General Paper Peer-Reviewed

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

Zhen Wan[a], Fei Cheng[a] and Sadao Kurohashi[a,b]

Large language models (LLMs) have exhibited impressive generalization through in-context learning (ICL), yet most studies focus on textual tasks, leaving the mechanisms that enable ICL to generalize across modalities largely unexplored. To bridge this gap, we propose a unified ICL framework that integrates task-aware demonstration retrieval and label-induced reasoning as two complementary components for improving both accuracy and interpretability.

We first validate the framework in textual relation extraction (RE), a representative structured prediction task that challenges LLMs to infer fine-grained entity–relation semantics. Task-aware retrieval ensures that retrieved examples are semantically aligned with the target instance, while label-induced reasoning enriches each demonstration with label-grounded explanatory logic. These mechanisms substantially narrow the performance gap between ICL and fully supervised models.

We then extend this framework to multimodal ICL, leveraging GPT-4o for visual question answering (VQA) and Whisper-large-v3 for audio question answering (AudioQA). Across both textual and multimodal benchmarks, our framework consistently outperforms GPT-3 and GPT-4 baselines and achieves competitive or superior results compared with fine-tuned models. These findings demonstrate that task-aware retrieval and label-induced reasoning together form a generalizable foundation for a unified in-context learning paradigm across modalities.

Key Words: *Relation Extraction, In-context Learning, Multimodal In-context Learning*

1 Introduction

The emergence of large language models (LLMs) such as GPT-3 (Brown et al. 2020) represents a significant advancement in natural language processing (NLP). Instead of following a pretraining-and-finetuning pipeline (Devlin et al. 2019; Beltagy et al. 2019; Raffel et al. 2019; Lan et al. 2019; Zhuang et al. 2021), 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)

[a] Graduate School of Informatics, Kyoto University
[b] National Institute of Informatics

The preliminary part on one particular textual task RE (GPT-RE) of this study was presented at EMNLP 2023 (Wan et al. 2023). This article extended our proposed two methods and proposed a unified framework cross modalities using an Omni model for multimodal ICL.

(C) The Association for Natural Language Processing. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

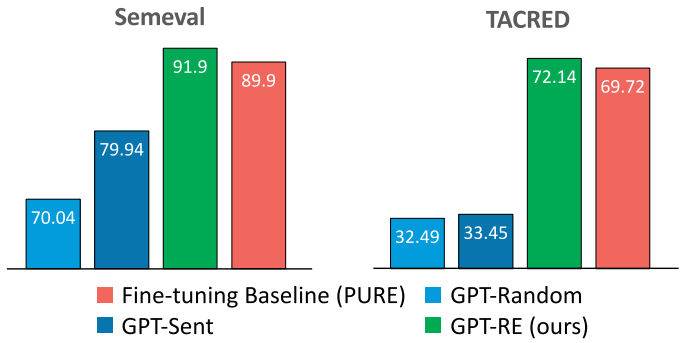


Figure 1 Micro F1 performances on two RE datasets. Previous GPT baselines (*GPT-Random*: randomly selected demonstrations and *GPT-Sent*: sentence-level demonstration retrieval) largely underperform fine-tuning baseline PURE while our *GPT-RE* substantially outperforms all baselines.

(Brown et al. 2020; Min et al. 2022) which formulates an NLP task under the paradigm of language generation and makes predictions by learning from a few demonstrations.

Recently, the success of multimodal large language models (MLLMs) such as GPT-4o and Gemini has extended this paradigm beyond text, demonstrating that ICL can also operate on vision–language and audio–language inputs (Tsimpoukelli et al. 2021; Alayrac et al. 2022; Zhu et al. 2023; Chen et al. 2024). These efforts aim to bring the strengths of LLMs into multimodal ICL by enabling models to process images and audio inputs using the same demonstration-based framework as in language tasks. Despite this progress, the investigation of multimodal in-context learning (MM-ICL) remains in its infancy. Existing works mainly investigate certain modality encoder-based retrieval strategies, while little attention has been paid to whether a unified ICL mechanism can operate consistently across modalities such as text, vision, and audio. These developments raise a fundamental question we intend to answer in this paper: can the mechanisms that make ICL effective for textual reasoning generalize to a unified multimodal framework as well?

Before addressing multimodal ICL, it is important to understand why ICL underperforms in structured textual tasks such as relation extraction (RE). RE is the central task for knowledge 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. Previous research (Gutiérrez et al. 2022) has sought to apply GPT-3

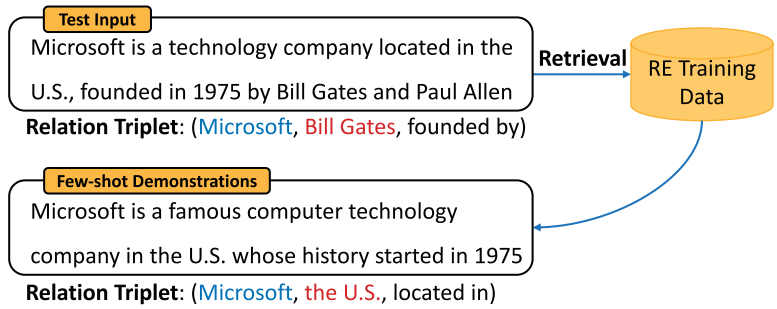


Figure 2 Retrieval without considering the task-aware triplet results in noisy demonstrations.

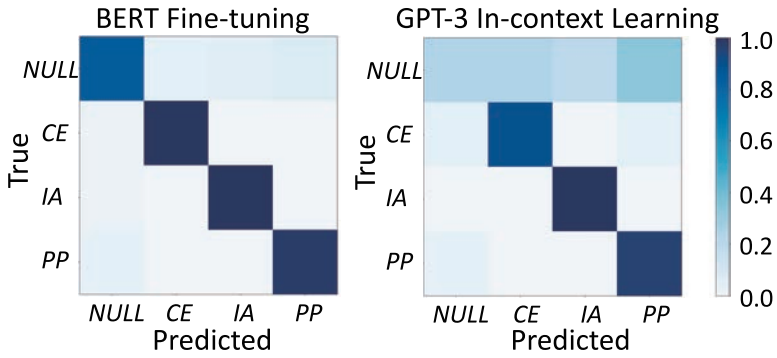


Figure 3 Confusion matrix on Semeval dataset with three selected relation labels. The NULL examples are overpredicted to other relations by GPT-3. CE: Cause-Effect, IA: Instrument-Agency, PP: Product-Producer.

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 (kNN) search based on sentence embedding (Liu et al. 2022b; Gutiérrez et al. 2022). Regrettably, kNN-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, the test input 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 kNN 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 (Wei et al. 2022; Wang et al. 2022; Kojima et al. 2022). But different 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.

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. To bridge this gap, we extend our ICL-based framework to two new modalities: image and audio. We instantiate the MM-ICL paradigm on one image understanding task—Visual Question Answering (VQA), and one audio task—automatic speech recognition (ASR), where each test instance is accompanied by a few image/audio-text demonstration triplets. Our goal is to investigate whether the same retrieval and reasoning techniques that enhance RE in NLP can be adapted to improve performance in vision and speech domains as well.

Specifically, we introduce GPT-MM, a multimodal extension of GPT-RE that applies our framework to visual and auditory tasks, incorporating image and audio demonstrations retrieved using task-specific similarity in each modality metrics. By injecting CoT-style reasoning adapted to each modality, we further enhance the model's ability to generalize from few-shot prompts. We evaluate our GPT-MM on two popular benchmarks: OKVQA for visual question answering and Earning22 for ASR. Our results show that GPT-MM achieves superior performance under few-

shot settings. These findings validate the generalizability of our retrieval and reasoning strategies and highlight the promise of in-context learning as a unified paradigm across modalities.

We summarize our contributions as follows:

- As our preliminary study, we diagnose why vanilla ICL underperforms on RE (low entity/relation relevance; no input–label justification) and propose GPT-RE, combining task-aware retrieval with gold label–induced reasoning. The experiment results show that our proposed methods signifiantly outperform previous GPT baselines
- We propose GPT-MM: a Unified multimodal ICL framework. We extend two above proposed methods to VQA and ASR (modality-specific similarity + CoT-style reasoning), showing few-shot gains in both visual and audio tasks over GPT baselines.
- We implement experiments on popular benchamrks to validate our proposed methods and also implement sufficient ablations and discussions for a deeper understanding of the improvements and guiding further researches.

2 Preliminary Study on Text: GPT-RE

2.1 Task Definition of RE

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 $y = \text{NULL}$.

2.2 Overview

We will first introduce the prompt construction to formalize RE as a language generation task in Sec. 2.3. Then to improve the ICL framework for RE, we will introduce two modules: (1) task-aware demonstration retrieval to select higher-quality demonstrations (Sec. 2.4); (2) gold label-induced reasoning to enrich each demonstration with explanations (Sec. 2.5). In Figure 4, we show the concrete workflow of processing a test input.

2.3 Prompt Construction

We construct a prompt for each given test example, which is fed to the GPT 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 predefined classes $\mathbb{R}$. The model is explicitly asked to output the relation, which belongs to the

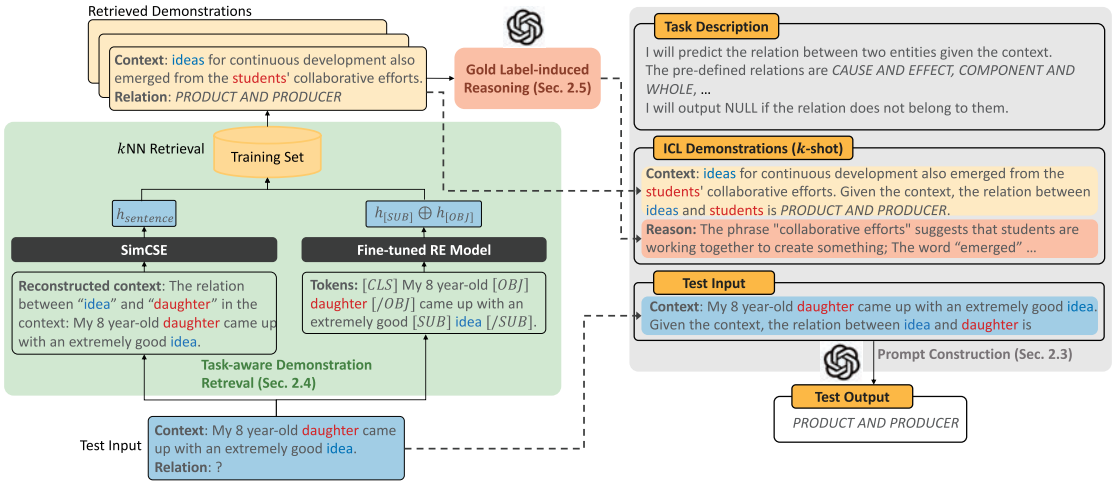


Figure 4 An illustration of GPT-RE. Given a test input, we first leverage two different task-aware retrieval methods to search for highly relevant demonstrations from the training set, and then incorporate the gold label-induced reasoning for each demonstration. Above contents will then be included in the prompt construction to make the prediction.

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 is expected to generate the corresponding relation y_{test}.

In summary, GPT-RE can be formulated as:

$$p\left(y_{test} \in \mathbb{R} \cup \{\text{NULL}\} | \mathcal{I}, \mathcal{D}, x_{test}\right) \tag{1}$$

2.4 Task-aware Demonstration Retrieval

ICL demonstrations closer to the test sample in the embedding space result in more consistent and robust performance (Liu et al. 2022b). Recent work (Gutiérrez et al. 2022; Liu et al. 2022b) employs the kNN to retrieve the most similar examples in the training set as the few-shot demonstrations for each test input. As kNN 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 kNN retrieval has a severe drawback: relation extrac-

tion 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. 2.4.1; (2) an advanced fine-tuned relation representation in Sec. 2.4.2.

2.4.1 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 (Gao et al. 2021) for computing sentence embedding-based similarity.

2.4.2 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 (Baldini Soares et al. 2019; Zhong and Chen 2021; Wan et al. 2022) 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 $p(y \in \mathbb{R} \cup \{\textsc{null}\} \mid \mathbf{Rel})$.

The entity markers have explicitly encoded subject and object entities and the relation representation **Rel** is naturally enriched with the entity information. We believe this approach can potentially compensate for the limitations of GPT in RE. While GPT 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 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.

2.5 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 induce the reasoning logic for each demonstration by the corresponding gold relation label. As shown in Figure 5, 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 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.

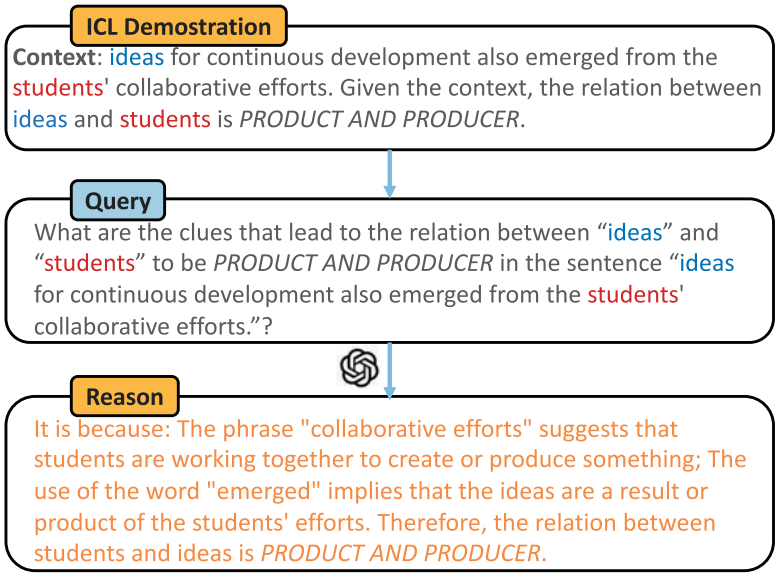


Figure 5 An illustration of adding reasoning.

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 1 Statistics of datasets.

3 Experiment Setup on RE

3.1 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, in our main results, we sample a subset (See Appendix C) from the original test set for two datasets: ACE05 and TACRED as shown in Table 1.

Semeval 2010 task 8 (Hendrickx et al. 2010) focuses on semantic relations between pairs of nominals collected from general domain resources.

TACRED (Zhang et al. 2017) is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text.

SciERC (Luan et al. 2018) collects AI paper abstracts and annotated relations, especially for scientific knowledge graph construction.

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

3.2 Baseline Methods

GPT baselines For GPT-3 and GPT-4 baselines and our methods, we select "`text-davinci-003`" for GPT-3 with maximal 4,097 input tokens, and default GPT-4 from OpenAI's API (for fair comparison, we use the same input length as GPT-3) and use the identical prompt construction (Sec. 2.3). We select GPT-4 instead of GPT-4o since in our preliminary experiments GPT-4 outperforms GPT-4o as a baseline. We implement two categories of GPT 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. In this work, our implementation adopted SimCSE (Gao et al. 2021), 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 (Zhong and Chen 2021), an entity marker-based fine-tuned model mentioned in Sec. 2.4.2 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 (Cohen et al. 2020) (reformulate RE as the question answering task) and TACRED (Wang et al. 2022) (extra pre-training to capture RE structure) datasets.

All implementation details are in Appendix A.

4 Experimental Results on RE

4.1 Main Results

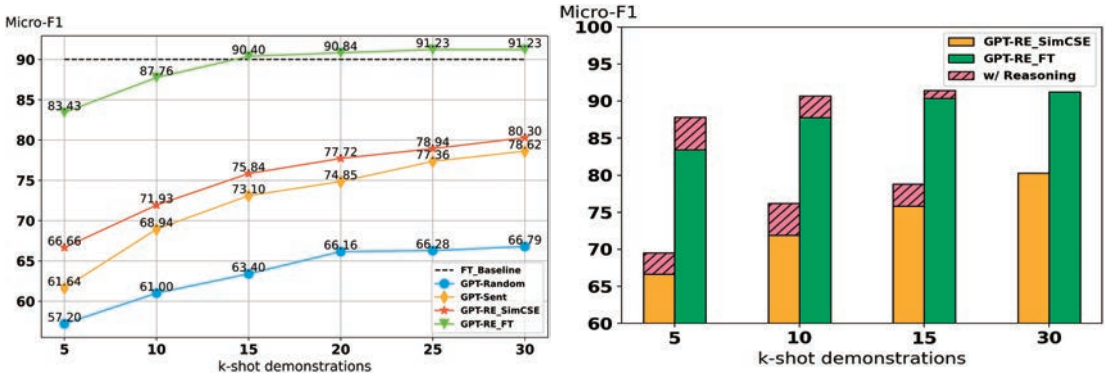
We compare our main experiment results with previous methods in Table 2. In general, we select the largest k avaialble in our experiment as the best k so that the input will not exceed the limited input lenth of GPT-3. Thus, for reasoning setup, the k will be smaller due to the longer length for each demonstration. As a result, we observe that GPT-4 consistently outperforms GPT-3 across most scenarios, although the margin of improvement remains modest, as RE is inherently a challenging task for the GPT-series models that have not been fine-tuned on RE. **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 *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 or GPT-4 has the potential to beat fine-tuning when the retriever has prior task knowledge. *GPT4-RE_FT* eventually achieves best results on Semeval and SciERC. (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 4.3.

Methods	Retriever	Semeval	TACRED	SciERC	ACE05
GPT Baselines (Best k-shot)					
GPT3-Random	—	70.04 (30)	32.49 (15)	17.92 (25)	9.04 (25)
GPT3-Sent	SimCSE	79.94 (30)	33.45 (15)	20.96 (25)	6.31 (25)
GPT4-Random	—	67.83 (30)	—	16.48 (25)	9.73 (25)
GPT4-Sent	SimCSE	77.64 (30)	—	21.60 (25)	10.04 (25)
Ours (Best k-shot)					
GPT3-RE_SimCSE	SimCSE	81.02 (30)	37.44 (15)	26.46 (25)	8.67 (25)
GPT3-RE_SimCSE*	SimCSE	77.49 (15)	31.58 (10)	—	—
+ Reasoning	SimCSE	79.88 (15)	33.18 (10)	—	—
GPT3-RE_FT	PURE	**91.90** (25)	72.14 (15)	**69.00** (30)	68.73 (25)
GPT3-RE_FT*	PURE	91.11 (15)	70.38 (10)	—	—
+ Reasoning	PURE	91.82 (15)	70.97 (10)	—	—
GPT4-RE_FT	PURE	**91.97** (25)	—	**69.12** (30)	69.13 (25)
+ Reasoning	PURE	**92.94** (25)	—	**69.93** (30)	69.76 (25)
Fine-tuned RE Baselines					
(Cohen et al. 2020)		**91.90**	—	—	—
(Wang et al. 2022)		—	♣**76.80**	—	—
PURE (Zhong and Chen 2021)		89.90	69.72	68.45	**70.09**

Table 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 expense, we did not conduct reasoning experiments on all datasets. ♣ 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.

4.2 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 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



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

Figure 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.

demonstrations.

4.3 Ablation Study on Reasoning Enhancing

We then check the influence of our proposed reasoning-enhanced demonstration, as shown in Figure 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 (Han et al. 2018; Geng et al. 2020; Liu et al. 2022a), which aims at recognizing novel relations with very few or no examples, and we leave this for future work.

4.4 Low-resource Scenario

We conduct the experiment for observing the low-resource performance in the general domain Semeval task. As shown in Figure 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 (Gutiérrez et al. 2022). We assume the difference could be caused by the domain and NULL proportion of the task.

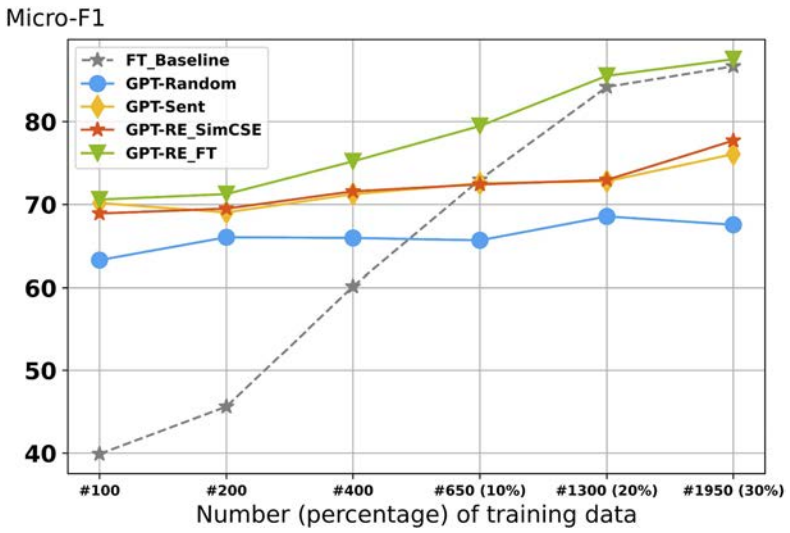


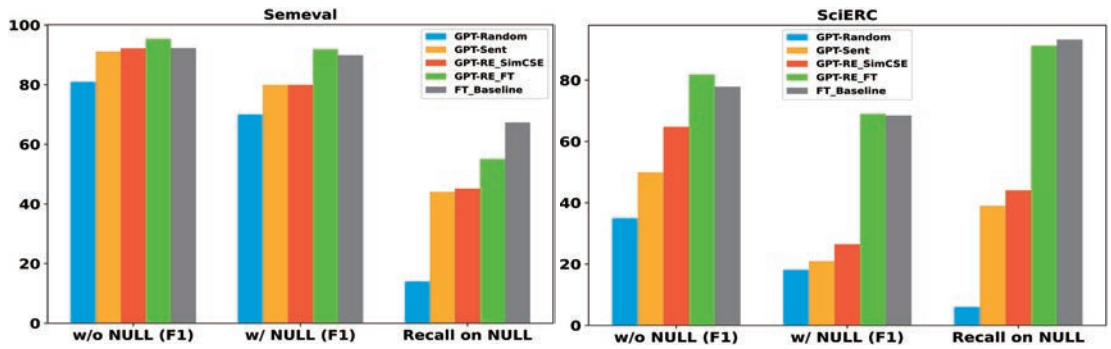
Figure 7 Low-resource Scenario on Semeval. We limit the percentage of training data for both fine-tuning and retrieval in GPT-RE.

5 Analysis and Case Study on RE

5.1 The Issue of "Overpredicting"

Recent work reveals another crucial problem named "overpredicting" as shown in Figure 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 (Gutiérrez et al. 2022; Blevins et al. 2022). 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.

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 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



(a) General domain with 17.4% NULL examples

(b) Scientific domain with 90.16% NULL examples

Figure 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.

knowledge. We think this ability of GPT-3 can be useful in more open fields, such as open RE (Banko and Etzioni 2008) which has no pre-defined relation classes.

5.2 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 9, 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

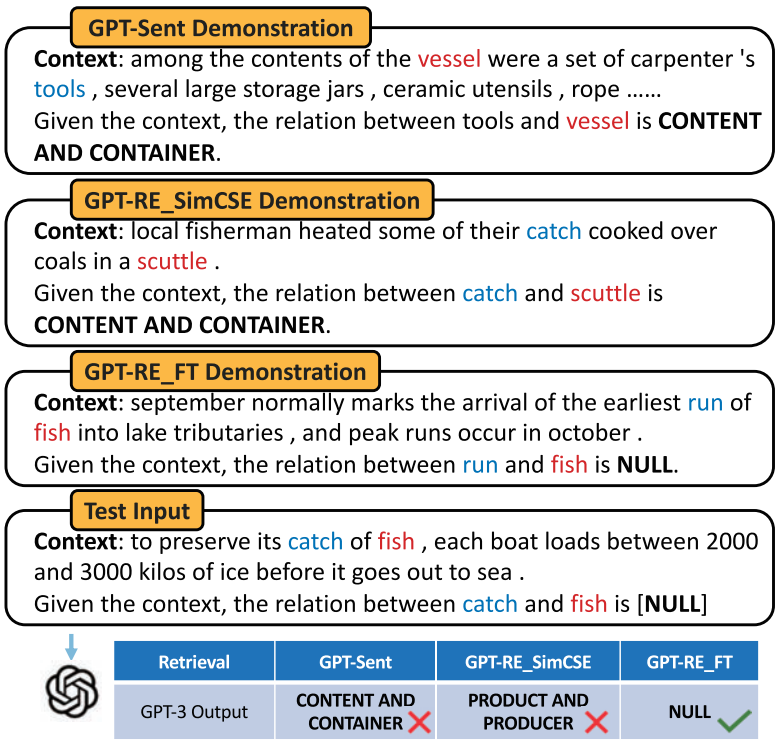


Figure 9 A case study of demonstration quality on Semeval. [NULL] is the gold label here.

enhancement with more entity information provided in retrieval. We also show some other case examples in Appendix B.

6 A Unified Multimodal Extension: GPT-MM

The preceding sections have established that the proposed task-aware retrieval and gold label-induced reasoning substantially enhance in-context learning (ICL) for textual relation extraction (RE). While these findings validate the effectiveness of our framework within language tasks, an open question remains: *can the same principles generalize beyond text to other modalities?*

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.

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

6.1 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:

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

6.2 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.

6.3 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. 6.4), and enrich each demonstration with modality-conditioned reasoning (Sec. 6.5). Each demonstration consists of an input-output-reasoning triplet: (x_i, y_i, r_i).

Test Input x_{test} We convert 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:

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

6.4 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 (Radford et al. 2021) for vision or Whisper (Radford et al. 2022) 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 (Xu et al. 2025) 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.

6.5 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:

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

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

7 Multimodal Experiment Results

7.1 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 (Marino et al. 2019) 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 (Rio et al. 2022) 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.

7.2 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 pre-

dicted 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.

7.3 Main Results

We report the main results of our multimodal in-context learning experiments in Table 3. 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.

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

Method	**OK-VQA (Accuracy)** ↑	**Earnings-22 (WER)** ↓
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 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.

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.

8 Related Work

In-context Learning Recent work shows that ICL of GPT-3 (Brown et al. 2020) 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 (Perez et al. 2021) for proper input to coherence calibration (Malkin et al. 2022) for tackling the diverse generated output. Another research path locates in the demonstration part, including ordered prompts (Lu et al. 2022) and retrieval-based demonstrations (Rubin et al. 2022; Liu et al. 2022b; Shin et al. 2021).

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.

Some recent work also extend ICL into the multimodal domains. Frozen (Tsimpoukelli et al. 2021) is the first attempt to exploit ICL ability in the vision-language model, and (Zhou et al. 2024) improves the performance on VQA by two-step demonstration retrieval to capture both image and text features. In speech, whisper-based ICL (Wang et al. 2024b) has also been widely explored, while some attend to investigate the ICL of speech-language models (Hsu et al. 2024; Pan et al. 2024; Borsos et al. 2023). For unified ICL with multimodal inputs, Bayesian example selection (Wang et al. 2024a) provides a general method for improving ICL in all modalities incuding text, audio and image. However, compared with strong improvements of ICL studies in the text domain, multimodal ICL still needs further research.

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 (Liu et al. 2022b; Shin et al. 2021; Rubin et al. 2022).

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 (Devlin et al. 2019), RoBERTa (Zhuang et al. 2021) KATE (Liu et al. 2022b), SimCSE (Gao et al. 2021), Sentence-BERT (Reimers and Gurevych 2019; Wolf et al. 2020). 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.

Relation to follow-up studies on GPT-RE Several studies have been proposed after our previous work GPT-RE, focusing on improving in-context learning (ICL) for textual relation extraction (RE) from complementary perspectives. (Han et al. 2024) leverage Abstract Meaning Representation (AMR) graphs to better capture the semantic similarity between instances, which is an intrinsic feature of information extraction. (Pang et al. 2023) introduce explicit task guidelines instead of training examples as demonstrations in the prompt to help large language models understand subtle RE rules and edge cases, which is similar to summarizing our gold label-induced reasoning on the whole dataset. (Mo et al. 2024) augment demonstrations by incorporating hard negative examples, let LLMs avoid potential errors in predictions, enabling more discriminative reasoning. (Ma et al. 2023) focuses on the reasoning side, they extract relation-indicative text spans and induce explicit evidence related to both the objective and subjective, and then integrate them into the CoT reasoning process.

Compared with these follow-up studies that further refine textual ICL within the information extraction domain through additional semantic representations, task-specific prompting heuristics, or data augmentation techniques, our work takes a broader, framework-level perspective. The proposed GPT-RE formulation was designed as a general ICL framework rather than a task-specific solution, making its core mechanisms of *task-aware retrieval* and *label-induced reasoning* readily transferable, and we have already extended this framework to all NLP tasks in work (Sun et al. 2023) and achieve significant improvements. In this study, we demonstrate that these mechanisms can not only be applied to other NLP tasks but can also generalize effectively across modalities. Specifically, we extend the same ICL framework to vision–language and audio–language scenarios, forming a unified and interpretable multimodal ICL paradigm.

9 Conclusions

This work presents a unified in-context learning (ICL) framework that enhances both textual and multimodal understanding through two complementary mechanisms—task-aware demon-

stration 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.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. (2022). "Flamingo: A Visual Language Model for Few-Shot Learning." *CoRR*, **abs/2204.14198**.

Baldini Soares, L., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). "Matching the Blanks: Distributional Similarity for Relation Learning." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2895–2905, Florence, Italy. Association for Computational Linguistics.

Banko, M. and Etzioni, O. (2008). "The Tradeoffs Between Open and Traditional Relation Extraction." In *Proceedings of ACL-08: HLT*, pp. 28–36, Columbus, Ohio. Association for Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (2019). "SciBERT: A Pretrained Language Model for Scientific Text." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Blevins, T., Gonen, H., and Zettlemoyer, L. (2022). "Prompting Language Models for Linguistic

Structure." *CoRR*, **abs/2211.07830**.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. (2023). "AudioLM: a Language Modeling Approach to Audio Generation." *CoRR*, **abs/2209.03143**.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). "A Large Annotated Corpus for Learning Natural Language Inference." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). "Language Models are Few-Shot Learners." In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc.

Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K. C., Li, J., Ghosh, S., Balam, J., and Ginsburg, B. (2024). "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)*, pp. 13521–13525.

Cohen, A. D. N., Rosenman, S., and Goldberg, Y. (2020). "Relation Extraction as Two-way Span-Prediction." *CoRR*, **abs/2010.04829**.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *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)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gao, T., Yao, X., and Chen, D. (2021). "SimCSE: Simple Contrastive Learning of Sentence Embeddings." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geng, X., Chen, X., Zhu, K. Q., Shen, L., and Zhao, Y. (2020). "MICK: A Meta-Learning Framework for Few-shot Relation Classification with Small Training Data." In d'Aquin, M., Dietze, S., Hauff, C., Curry, E., and Cudré-Mauroux, P. (Eds.), *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event,*

Ireland, October 19–23, 2020, pp. 415–424. ACM.

Gutiérrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., and Su, Y. (2022). "Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again." *CoRR*, **abs/2203.08410**.

Han, P., Pereira, L. K., Cheng, F., She, W. J., and Aramaki, E. (2024). "AMR-RE: Abstract Meaning Representations for Retrieval-Based In-Context Learning in Relation Extraction." *CoRR*, **abs/2406.10432**.

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., and Sun, M. (2018). "FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals." In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Hsu, M.-H., Chang, K.-W., Li, S.-W., and yi Lee, H. (2024). "Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks." *CoRR*, **abs/2310.12477**.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). "Large Language Models are Zero-Shot Reasoners." *CoRR*, **abs/2205.11916**.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." *CoRR*, **abs/1909.11942**.

Liu, F., Lin, H., Han, X., Cao, B., and Sun, L. (2022a). "Pre-training to Match for Unified Low-shot Relation Extraction." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5785–5795, Dublin, Ireland. Association for Computational Linguistics.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2022b). "What Makes Good In-Context Examples for GPT-3?" In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ma, X., Li, J., and Zhang, M. (2023). "Chain of Thought with Explicit Evidence Reasoning for Few-shot Relation Extraction." In Bouamor, H., Pino, J., and Bali, K. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2334–2352, Singapore. Association for Computational Linguistics.

Malkin, N., Wang, Z., and Jojic, N. (2022). "Coherence Boosting: When Your Pretrained Language Model is not Paying Enough Attention." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. (2019). "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge." *CoRR*, **abs/1906.00067**.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" *CoRR*, **abs/2202.12837**.

Mo, Y., Liu, J., Yang, J., Wang, Q., Zhang, S., Wang, J., and Li, Z. (2024). "C-ICL: Contrastive In-context Learning for Information Extraction." In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10099–10114, Miami, Florida, USA. Association for Computational Linguistics.

Pan, J., Wu, J., Gaur, Y., Sivasankaran, S., Chen, Z., Liu, S., and Li, J. (2024). "COSMIC: Data Efficient Instruction-tuning For Speech In-Context Learning." *CoRR*, **abs/2311.02248**.

Pang, C., Cao, Y., Ding, Q., and Luo, P. (2023). "Guideline Learning for In-Context Information Extraction." In Bouamor, H., Pino, J., and Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15372–15389, Singapore. Association for Computational Linguistics.

Perez, E., Kiela, D., and Cho, K. (2021). "True Few-Shot Learning with Language Models." In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11054–11070.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *CoRR*, **abs/2103.00020**.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). "Robust Speech Recognition via Large-Scale Weak Supervision.".

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *CoRR*, **abs/1910.10683**.

Reimers, N. and Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rio, M. D., Ha, P., McNamara, Q., Miller, C., and Chandra, S. (2022). "Earnings-22: A Practical Benchmark for Accents in the Wild." *CoRR*, **abs/2203.15591**.

Rubin, O., Herzig, J., and Berant, J. (2022). "Learning To Retrieve Prompts for In-Context Learning." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, Seattle, United States. Association for Computational Linguistics.

Shin, R., Lin, C., Thomson, S., Chen, C., Roy, S., Platanios, E. A., Pauls, A., Klein, D., Eisner, J., and Van Durme, B. (2021). "Constrained Language Models Yield Few-Shot Semantic Parsers." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sun, X., Dong, L., Li, X., Wan, Z., Wang, S., Zhang, T., Li, J., Cheng, F., Lyu, L., Wu, F., and Wang, G. (2023). "Pushing the Limits of ChatGPT on NLP Tasks." *CoRR*, **abs/2306.09719**.

Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., and Hill, F. (2021). "Multimodal Few-Shot Learning with Frozen Language Models." *CoRR*, **abs/2106.13884**.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., and Kurohashi, S. (2023). "GPT-RE: In-context Learning for Relation Extraction using Large Language Models." In Bouamor, H., Pino, J., and Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3534–3547, Singapore. Association for Computational Linguistics.

Wan, Z., Liu, Q., Mao, Z., Cheng, F., Kurohashi, S., and Li, J. (2022). "Rescue Implicit and Long-tail Cases: Nearest Neighbor Relation Extraction." *CoRR*, **abs/2210.11800**.

Wang, C., Liu, X., Chen, Z., Hong, H., Tang, J., and Song, D. (2022). "DeepStruct: Pretraining of Language Models for Structure Prediction." In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 803–823, Dublin, Ireland. Association for Computational Linguistics.

Wang, S., Yang, C.-H. H., Wu, J., and Zhang, C. (2024a). "Bayesian Example Selection Improves In-Context Learning for Speech, Text, and Visual Modalities." *CoRR*, **abs/2404.14716**.

Wang, S., Yang, C.-H. H., Wu, J., and Zhang, C. (2024b). "Can Whisper Perform Speech-based In-context Learning?" *CoRR*, **abs/2309.07081**.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., and Zhou, D. (2022). "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *CoRR*, **abs/2203.11171**.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. (2022). "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *CoRR*, **abs/2201.11903**.

Williams, A., Nangia, N., and Bowman, S. (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online. Association for Computational Linguistics.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. (2025). "Qwen2.5-Omni Technical Report." *CoRR*, **abs/2503.20215**.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017). "Position-aware Attention and Supervised Data Improve Slot Filling." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zhong, Z. and Chen, D. (2021). "A Frustratingly Easy Approach for Entity and Relation Ex-

traction." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 50–61, Online. Association for Computational Linguistics.

Zhou, Y., Li, X., Wang, Q., and Shen, J. (2024). "Visual In-Context Learning for Large Vision-Language Models." *CoRR*, **abs/2402.11574**.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models." *CoRR*, **abs/2304.10592**.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). "A Robustly Optimized BERT Pre-training Approach with Post-training." In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pp. 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Appendix

A Hyperparameters

A.1 GPT-3 Hyperparameters

We use the GPT-3 API during the experiments and set the hyperparameters as in Table 4. 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 5.

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 4 GPT-3 Hyperparamters.

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

Table 5 Search range for each dataset.

A.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* (Beltagy et al. 2019) as the base encoder for SciERC and *bert-base-uncased* (Devlin et al. 2019) for the remaining three general domain datasets. We follow hyper-parameters in their paper. We used 2 NVIDIA RTX3090 for training. As a further option, a widely used method contrastive learning for semi-supervised embedding training, to validate that supervised finetuining has been sufficient in our task, we implemented to train the retrieval model with contrastive leaning directly on the Semeval dataset and the the results show to be worse than PURE fine-tuning as the baseline (82.18 v.s. 89.90), thus we think contrastive learning alone may be a better selection in further demonstration retrieval.

A.3 Sentence Embedding Methods

Recent work (Gutiérrez et al. 2022) uses the [CLS] of RoBERTa-large as the representation in retrieval, (Liu et al. 2022b) fine-tunes RoBERTa-large on two natural language inference (NLI) datasets: SNLI (Bowman et al. 2015) and MultiNLI (Williams et al. 2018) 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`.

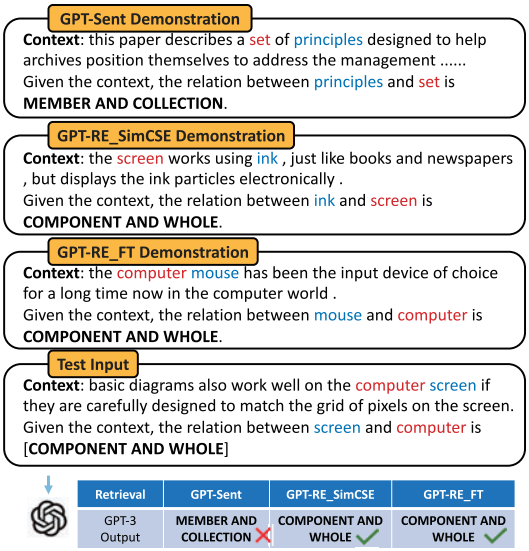
B Case Study

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

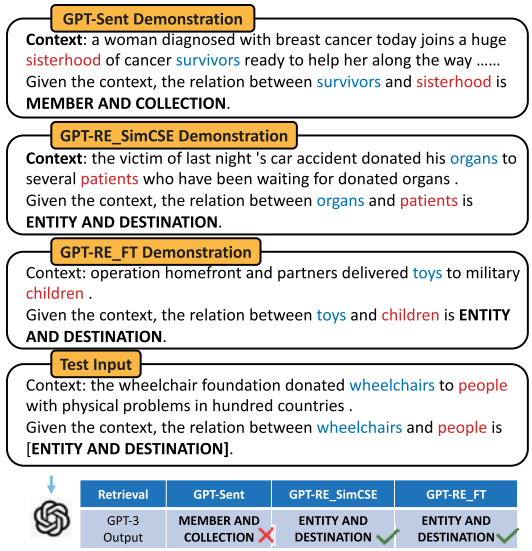
For Figure 10a, *GPT-Sent* retrieves a demonstration that shares the same semantic meaning of "design" with the test input. However, the entity pair is irrelevant to the concept "design" resulting in a noisy demonstration. Instead, *GPT-RE_SimCSE* retrieves a more relative demonstration with closer pair 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 10b shows a similar phenomenon.

C Subset

The number of sampled examples is not only related to the size of the training data itself. A more important factor is the proportion of NULL. We have to maintain the original label



(a) **[COMPONENT AND WHOLE]** denotes the gold label



(b) **[ENTITY AND DESTINATION]** denotes the gold label

Figure 10 More cases.

Label	# Num
PHYS	28
GEN-AFF	12
PER-SOC	11
GEN-AFF	33
PART-WHOLE	13
ART	19
NULL	2329

Table 6 ACE05

distribution in datasets with a high proportion of NULL. Thus, the rule to sample the subset is to keep the proportion of each relation label consistent with the original test set. Table 6 and Table 7 are label distributions of two subsets.

GPT-RE_FT on TACRED surpasses the supervised baseline in the current subset. As we show above, some labels in TACRED are indeed not well presented (only 1 example), since TACRED dataset contains some long-tail labels. We decided to add additional results of GPT-RE_FT by enlarging our sampled set to # 3200 (2 times the current version), and the performance of GPT-RE_FT ($k = 15$) is 73.16 while the performance of PURE is 70.48.

Label	# Num
Per:title	40
PER:city_of_death	1
Org:shareholders	2
Per:origin	12
Org:top_members/employees	36
Org:city_of_headquarters	11
Per:religion	4
Per:city_of_birth	1
Per:employee_of	27
Per:data_of_death	3
Per:other_family	5
Org:website	6
Per:cause_of_death	3
Org:subsidiaries	4
Org:stateorprovince_of_headquarters	5
Per:countries_of_residence	10
Per:siblings	5
Per:stateorprovinces_of_residence	11
Org:alternate_names	27
Per:spouse	4
Per:parents	7
Org:country_of_headquarters	9
Per:age	21
Per:date_of_birth	1
Per:country_of_death	1
Per:schools_attended	4
Org:member_of	3
Per:children	5
Org:parents	7
Per:cities_of_residence	24
Per:stateorprovince_of_brith	1
Per:charges	12
Org:founded	2
Org:country_founded_by	5
Per:stateorprovince_of_death	1
Org:members	4
Per:country_of_birth	1
Per:alternate_names	1
Org:number_of_employees/members	1
Org:dissolved	1
Org:political/religious_affiliation	1
NULL	1271

Table 7 TACRED

Zhen Wan: received his B.S. in Energy Engineering from Zhejiang University in 2018, and his M.S. in Informatics from Kyoto University in 2023. He is currently a Ph.D. student at Kyoto University. His research interests include natural language processing, in particular inference-time emergent abilities and agentic multimodal systems.

Fei Cheng: received his B.S. degree in Applied Physics from Donghua University in 2005 and his Ph.D. in Engineering from the Nara Institute of Science and Technology in 2018. His research interests include information extraction, numerical reasoning, large language models, and a broad range of natural language processing topics. He is currently a Program-specific Senior Lecturer at Kyoto University.

Sadao Kurohashi: received a PhD in Electrical Engineering from Kyoto University in 1994. Since 2023, he has served as the Director-General of the National Institute of Informatics, Japan. His research interests include natural language processing, knowledge infrastructure, and open science. He received the 2017 Commendation for Science and Technology from the Minister of Education, and he was named an ACL Fellow in 2025.

(Received August 1, 2025)
(Revised November 8, 2025)
(Accepted December 11, 2025)