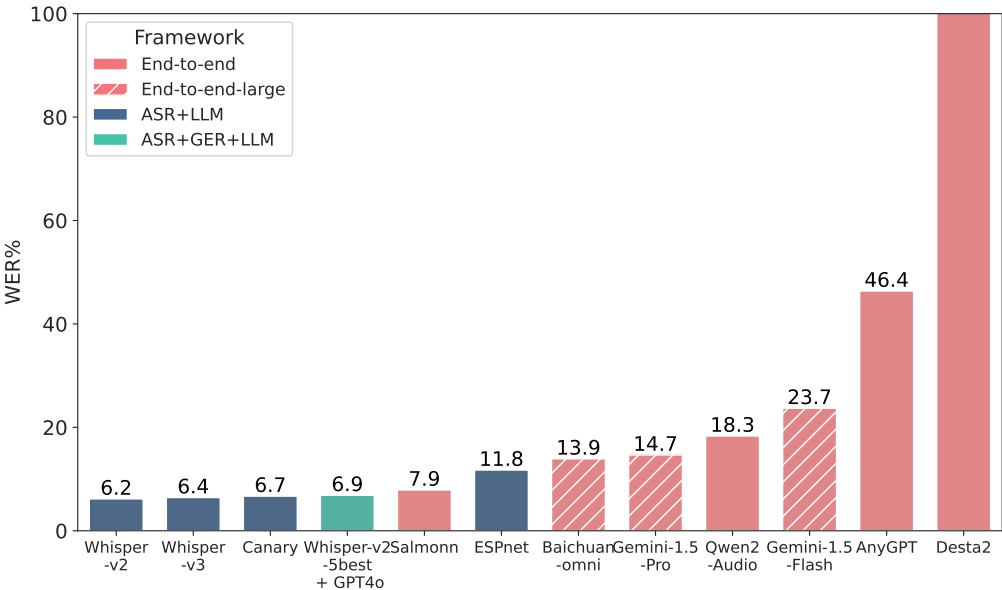


Figure C.2: WER on medasr

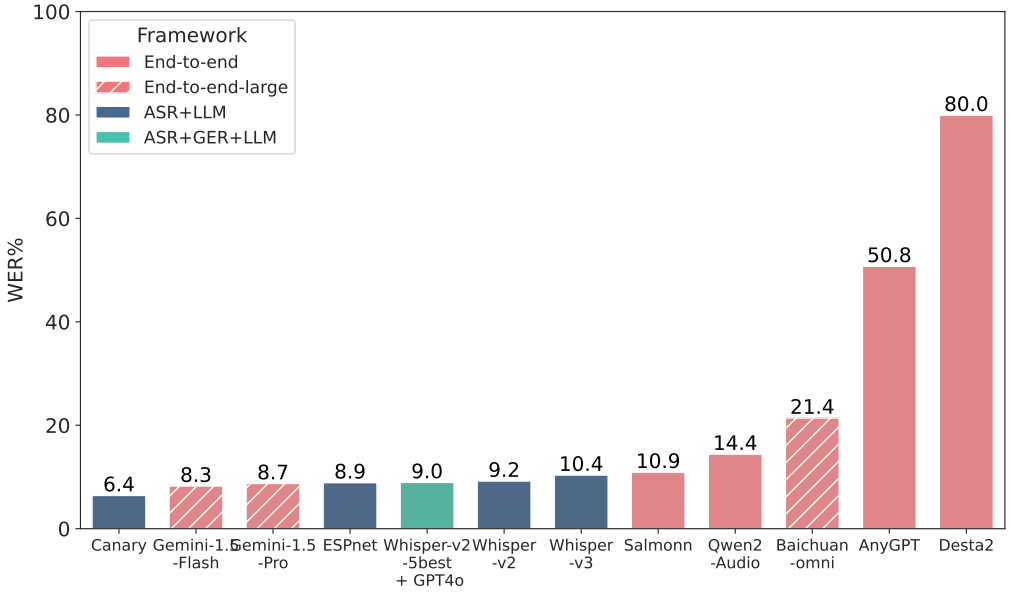


Figure C.3: WER on voxpopuli

C.1.3 QA Accuracy Results

As in Figure C.7 C.8 C.9, a interesting tendency occurs that end-to-end-large LLM_{Voice}s show the best performances in solving multi-choice questions while smaller end-to-end LLM_{Voice}s show the worst performances, with cascaded LLM_{Voice}s lying on the mid. This may indicate that modality alignment need scaling-up data size to train robust speech-text models and better than their cascaded foundation variants.

C.2 Scaling Law Normalization

We realize that normalize the SIQ based on computations in training related to both model size and data size will be a significant step for fairly comparing various models. This normalization matches the age-aware normalization in computing human IQ, since models trained with less computations should be rewarded more SIQ if perform similar or better than larger models, representing the better training strategies in modality alignment. Thus, in our future work we will make

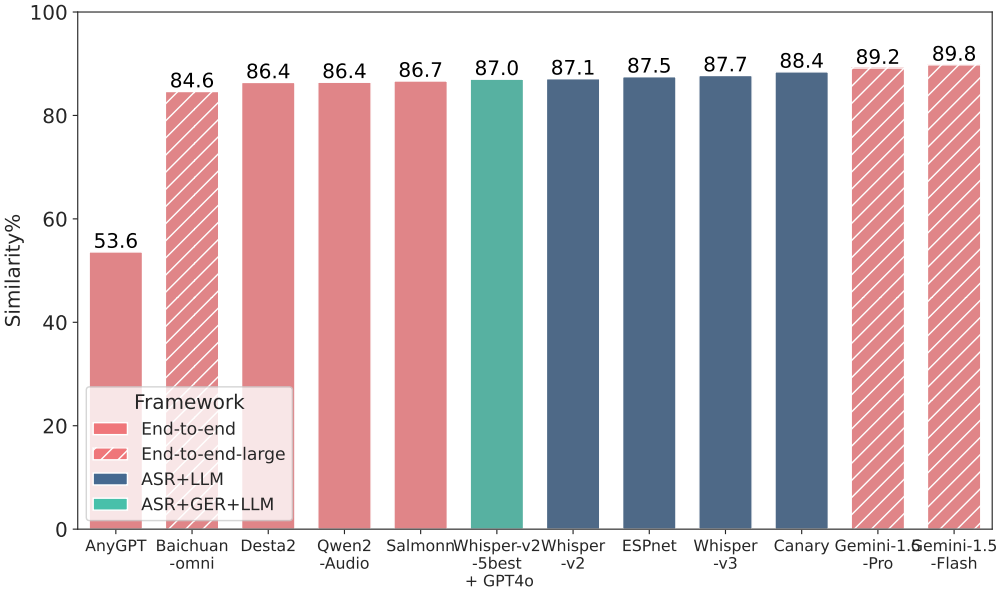


Figure C.4: Similarity on earning22

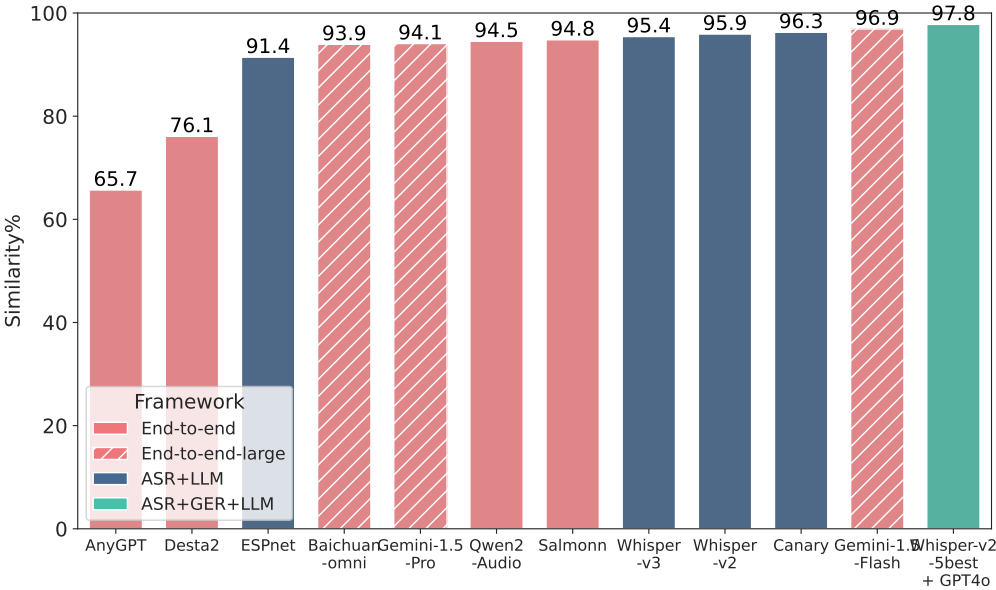


Figure C.5: Similarity on medasr

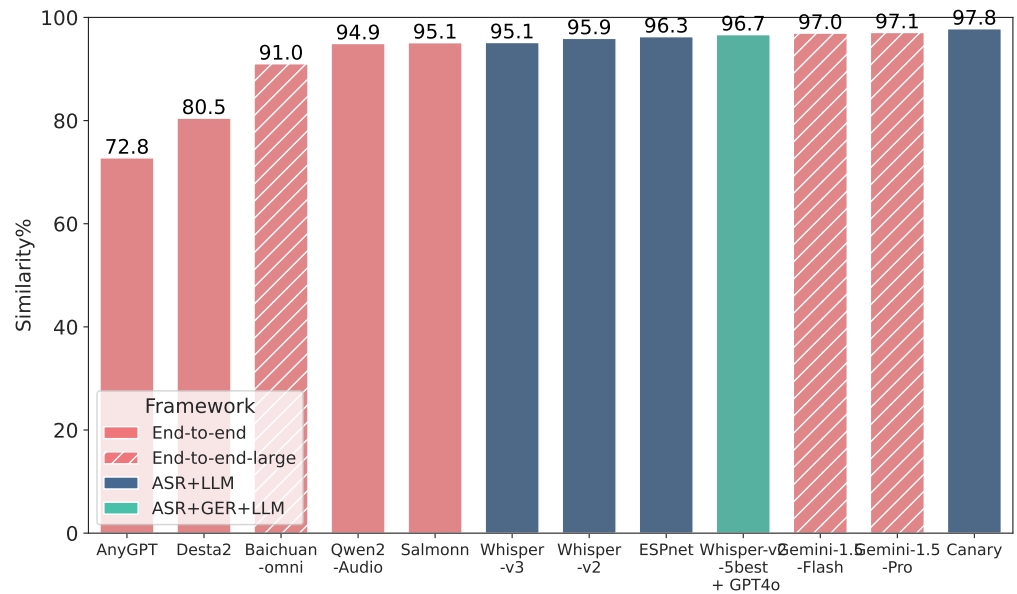


Figure C.6: Similarity on voxpopuli

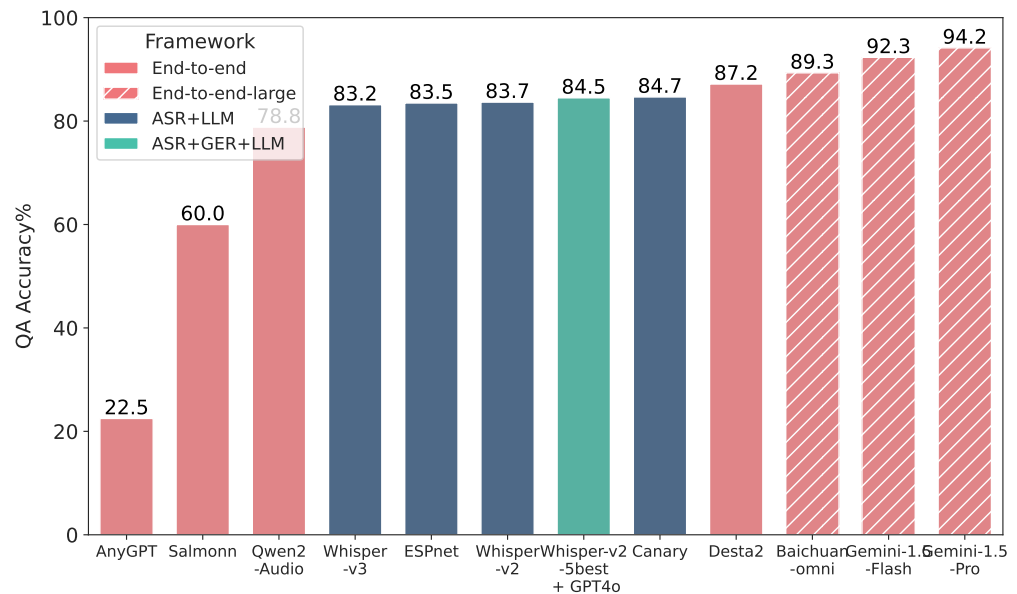


Figure C.7: QA accuracy on earning22

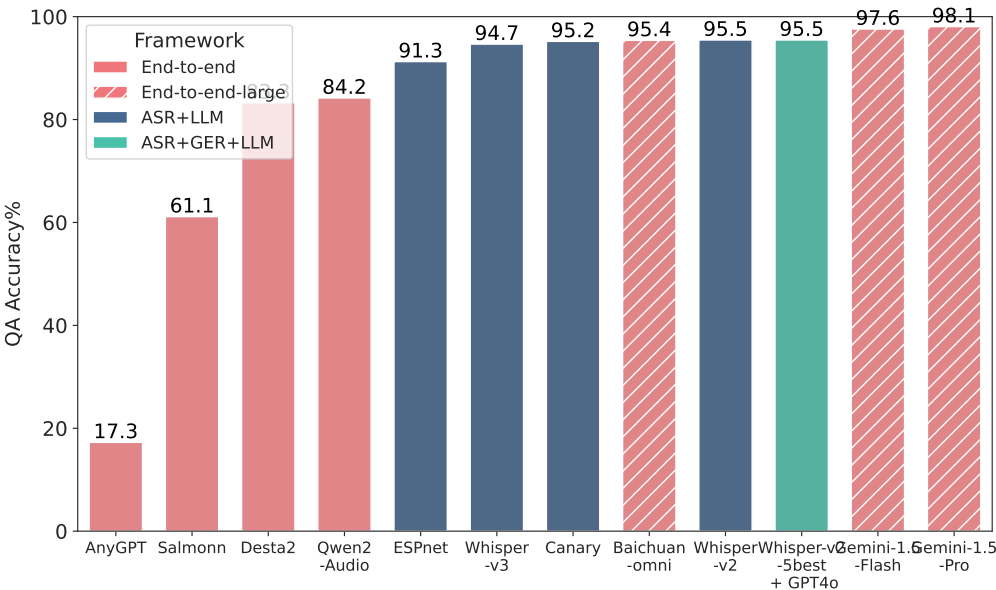


Figure C.8: QA accuracy on medasr

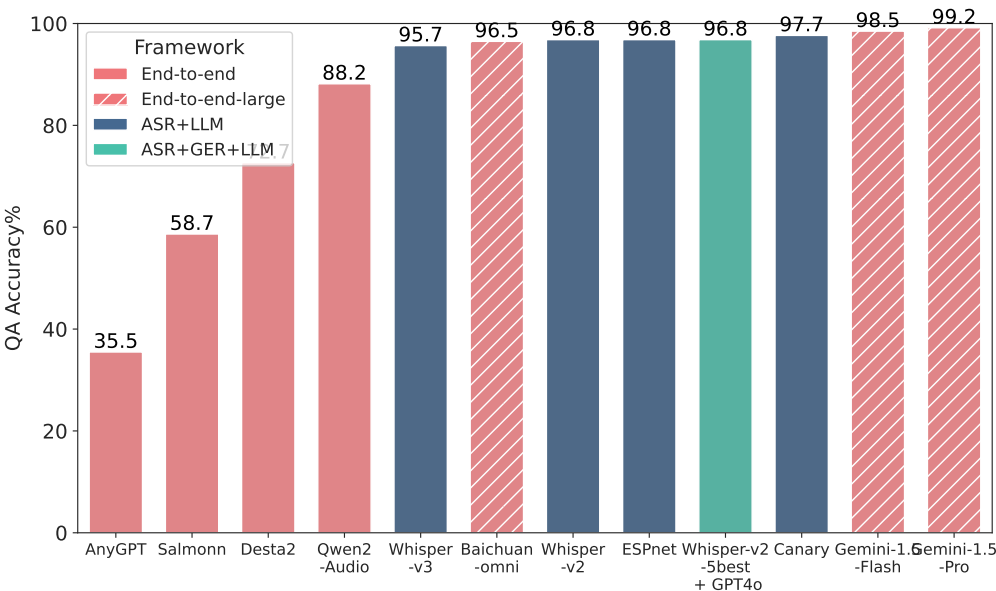


Figure C.9: QA accuracy on voxpopuli

efforts to include the scaling law normalization, even considering the fact that the computation cost is sometimes closed to acquire.

Appendix D

Additional Experimental Details of Speech-Hands

D.1 Failure Case of Multimodal GER

D.1.1 Amplifying Hallucination

This is often a situation where external audio perception is not strong enough or N-best decoding introduces noise for GER. Especially when the omni model is encouraged to take all hypotheses in consideration.

ASR Failure Case

Internal: insane

External (Whisper 5-best): [' Gimseeinnnnnn', ' You can say.', ' Insta.', ' Wednesday.', " I'm from Phelps County, I'm gonna see what this guy's doing."]

GER Pred: I'm from Phelps County, I'm gonna see what this guy's doing

Gold: insane

D.1.2 Overcorrection

Due the language modeling nature, LLMs tend to revise the transcription to be more like a complete sentence, which sometimes caused “overcorrection.”

ASR Failure Case

Internal: you in the way marguerite but how

External (Whisper 5-best): [' you ll in the way marguerite but how', 'you in the way marguerite but how', 'you in the way marguerite but how.', ' you in the way marguerite but how.', " you in the way marguerite but how."]

GER Pred: you are in the way marguerite but how

Gold: you in the way marguerite but how

D.2 Dataset Details for DCASE2025 AudioQA

Subset	#Train / #Dev	<in>/<ex>/<re>
Bio-acoustic QA	0.7K / 0.2K	338/234/168
Soundscape QA	1.0K / 0.6K	604/182/252
Complex QA	6.4K / 1.6K	4,267/1,785/391

Table D.1: Statistics of the three DCASE2025 AudioQA subsets used in our experiments. <in>/<ex>/<re> shows the token distribution in training w/ majority sampling.

The DCASE2025 AudioQA benchmark comprises three complementary multiple-choice question-answering subsets, each designed to evaluate a different aspect of audio reasoning.

D.2.1 Bioacoustics QA

This subset targets perceptual and cognitive grounding in marine bioacoustics. It includes questions about 31 marine mammal species with diverse acoustic ranges,

habitats, and vocalization durations. Tasks include species classification, vocalization type recognition, factual retrieval, interpretation of acoustic features, and comparative reasoning. The dataset includes approximately 0.7K training and 0.2K development QA pairs. Audio clips range in sample rate from 600 Hz to 160 kHz and in duration from 0.4 s to 625 s, allowing evaluation under highly varied acoustic conditions. All audio is sourced from the Watkins Marine Mammal Sound Database (Woods Hole Oceanographic Institution; New Bedford Whaling Museum), and usage of audio beyond the provided splits is strictly prohibited.

D.2.2 Temporal Soundscapes QA

This subset focuses on temporal reasoning over sound events, encompassing 26 event classes. Questions require identifying active sound classes, temporal ordering, timestamp estimation (onset, offset, duration), and event comparison. The subset comprises approximately 1.0K training and 0.6K development QA pairs. Audio clips are mono-channel, 10 seconds long, and sampled at 32–48 kHz. Most clips correspond to a single QA item, while a small portion supports multiple questions. All annotations are manually verified, which include event types, timestamps, and answers. Audio is sourced from NIGENS, L3DAS23 Challenge [117], and TAU Spatial Sound Events 2019 datasets [2].

D.2.3 Complex QA (MMAU)

This subset evaluates high-level reasoning over natural sound scenes. Each instance consists of a 10-second, 16 kHz audio clip paired with a question requiring reasoning over acoustic, temporal, and contextual cues (*i.e.*, overlapping events, implied sequences, or abstract relationships.) Tasks are inspired by the MMAU Sound benchmark [155] and extend its principles to more diverse soundscapes. The data include approximately 6.4K training and 1.6K development QA pairs. Audio clips are sourced from AudioSet and Mira.

D.3 Cases On ASR

We provide several examples to show the effectiveness of our Speech-Hands in ASR tasks.

ASR Case 1

Q. What is the correct transcription of this audio?

Internal: water and fresh stalls to take on there

External: we've water and fresh stores to tick on there

Rewrite: we've water and fresh stores to take on there

Pred: <rewrite> we've water and fresh stores to take on there

Gold: we've water and fresh stores to take on there

ASR Case 2

Q. What is the correct transcription of this audio?

Internal: then has nothing happened with a voice like a whip mr wicker said start at once

External: then as nothing happened with a voice like a whip mister wigger said start at once

Rewrite: then as nothing happened with a voice like a whip mister wuthers said start at once

Pred: <external> then as nothing happened with a voice like a whip mister wigger said start at once

Gold: then as nothing happened with a voice like a whip mister wicker said start at once

D.4 Cases On AudioQA

For each token we pick up 5 examples to illustrate how action tokens influence the final correctness. For `<rewrite>` token, it works not only when the internal prediction is conflicted with the external prediction, but also show robust and accurate decision when hypotheses generate the same but wrong prediction. This is extremely important when facing challenging tasks, the omni-agent should be confident to refuse wrong hypotheses. Also for `<internal>` and `<external>` tokens that have more frequencies, they should be accurately distinguished when conflicts, and action tokens complete a good job on recognizing the true answer.

AudioQA Case 1

Q. Based on the audio, which natural phenomenon could be occurring?

Options:

- A. Earthquake
- B. Thunderstorm
- C. Forest fire
- D. Snowstorm

Internal pred: B. Thunderstorm

External pred: B. Thunderstorm

Pred: `<rewrite>` C. Forest fire

Gold: `<rewrite>` C. Forest fire

AudioQA Case 2

Q. What activity is likely occurring based on the described sounds?

Options:

- A. Driving a car
- B. Planting a garden
- C. Painting a picture
- D. Cooking a meal

Internal pred: A. Driving a car

External pred: B. Planting a garden

Pred: <rewrite> D. Cooking a meal

Gold: <rewrite> D. Cooking a meal

AudioQA Case 3

Q. Based on the audio, what is the most likely environment in which the sounds are occurring?

Options:

A. An indoor setting

B. A rural countryside

C. A busy city street

D. A quiet residential neighborhood

Internal pred: A. An indoor setting

External pred: B. A rural countryside

Pred: <rewrite> C. A busy city street

Gold: <rewrite> C. A busy city street

AudioQA Case 4

Q. What background sound is present in the audio clip?

Options:

A. Waves crashing

B. Birds chirping

C. Bells ringing

D. Running car engine

Internal pred: A. Waves crashing

External pred: D. Running car engine

Pred: <rewrite> C. Bells ringing

Gold: <rewrite> C. Bells ringing

AudioQA Case 5

Q. What might the purpose of tapping the metal object in the background be?

Options:

- A. To emphasize the speaker's instructions
- B. To distract from the speaker's sad tone
- C. To demonstrate its material quality
- D. To create a rhythmic background

Internal pred: A. To emphasize the speaker's instructions

External pred: A. To emphasize the speaker's instructions

Pred: <rewrite> C. To demonstrate its material quality

Gold: <rewrite> C. To demonstrate its material quality

AudioQA Case 6

Q. What type of mood is conveyed through the musical elements in this audio?

Options:

- A. Calm and soothing
- B. Angry and aggressive
- C. Joyful and uplifting
- D. Sad and reflective

Internal pred: B. Angry and aggressive

External pred: D. Sad and reflective

Pred: <external> D. Sad and reflective

Gold: <external> D. Sad and reflective

AudioQA Case 7

Q. Based on the audio description, what is likely happening in the background?

Options:

- A. A calm evening
- B. A windy day
- C. An earthquake
- D. A quiet morning

Internal pred: B. A windy day

External pred: C. An earthquake

Pred: <external> C. An earthquake

Gold: <external> C. An earthquake

AudioQA Case 8

Q. What type of sound is present in the background of the audio clip?

Options:

- A. Car engine
- B. Ocean waves
- C. Upbeat synthesized music
- D. Bird chirping

Internal pred: B. Ocean waves

External pred: C. Upbeat synthesized music

Pred: <external> C. Upbeat synthesized music

Gold: <external> C. Upbeat synthesized music

AudioQA Case 9

Q. Based on the audio description, what is the primary focus of the sounds?

Options:

- A. A quiet library setting
- B. A peaceful nature scene

C. A busy city street

D. A defense attack scenario

Internal pred: A. A quiet library setting

External pred: D. A defense attack scenario

Pred: <external> D. A defense attack scenario

Gold: <external> D. A defense attack scenario

AudioQA Case 10

Q. Based on the audio description, what type of activity is most likely taking place?

Options:

A. Gardening

B. Woodworking

C. Lock-picking

D. Cooking

Internal pred: B. Woodworking

External pred: C. Lock-picking

Pred: <external> C. Lock-picking

Gold: <external> C. Lock-picking

AudioQA Case 11

Q. What element in the audio contributes to the emotional depth of the song besides the vocals?

Options:

A. The language spoken

B. The steady drum beats

C. The groovy bass line

D. The keyboard harmony

Internal pred: D. The keyboard harmony

External pred: C. The groovy bass line

Pred: <internal> D. The keyboard harmony

Gold: <internal> D. The keyboard harmony

AudioQA Case 12

Q. Based on the audio description, what type of environment is suggested by the background sounds?

Options:

- A. A serene forest
- B. A quiet library
- C. A beach with waves
- D. A busy city street

Internal pred: D. A busy city street

External pred: D. A busy city street

Pred: <internal> D. A busy city street

Gold: <internal> D. A busy city street

AudioQA Case 13

Q. What might the purpose of the whistle and crinkling leaves in the background be?

Options:

- A. To create a suspenseful atmosphere
- B. To signal the attention of someone nearby
- C. To mimic a bustling city environment
- D. To indicate the presence of wildlife

Internal pred: B. To signal the attention of someone nearby

External pred: B

Pred: <internal> B. To signal the attention of someone nearby

Gold: <internal> B. To signal the attention of someone nearby

AudioQA Case 14

Q. Why does the conversation feature electronic beats and rhythmic cymbal sounds?

Options:

- A. To mimic the sounds of a busy environment
- B. To create a sense of urgency in the interaction
- C. To drown out background noise
- D. To enhance the emotional depth of the speech

Internal pred: B. To create a sense of urgency in the interaction

External pred: D. To enhance the emotional depth of the speech

Pred: <internal> B. To create a sense of urgency in the interaction

Gold: <internal> B. To create a sense of urgency in the interaction

AudioQA Case 15

Q. What sound appears earliest in the audio?

Options:

- A. Ticking
- B. Accelerating, revving, vroom
- C. Idling
- D. Car

Internal pred: C. Idling

External pred: D. Car

Pred: <internal> C. Idling

Gold: <internal> C. Idling

List of Publications

- [1] Zhen Wan*, Qianying Liu*, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi and Jiwei Li. Rescue Implicit and Long-tail Cases: Nearest Neighbor Relation Extraction. EMNLP 2022.
- [2] Zhen Wan, Fei Cheng, Qianying Liu, Zhuoyuan Mao, Haiyue Song, Sadao Kurohashi. Relation Extraction with Weighted Contrastive Pretraining on Distant Supervision. EACL 2023 Findings.
- [3] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, Sadao Kurohashi. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. EMNLP 2023.
- [4] Yikun Sun*, Zhen Wan*, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, Sadao Kurohashi. Rapidly Developing High-quality Instruction Data and Evaluation Benchmark for Large Language Models with Minimal Human Effort: A Case Study on Japanese. Coling 2024.
- [5] Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, Sadao Kurohashi. Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain. ACL 2024 Findings.
- [6] Zhen Wan, Chao-Han Huck Yang, Yahan Yu, Jinchuan Tian, Sheng Li, Ke Hu, Zhehuai Chen, Shinji Watanabe, Fei Cheng, Chenhui Chu, Sadao Kurohashi. SpeechIQ: Speech-Agentive Intelligence Quotient Across Cognitive Levels in Voice Understanding by Large Language Models. ACL 2025.
- [7] Zhen Wan, Fei Cheng, Sadao Kurohashi. GPT-MM: Improving Multimodal In-context Learning with Task-specific Retrieval and Reasoning. In *Journal of Natural Language Processing* 2026.
- [8] Zhen Wan, Chao-Han Huck Yang, Jinchuan Tian, Hanrong Ye, Ankita Pasad, Szu-Wei Fu, Arushi Goel, Ryo Hachiuma, Shizhe Diao, Kunal Dhawan, Sreyan Ghosh, Yusuke Hirota, Zhehuai Chen, Rafael Valle, Shinji Watanabe, Boris Ginsburg, Yu-Chiang Frank Wang. Speech-Hands: A Self-Reflection

Voice Agentic Approach to Speech Recognition and Audio Reasoning with Omni Perception. Under submission to ACL 2026.