Zero-Shot Relation Classification with Large Language Models: Binary Prompting and Relevance Elicitation

Anonymous ACL submission

1 Abstract

Relation classification (RC) transforms unstructured text into structured knowledge by identifying semantic relationships between entities. As natural language data volumes rapidly increase, manual knowledge extraction has become impractical. Existing zero-shot LLM-based approaches face limitations due to sparse RC task representation in instruction-tuning datasets and inherent biases toward certain relation types. To address these challenges, we propose Binary Relevance Elicitation (BRE), a true/false prompting framework that systematically queries an LLM for each candidate relation between two entities from a closed set of relation types. BRE incorporates relevance scoring to identify the most likely relation among those classified as "true," reducing false positive predictions. We evaluate BRE using multiple open-source LLMs and achieve state-of-the-art zero-shot RC accuracy across the TACRED family of datasets. Our results demonstrate performance improvements with newer model releases that are independent of parameter count, suggesting information representation advances within LLMs drive zero-shot RC capabilities.

2 Introduction

Social network analysis reveals how entities (people, organizations, locations) connect and how these connections influence social behavior and outcomes. Relation classification (RC) identifies semantic relationships between entities in sentences, converting unstructured text into structured data. These triplets enhance knowledge classification, improve information retrieval and question answering, and generate knowledge graphs (Zhao et al., 2024a; Nayak et al., 2021).

RC has achieved state-of-the-art performance by leveraging pre-trained language models (PLMs) to extract relation triplets (Park and Kim, 2021; Han et al., 2022; Zhou and Chen, 2022; Huguet Cabot et al., 2021). Yet a fundamental misalignment persists: top-performing PLMs rely on pre-training objectives like masked language modeling that fail to capture the structured semantic relationships between predefined entity pairs (Devlin, 2018; Liu, 2019). While some efforts have reframed RC as textual entailment or natural language inference (Obamuyide and Vlachos, 2018; Chanthran et al., 2026), PLM-based approaches remain critically dependent on large volumes of high-quality labeled training data (Huguet Cabot et al., 2021; Elsahar et al., 2018; Zhao et al., 2024a). Language ambiguity makes annotating semantic relations inherently challenging, producing incomplete or inconsistent ground-truth data that varies by annotator perspective and dataset domain. The absence of standardized, closed-set training and testing datasets represents a fundamental obstacle to RC advancement,

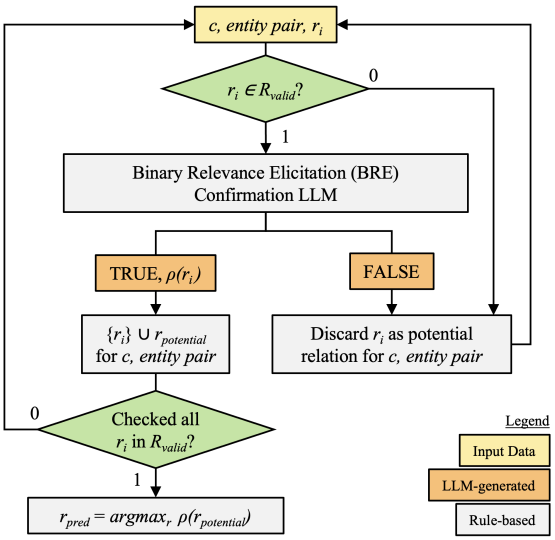

Figure 1: Zero-shot BRE prompting pipeline using the relevance elicitation measure (ρ) to select the relation for a given context (c) and entity pair. The predicted relation is the one with the highest relevance measure from the set of possible relations ($r_{pred} \in r_{potential} \subseteq R$).

making zero-shot approaches not merely desirable but essential for domain-specific applications.

Large language models (LLMs) have shown promise with minimal training data (Hang et al., 2025), yet they falter in critical scenarios: capturing implied knowledge in long-tail cases where entity relations must be globally inferred (Zhao et al., 2024b), and handling complex multi-relation sentences without handcrafted templates for each relation type (Li et al., 2024; Geroimenko, 2025; Aguda et al., 2025).

We introduce a zero-shot approach that fundamentally rethinks RC through true/false binary prompting to not rely on manual customization. We propose **relevance elicitation** to provide a quantitative ranking mechanism that identifies optimal relation types while directly addressing LLMs' systematic over-prediction bias. By treating each potential relation as an independent binary query with relevance-based selection, we harness LLM strengths while neutralizing their limitations. This approach mirrors the successful binary relevance filtering demonstrated by Zhuang et al. (Zhuang et al., 2024) for document ranking. Our method achieves state-of-the-art zero-shot RC performance across the TACRED family of datasets (Zhang et al., 2017) using various open-source LLMs.

3 Related Work

3.1 From Supervised to Zero-Shot Closed RC

RC methods vary in supervision requirements and output structures. Traditional supervised RC approaches using deep neural networks require substantial annotated training data, limiting their applicability to new domains or relation types. Unsupervised RC methods eliminate this annotation burden, primarily employing clustering-based techniques that extract entities via named-entity taggers and group entity pairs based on contextual similarity (Detroja et al., 2023). However, these methods produce open-ended clusters rather than assignments to predefined relation labels, making interpretation ambiguous and challenging to evaluate when sentences contain entity pairs with multiple relations. Zero-shot closed RC addresses these limitations by enabling users to specify a predefined set of relations of interest without requiring annotated training examples. This approach combines the flexibility of unsupervised methods (no annotation requirement) with the interpretability of supervised methods (explicit, user-defined label assignments), making it particularly suitable for domain-specific applications where relation types are known but labeled data is scarce.

3.2 Pre-Trained Language Models (PLMs)

PLMs have advanced RC by eliciting rich knowledge from large pre-trained models (Huguet Cabot et al., 2021; Tang et al., 2022; Sui et al., 2023). Models like BERT (Devlin, 2018), RoBERTa (Liu, 2019), and ELMo (Peters et al., 2018) produce contextualized representations where token meaning depends on surrounding context, enabling nuanced understanding of entity relationships within sentences. PLM-based RC approaches employ fine-tuning to adapt these models for specific RC tasks. However, standard pre-training objectives (e.g., masked language modeling, next-token prediction) do not naturally capture structured semantic relationships, creating reliance on task-specific training data (Hang et al., 2025). This creates two key limitations that zero-shot closed RC addresses: significant computational requirements for training and inference, and tendency to overfit on domain-specific datasets (Detroja et al., 2023).

3.3 Large Language Models (LLMs)

LLMs have transformed zero-shot RC through two primary approaches: direct RC and synthetic data generation for training downstream models. For direct classification, prompt engineering approaches typically require no additional fine-tuning, simplifying deployment and eliminating the need for annotated few-shot data (Chia et al., 2022; Li et al., 2024). Alternatively, LLMs can generate synthetic training data as supplementary samples for unseen relations (Xu et al., 2023). However, effective prompt engineering for direct classification requires considerable subject matter expertise specific to the RC-employed domain (Geroimenko, 2025), and out-of-the-box LLM performance is limited by general language understanding capabilities and training biases (Aguda et al., 2025; Dai et al., 2024).

Recent efforts show LLM performance varies significantly with prompt templates (Liu et al., 2023; Zhang et al., 2023a; Xu et al., 2024). To address this, approaches have incorporated additional context (Wan et al., 2023), entity type information (Zhang et al., 2023a), multi-step prompting for chain-of-thought control (Layegh et al., 2024; Hannah et al., 2025), and logical searches for discrete prompts to avoid manual design (Xu et al.,

2024). A common observation is that task simplification correlates with improved LLM performance (Zhou et al.; Khot et al.). Our approach uniquely simplifies RC to a binary decision and adapts the LLM-as-a-judge framework (Gu et al., 2024) to eliminate domain-specific prompt-tuning while improving performance.

4 Methodology

Our Binary Relevance Elicitation (BRE) method, described in Algorithm 1, simplifies RC to a series of binary decisions while enabling principled relation selection through relevance scoring. For each entity pair (e_h, e_t) in context c, BRE iterates through type-compatible relations from a predefined set R, prompting the LLM to evaluate whether each relation r_i holds and rate its relevance ρ to the context on a 0-10 scale. The prompt provides the relation definition d_i and asks the LLM to assess the statement "e_h and e_t are somehow connected through the relationship of 'r_i'."

The binary classification with numerical relevance rating is critical: while the binary decision identifies plausible relations, the relevance score ρ_i enables ranking when multiple relations are deemed true. BRE returns the highest-scoring relation from all 'true' classifications (a tie-breaker when possible[1]), or 'no relation' if none qualify. This relevance elicitation mechanism directly addresses the multi-relation challenge in RC by providing a principled, interpretable basis for selecting the most contextually appropriate relation without requiring domain-specific prompt engineering or few-shot examples. The prompt template can be found in the Appendix A.1.

4.1 Closed RC Definition

The RC task considers a given sentence c, a set of entities $E \subseteq c$, entity pairs $(e_h, e_t) \in E \times E$ where $e_h \neq e_t$, and a set of relation types R. The task identifies the semantic relations between entity pairs in c. The RC output is a set of sentence-level triplets $\langle e_h, r_{pred}, e_t \rangle$ where $r_{pred} \in R \cup \{\text{'no relation'}\}$. Zero-shot RC employs no fine-tuning on the target relations, and no examples are included in the prompt. Zero-shot RC performance varies substantially across LLMs even under identical task framing. By holding the dataset and prompt fixed, this work isolates and evaluates these cross-model differences. Figure 1 shows the flow to identify the predicted relation for a valid entity pair.

[1]If two relations are assigned the same relevance score, the selected relation is randomly chosen for this evaluation.

Algorithm 1 Binary Relevance Elicitation (BRE)

Require: sentence c, entity pair (e_h, e_t), entity types (e_h^{type}, e_t^{type}), relation set R, valid type pairs T_r, definitions D, LLM θ
$R_{valid} \leftarrow \{r \in R \mid (e_h^{type}, e_t^{type}) \in \mathcal{T}_r\}$
$r_{pot} \leftarrow \emptyset$ ▷ Store LLM-selected relations
for $r_i \in R_{valid}$ **do**
  $prompt \leftarrow$ construct_prompt$(c, e_h, r_i, e_t, D(r_i))$
  $(y_i, \rho_i) \leftarrow \theta(prompt)$
  ▷ $y_i \in \{\text{TRUE}, \text{FALSE}\}, \rho_i \in [0, 10]$
  if $y_i =$ TRUE **then**
    $r_{pot} \leftarrow r_{pot} \cup \{(r_i, \rho_i)\}$
  end if
end for
return $|r_{pot}| = 0$? 'no relation' : $argmax_{(r,\rho) \in r_{pot}} \rho$

4.2 Entity-Type Aware Querying

Entity type constraints critically reduce the search space of potential relation triplets. By defining valid entity type pairs (e_h^{type}, e_t^{type}) for each relation, we systematically filter candidate relations before LLM inference. Specifically,

$$R_{valid} = \{r \in R \mid (e_h^{type}, e_t^{type}) \in T_r\}$$

where T_r is the set of valid type pairs for relation r as defined by TACRED dataset annotation guidelines (Ellis et al., 2015). This leverages domain knowledge to eliminate semantically invalid combinations (e.g., testing "founded by" for two location entities) (Zhang et al., 2023a). This optional type-aware filtering improves computational efficiency and extraction precision by focusing on semantically plausible candidates, yielding more reliable and cost-effective RC systems.

4.3 Relation Definition Context

Including relation definitions in prompts serves as a disambiguation mechanism (Zhou et al., 2024). Natural language descriptions establish clear semantic boundaries for interpreting relation types and contextual evidence. This is particularly valuable for ambiguous or overlapping relations (e.g., distinguishing "employed by" from "affiliated with"), where definitions provide the conceptual framework for accurate classification. This approach reduces semantic confusion and improves extraction accuracy by grounding reasoning in explicit definitions, ensuring consistent relation interpretation across varying contexts rather than relying on context-dependent implicit associations.

Model	Huggingface Model ID	Parameter Size (B)	Release Date
Llama-3.2-3B	meta-llama/Llama-3.2-3B-Instruct	3	09/25/2024
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct	8	07/23/2024
Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct	70	07/23/2024
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct	70	12/06/2024
Llama-4-Scout-17B-16E	meta-llama/Llama-4-Scout-17B-16E-Instruct	109	04/05/2025
Mistral-7B	mistralai/Mistral-7B-Instruct-v0.2	7	12/25/2023
Mixtral-8x7B	mistralai/Mixtral-8x7B-Instruct-v0.1	56	12/10/2023
Mixtral-8x22B	mistralai/Mixtral-8x22B-Instruct-v0.1	176	04/17/2024
Qwen2.5-VL-32B	Qwen/Qwen2.5-VL-32B-Instruct	32	02/19/2025
DeepSeek-R1-Distill-Qwen-32B	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	32	02/23/2025

Table 1: Evaluated LLM version details including their corresponding Huggingface model IDs, their parameter sizes, and release dates in the format MM/DD/YYYY.

Specifically, we define $D = \{r_i : d_i\}$ such that d_i is the definition for r_i as described in the TACRED family dataset annotation guidelines (Ellis et al., 2015; Stoica et al., 2021).

5 Evaluation

5.1 Experimental Setup

LLM Models We investigate open source LLMs from different organizations such as Llama, Mistral, Qwen and DeepSeek models, detailed in Table 1. We explore models with varying parameter counts to observe correlations between parameter size, LLM organization class, and model release date to the RC task. Lastly, the exact same prompt template was used for all LLMs under evaluation.

Computational Usage We implemented our approach using VLLM on NVIDIA H100 GPUs, scaling resources by model size (up to four H100s for the largest model). We performed five inference runs per model and report averaged results to account for stochastic variation.

NoTA Evaluation LLM-based RC systems exhibit a critical bias: they over-assign named relations to entity pairs rather than predicting "no relation" or the none-of-the-above (NoTA) class, even when no valid relationship exists (Wan et al., 2023; Ma et al.; Zhang et al., 2023b). This bias is amplified by the asymmetric nature of many relations, which requires evaluating numerous ordered entity pairs with potentially no predefined relationship (Qin et al., 2021; Baek and Choi, 2022). The resulting severe class imbalance increases false positives and reduces precision. Our relevance elicitation method mitigates this by ranking potential false positives through comparative relevance scoring across candidate relations. Gutierrez et al. (Gutierrez et al., 2022) demonstrated that early LLM underperformance on RC tasks was largely due to poor NoTA handling. To evaluate zero-shot methods, we report both NoTA-included and NoTA-excluded metrics for comparison with prior work.

Parsing LLM Outputs All evaluated LLM outputs except Llama-4-Scout-17B-16E followed the specified format: starting the answer with "true" or "false" followed by the numerical relevance score, enabling uniform parsing. Llama-4-Scout-17B-16E consistently placed its binary output and relevance score at the end of its response, requiring a modified parsing script to extract the final answer. Complete parsing scripts are provided in Appendix A.2.

5.2 Baselines

Few-Shot Baselines ODDA (Zhong et al., 2025) employs data augmentation to iteratively generate few-shot examples based on the RE model's predictions for target relations, then produces diverse training data using attribute-based constraints. ConsistRE (Zheng et al., 2024) uses relation-specific keywords as generation constraints to encourage relation consistency. However, ConsistRE requires access to a sufficiently informative external corpus for keyword extraction, an assumption that may not be feasible in all scenarios. It is worth noting that both approaches use DeepSeek-v3 (671B parameters) along with its PLM backbone.

Zero-Shot Baselines QA4RE (Zhang et al., 2023a) reformulates RC as multiple-choice question-answering by incorporating valid entity pairs into pre-defined relation templates as answer options. SUMASK (Li et al., 2023) takes a different approach, combining LLMs with natural language inference and chain-of-thought reasoning to sequentially query whether each candidate relation exists between entity pairs through iter-

		macro-average			weighted-average			micro-F_1
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Accuracy
Re-TACRED	Llama-3.2-3B	17.5	41.3	20.9	58.3	46.2	48.9	42.1
	Llama-3.1-8B	22.1	55.8	28.4	64.3	51.4	50.5	44.0
	Llama-3.1-70B	61.1	53.3	54.1	80.5	76.0	75.8	79.1
	Llama-3.3-70B	58.5	61.3	**57.3**	73.1	72.3	71.3	**79.5**
	Llama-4-Scout-17B-16E	57.6	72.7	**61.5**	86.4	83.3	**83.9**	**83.2**
	Mistral-7B	38.1	56.5	40.6	67.0	63.1	63.4	65.7
	Mixtral-8x7B	47.0	60.0	47.8	75.1	73.3	72.9	72.4
	Mixtral-8x22B	50.7	62.0	52.7	81.8	77.7	**78.2**	79.4
	Qwen2.5-VL-32B	69.6	64.9	**63.0**	87.4	82.5	**83.3**	**85.6**
	DeepSeek-R1-Distill-Qwen-32B	44.2	61.5	48.5	76.3	74.3	74.3	74.5
TACREV	Llama-3.2-3B	21.0	45.8	24.2	75.7	54.6	61.3	49.7
	Llama-3.1-8B	22.0	65.3	28.8	79.3	52.8	59.1	48.1
	Llama-3.1-70B	57.2	59.8	54.6	85.7	87.5	86.1	85.8
	Llama-3.3-70B	56.6	69.7	**59.6**	87.9	88.2	**87.8**	**86.4**
	Llama-4-Scout-17B-16E	49.2	76.5	**56.8**	90.8	88.3	**88.8**	**86.7**
	Mistral-7B	38.2	68.4	45.4	84.7	74.7	78.0	73.5
	Mixtral-8x7B	46.7	70.4	53.1	86.3	80.5	82.6	78.7
	Mixtral-8x22B	47.3	71.6	54.4	88.3	87.8	87.6	85.9
	Qwen2.5-VL-32B	62.2	72.9	**63.9**	91.7	90.6	**90.6**	**90.0**
	DeepSeek-R1-Distill-Qwen-32B	42.3	70.1	49.3	85.7	83.7	84.1	80.8
TACRED	Llama-3.2-3B	22.8	46.0	25.5	75.2	54.3	61.2	50.0
	Llama-3.1-8B	23.5	59.6	30.1	78.1	51.9	58.6	48.1
	Llama-3.1-70B	54.8	50.1	50.2	84.0	84.8	83.9	83.4
	Llama-3.3-70B	52.9	59.1	**54.2**	86.7	84.7	**85.5**	**83.9**
	Llama-4-Scout-17B-16E	49.8	64.0	**54.1**	88.3	85.2	**86.3**	**84.4**
	Mistral-7B	40.8	61.6	44.9	83.3	73.2	76.9	72.4
	Mixtral-8x7B	47.7	59.8	49.9	84.4	79.0	81.1	77.7
	Mixtral-8x22B	47.5	61.2	51.5	86.4	84.2	85.0	83.0
	Qwen2.5-VL-32B	59.4	61.2	**57.6**	88.5	86.8	**87.3**	**86.8**
	DeepSeek-R1-Distill-Qwen-32B	43.1	60.6	47.6	83.8	81.5	82.2	79.3

Table 2: NoTA-included zero-shot RC performance on the TACRED family of datasets. The **bolded** scores indicate the top three performing F_1 scores and the **underlined** score is the highest for that category.

ative decomposition. REA (Layegh et al., 2024) employs a more elaborate four-stage prompting strategy: identifying entity types, filtering relation candidates by entity type compatibility, generating hypothesis sentences with confidence scores for each candidate, and selecting the final relation based on accumulated evidence.

5.3 Datasets

TACRED (Zhang et al., 2017) is a widely-used dataset for sentence-level relation extraction with 42 relation types, containing annotated news articles and web documents with entities and predefined relations between persons and organizations for knowledge base construction. Updated versions include TACREV (Alt et al., 2020), which corrects systematic labeling errors (particularly false "no relation" labels), and Re-TACRED (Stoica et al., 2021), which employed improved crowd-sourced annotation and relabeled 23.9% of the original dataset for enhanced accuracy. Evaluating all three datasets (entire test sets) demonstrates our methodology's robustness to annotation inconsistencies.

6 Results

6.1 Relation Classification Performance

Qwen2.5-VL-32B consistently shows high RC performance due to its advanced attention mechanisms and multimodal pre-training which enhance structured relationship understanding (Wang et al., 2024). MoE architectures (e.g., Llama-4-Scout and Mixtral) with specialized tuning showed high accuracy with balanced classification performance (Shen et al., 2023). Both represent architectural innovations beyond standard decoder-only transformers, specifically beneficial for tasks requiring precise semantic relationship extraction.

Observation: DeepSeek-R1-Distill-Qwen-32B showed inferior expected performance suggesting distillation disproportionately affected semantic relationship identification.

6.2 Baseline Comparisons

BRE prompting demonstrates substantial improvements over existing zero-shot baselines across all

	Re-TACRED	TACREV	TACRED
Qwen2.5-VL-32B	**74.7**	**82.4**	**73.9**
Llama-4-Scout-17B-16E	**78.8**	**86.3**	**78.8**
Llama-3.3-70B	65.6	78.9	**73.9**
Mixtral-8x22B	69.0	79.4	70.7
Llama-3.1-70B	62.1	68.3	62.7
DeepSeek-R1-Distill-Qwen-32B	66.7	75.7	68.5
Mixtral-8x7B	65.9	71.3	68.0
REA + Mixtral-8x7B	-	-	58.0
REA + GPT-3.5-Turbo	-	-	63.5
QA4RE + FLAN-T5 XXL	67.3	54.3	54.5
QA4RE + GPT-3.5-Turbo	61.2	59.4	59.4
SUMASK + GPT-3.5-Turbo	-	-	57.7

Table 3: NoTA-excluded comparison to state-of-the-art zero-shot RC methods. The **bolded** scores indicate the top two micro-F_1 scores and the **underlined** score is the highest.

	Re-TACRED	TACREV	TACRED
Qwen2.5-VL-32B	**85.6**	**90.0**	**86.8**
Llama-4-Scout-17B-16E	83.2	**86.7**	**84.4**
Llama-3.3-70B	79.5	86.4	83.9
Mixtral-8x22B	79.4	85.9	83.0
Llama-3.1-70B	79.1	85.8	83.4
DeepSeek-R1-Distill-Qwen-32B	74.5	80.8	79.3
Mixtral-8x7B	72.4	78.7	77.7
SUMASK + GPT-3.5-Turbo	73.8	75.1	79.6
ODDA/TYP Marker (8-shot)	76.8	-	75.8
ODDA/KnowPrompt (8-shot)	**84.1**	-	81.0
ConsistRE/TYP Marker (8-shot)	71.8	-	72.6
ConsistRE/KnowPrompt (8-shot)	75.8	-	80.1

Table 4: NoTA-included comparison to state-of-the-art zero- and few-shot RC methods. All approaches in this table are zero-shot unless otherwise specified. The **bolded** scores indicate the top two micro-F_1 scores and the **underlined** score is the highest.

evaluated datasets[2]. Most notably, BRE outperforms REA (Layegh et al., 2024) even when both methods employ the same foundational Mixtral 8x7B model, achieving 17% relative increased accuracy on TACRED. The consistent gains across datasets highlights that our approach effectively addresses the over-prediction bias inherent in LLM-based RC.

Using Qwen2.5-VL-32B and Llama-4-Scout, BRE matches or surpasses the performance of state-of-the-art few-shot baselines which require labeled

[2]Reported baseline performance numbers are from original efforts. Missing reported values were not included in the original work.

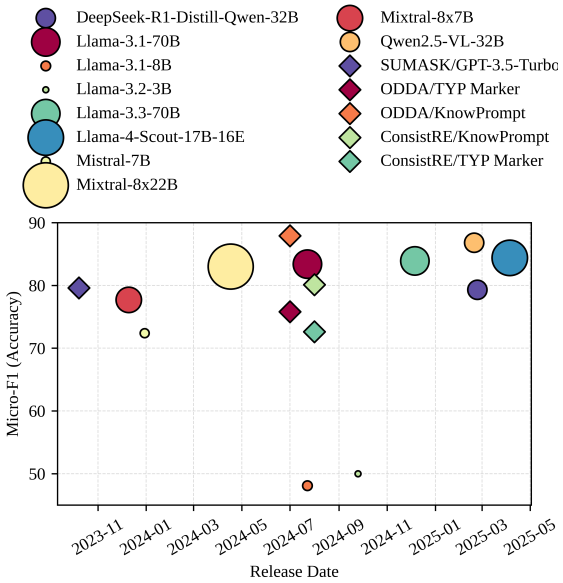

Figure 2: NoTA-included micro-F_1 comparison to state-of-the-art zero-shot RC methods on TACRED dataset.

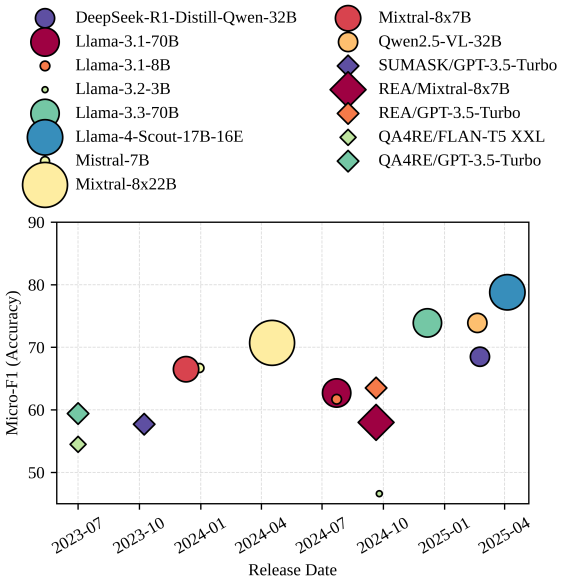

Figure 3: NoTA-excluded micro-F_1 comparison to state-of-the-art zero-shot RC methods on TACRED dataset.

data or iterative data augmentation. However, analysis of NoTA-excluded metrics reveals an important distinction between our top-performing models: Llama-4-Scout outperforms Qwen2.5-VL-32B in NoTA-excluded scenarios.

Observation: Qwen2.5-VL-32B's high overall scores stem largely from strong NoTA classification, while Llama-4-Scout maintains more balanced performance.

In Figures 2 and 3, circles (○) represent BRE, and diamonds (◇) represent the baseline methods.

The size of the shape indicates the parameter size in the foundational model[3]. The baseline release dates included are those of the research publication. These figures show the positive correlation of RC task performance over time, especially in Figure 3 without the performance inflation provided by the NoTA relations.

Observation: Better natural language representations through techniques like extended context (Chen et al., 2023) and enhanced reasoning (Wang et al., 2024) improve RC performance independent of model size.

6.3 Relation-Specific Analysis

Qwen2.5-VL-32B and Llama-4-Scout achieved the highest accuracy across all three datasets when paired with BRE prompting. However, the discrepancy between weighted-average and macro-average scores indicates that performance varies significantly across different relation types. Figure 4 presents per-relation F_1 scores on the TACRED dataset, revealing that all models exhibit consistently low performance on the same subset of relations regardless of overall accuracy differences. Similar heatmaps for the TACREV and Re-TACRED dataset are included in Appendix A.3.

Relations such as 'members of', 'members,' and 'shareholders' prove particularly difficult, likely due to semantic overlap between relation definitions. When multiple relations share similar meanings, the relevance elicitation process may assign comparable scores across overlapping categories for a given entity pair, making correct distinction difficult. Notably, BRE's relevance elicitation identifies these challenging cases by providing interpretable scores, enabling more transparent failure analysis compared to baseline approaches.

Observation: Relevance elicitation enables transparent identification of challenging relation types, particularly those with semantic overlap that confound classification.

6.4 Ablation Study

Impact of Relation Definitions The impact of relation-specific challenges are further evidenced in our ablation study, where removing relation definitions resulted in an average relative accuracy degradation of 4% across the TACRED family of datasets (p-value $= 0.01 < 0.05$). Including relation definitions serves as a semantic anchor grounds the binary classification task and provides explicit criteria for what constitutes a valid relationship beyond mere label matching. It also contributes more context regarding what a potential relation may include in scenarios where the sentence or relation context are ambiguous.

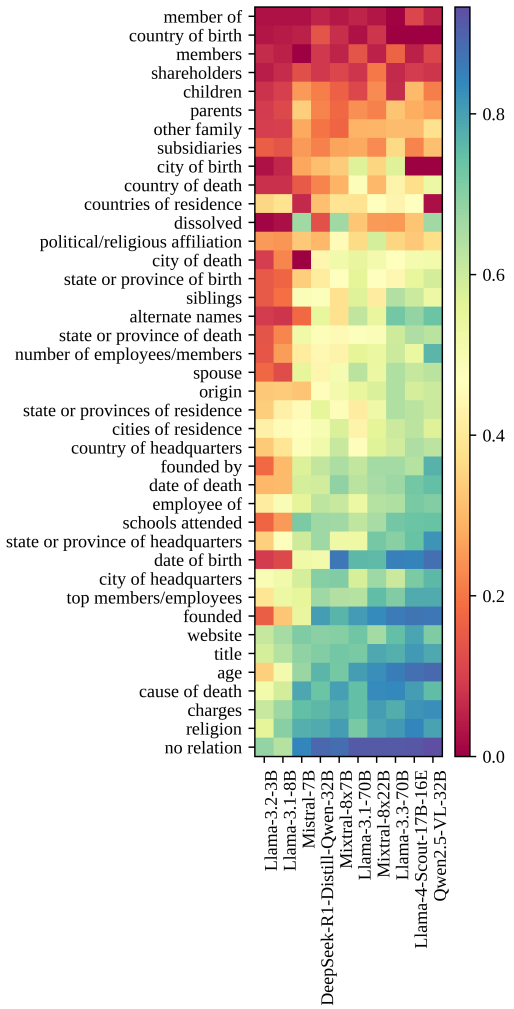

Figure 4: F_1 scores for the varying relation types in the TACRED dataset. The relations are sorted based on their F_1 and the models are sorted based on their macro-F_1 scores.

Impact of Entity-Type Awareness Without entity-type awareness, the model must evaluate all relations for each entity pair regardless of semantic plausibility. Naturally, precision suffers from false positives as the expanded search space introduces implausible relation candidates that dilute relevance elicitation. However, TACRED's single-relation constraint implies that BRE always selects

[3]Figure 2 baseline parameter sizes are not scaled since ODDA and ConsistRE with the DeepSeek-v3 (671B parameters) backbones were too large to scale accordingly.

		macro-average			weighted-average			micro-F_1
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Accuracy
Re-TACRED	Qwen2.5-VL-32B	66.6	57.4	56.4 (-6.6)	85.2	77.4	78.2 (-5.1)	82.7 (-2.9)
	Llama-4-Scout-17B-16E	50.0	64.9	51.4 (-10.1)	83.4	77.5	78.9 (-5.0)	77.8 (-5.4)
	Llama-3.3-70B	54.3	65.9	53.1 (-4.2)	82.8	73.1	74.7 (+3.4)	76.9 (-2.6)
	Mixtral-8x22B	45.2	60.9	47.4 (-5.3)	78.3	71.9	72.8 (-5.4)	74.8 (-4.6)
	Llama-3.1-70B	45.9	59.4	46.8 (-7.3)	79.7	73.4	74.2 (-1.6)	74.5 (-4.6)
TACREV	Qwen2.5-VL-32B	62.5	69.9	62.0 (-1.9)	90.5	89.7	89.5 (-1.1)	89.2 (-0.8)
	Llama-4-Scout-17B-16E	45.9	74.6	52.9 (-3.9)	89.5	85.0	86.1 (-2.7)	83.0 (-3.7)
	Llama-3.3-70B	50.7	69.1	54.8 (-4.8)	88.0	85.6	86.2 (-1.6)	84.0 (-2.4)
	Mixtral-8x22B	41.7	72.7	49.9 (-4.5)	88.1	83.5	84.9 (-2.7)	81.7 (-4.2)
	Llama-3.1-70B	44.5	69.0	49.9 (-4.7)	85.5	83.7	83.9 (-2.2)	80.8 (-5.0)
TACRED	Qwen2.5-VL-32B	59.9	56.4	54.5 (-3.1)	87.8	86.1	86.2 (-1.1)	86.1 (-0.7)
	Llama-4-Scout-17B-16E	44.9	62.6	49.3 (-4.8)	87.1	82.2	83.8 (-2.5)	81.0 (-3.4)
	Llama-3.3-70B	51.0	62.1	53.0 (-1.2)	87.1	82.7	84.2 (-1.3)	81.8 (-2.1)
	Mixtral-8x22B	43.7	65.6	49.9 (-1.6)	86.1	81.9	83.4 (-1.6)	80.6 (-2.4)
	Llama-3.1-70B	44.9	60.7	48.7 (-1.5)	84.4	81.3	82.3 (-1.6)	79.1 (-4.3)
	Average Δ	**-5.51**	**+0.73**	**-4.37**	**-0.27**	**-2.99**	**-2.14**	**-3.27**

Table 5: Ablation 1 - Excluding relation definitions during prompting

		macro-average			weighted-average			micro-F_1
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Accuracy
Re-TACRED	Qwen2.5-VL-32B	46.8	8.4	12.0 (-51.0)	68.6	61.0	54.6 (-28.7)	71.0 (-14.6)
	Llama-4-Scout-17B-16E	35.0	9.4	12.7 (-48.8)	62.0	60.2	54.1 (-29.8)	69.5 (-13.7)
	Llama-3.3-70B	41.7	8.9	12.3 (-45.0)	63.2	59.7	53.0 (-18.3)	69.9 (-9.6)
	Mixtral-8x22B	21.4	6.8	8.6 (-44.1)	57.1	58.9	52.1 (-26.1)	69.0 (-10.4)
	Meta-Llama-3.1-70B	34.3	6.8	9.0 (-45.1)	60.1	59.5	52.3 (-23.5)	69.8 (-9.3)
TACREV	Qwen2.5-VL-32B	34.7	8.4	11.6 (-52.3)	75.5	80.8	74.9 (-15.7)	81.9 (-8.1)
	Llama-4-Scout-17B-16E	24.5	7.9	9.8 (-47.0)	73.2	79.6	74.1 (-14.7)	80.3 (-6.4)
	Llama-3.3-70B	30.2	7.6	10.0 (-49.6)	75.5	80.2	74.6 (-13.2)	81.1 (-5.3)
	Mixtral-8x22B	17.5	6.5	7.9 (-46.5)	70.9	79.3	73.3 (-14.3)	80.4 (-5.5)
	Llama-3.1-70B	28.2	6.9	9.2 (-45.4)	74.2	80.0	74.1 (-12.0)	81.0 (-4.8)
TACRED	Qwen2.5-VL-32B	33.4	7.3	10.0 (-47.6)	74.2	79.0	72.9 (-14.4)	79.9 (-6.9)
	Llama-4-Scout-17B-16E	24.8	6.6	8.7 (-45.4)	71.7	78.2	72.3 (-14.0)	78.6 (-5.8)
	Llama-3.3-70B	29.8	6.5	8.8 (-45.4)	74.3	78.6	72.7 (-12.8)	79.3 (-4.6)
	Mixtral-8x22B	16.2	5.9	6.9 (-44.6)	69.0	77.7	71.5 (-13.5)	78.4 (-4.6)
	Llama-3.1-70B	22.9	6.3	7.7 (-42.5)	71.8	78.5	72.2 (-11.7)	79.2 (-4.2)
	Average Δ	**-26.20**	**-56.67**	**-46.69**	**-16.41**	**-11.25**	**-17.51**	**-7.59**

Table 6: Ablation 2 - Excluding entity-type aware filtering such that all relations were evaluated for each entity pair

one final relation, regardless of how many relations are equally relevant to a sentence. Thus, BRE randomly selects a relation among the highest scores, and with diluted relevance scores, this random selection significantly reduces recall.

7 Conclusion and Future Work

We introduce Binary Relevance Elicitation (BRE), a zero-shot RC approach that addresses systematic over-prediction bias in LLM-based RC through simplified binary prompting and relevance-based filtering. BRE decomposes multi-class classification into independent binary decisions with quantitative relevance scoring, achieving state-of-the-art zero-shot performance on the TACRED family of datasets and matching or exceeding few-shot baselines. Notably, our results reveal that performance improvements stem from architectural advances in newer models rather than increased parameter counts, suggesting that enhanced semantic representations within LLMs are the primary drivers of zero-shot relation classification capability.

While BRE substantially improves zero-shot RC performance and provides a robust framework for NoTA handling, performance variation across relation types remains an open challenge. Future work should explore relation-specific solutions and hybrid approaches combining BRE's filtering strengths with retrieval-augmented generation for ambiguous sentence-level cases. Despite these limitations, our work demonstrates that binary decomposition and systematic relevance filtering represent a significant advancement toward more reliable and interpretable zero-shot RC.

Limitations

BRE achieves state-of-the-art RC performance when dataset schemas provide well-documented relation definitions and entity type specifications. Since relation definitions are incorporated directly into the prompt and valid relations are filtered based on entity types, incomplete or inconsistent metadata may reduce prompt quality and system performance, as demonstrated in our ablation studies.

Additionally, our sentence-level evaluation strategy introduces substantial computational costs. For each entity pair in a sentence, BRE queries each valid relation type individually, resulting in worst-case complexity of $\mathcal{O}(|C| \times |E| \times |R|)$, where C represents the set of sentences, E the set of entity pairs, and R the set of candidate relations. While this granular approach enables interpretable relevance scoring, it may become prohibitive for large-scale deployment without optimization strategies such as batching or parallel processing.

Ethical Considerations

Our approach leverages LLM reasoning capabilities but inherits biases from their pre-training corpora, which predominantly reflect internet content from Western, English-speaking populations (Bender et al., 2021). These biases produce systematic performance disparities across languages, domains, and demographic groups (Navigli et al., 2023). Since RC commonly populates knowledge bases for downstream applications, unmitigated biases may propagate through entire pipelines.

References

Toyin Aguda, Erik Wilson, Allan Anzagira, Simerjot Kaur, and Charese Smiley. 2025. Conservative bias in large language models: Measuring relation predictions. *arXiv preprint arXiv:2506.08120*.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Tacred revisited: A thorough evaluation of the tacred relation extraction task. *arXiv preprint arXiv:2004.14855*.

Hyeong-Ryeol Baek and Yong-Suk Choi. 2022. Enhancing targeted minority class prediction in sentence-level relation extraction. *Sensors*, 22(13):4911.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Mohan Raj Chanthran, Soon Lay Ki, Ong Huey Fang, and Bhawani Selvaretnam. 2026. Document-level zero-shot relation extraction with entity side information. *arXiv preprint arXiv:2601.07271*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *Findings of the Association for Computational Linguistics: ACL 2022*.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

Kartik Detroja, CK Bhensdadia, and Brijesh S Bhatt. 2023. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joe Ellis, Jeremy Getman, Heather Simpson, Kira Griffitt, Hoa Trang Dang, Ralph Grishman, Heng Ji, Catherine DePrince, Thomas Riese, and Neil Kuster. 2015. TAC KBP 2015 Slot Descriptions. Technical report, Linguistic Data Consortium. Version 1.0.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vladimir Geroimenko. 2025. Key challenges in prompt engineering. In *The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks*, pages 85–102. Springer.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *The Innovation*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. 2022. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3170–3185.

Tingting Hang, Shuting Liu, Jun Feng, Hamza Djigal, and Jun Huang. 2025. Few-shot relation extraction based on prompt learning: A taxonomy, survey, challenges and future directions. *ACM Computing Surveys*, 58(2):1–38.

George Hannah, Jacopo de Berardinis, Terry R Payne, Valentina Tamma, Andrew Mitchell, Ellen Piercy, Ewan Johnson, Andrew Ng, Harry Rostron, and Boris Konev. 2025. Relrae: Llm-based relationship extraction, labelling, refinement, and evaluation. *arXiv preprint arXiv:2507.03829*.

Pere-Lluís Huguet Cabot, Roberto Navigli, et al. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics. EMNLP 2021*, pages 2370–2381. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Amirhossein Layegh, Amir H Payberah, and Mihhail Matskin. 2024. Rea: Refine-estimate-answer prompting for zero-shot relation extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 301–316. Springer.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892.

Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024. Meta in-context learning makes large language models better zero and few-shot relation extractors. *arXiv preprint arXiv:2404.17807*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yubo Ma, Yixin Cao, Yong Ching Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Tapas Nayak, Navonil Majumder, Pawan Goyal, and Soujanya Poria. 2021. Deep neural approaches to relation triplets extraction: A comprehensive survey. *Cognitive Computation*, 13(5):1215–1232.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pages 72–78.

Seongsik Park and Harksoo Kim. 2021. Improving sentence-level relation extraction through curriculum learning. *arXiv preprint arXiv:2107.09332*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Yongbin Qin, Weizhe Yang, Kai Wang, Ruizhang Huang, Feng Tian, Shaolin Ao, and Yanping Chen. 2021. Entity relation extraction based on entity indicators. *Symmetry*, 13(4):539.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13843–13850.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE transactions on neural networks and learning systems*, 35(9):12784–12795.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. Unirel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7087–7099.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 3534–3547.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynre: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207.

Liang Xu, Xiaoxuan Bu, and Xuetao Tian. 2024. Dynamic Prompt-Driven Zero-Shot Relation Extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2900–2912.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023a. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the association for computational linguistics: ACL 2023*, pages 794–812.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023b. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024a. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024b. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.

Yifan Zheng, Wenjun Ke, Qi Liu, Yuting Yang, Ruizhuo Zhao, Dacheng Feng, Jianwei Zhang, and Zhi Fang. 2024. Making llms as fine-grained relation extraction data augmentor. In *IJCAI*, pages 6660–6668.

Yijie Zhong, Yunfan Gao, Xiaolian Zhang, and Haofen Wang. 2025. Odda: An ooda-driven diverse data augmentation framework for low-resource relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 267–285.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Sizhe Zhou, Yu Meng, Bowen Jin, and Jiawei Han. 2024. Grasping the Essentials: Tailoring Large Language Models for Zero-Shot Relation Extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13462–13486. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (volume 2: short papers)*, pages 358–370.

A Appendix

A.1 BRE Prompt Template

Input Prompt: Using only the context text provided, label the provided statements 'True' or 'False' followed by a numerical rating representing how relevant the statement is to the main idea in the context text.

Answer 'True' if the statement is explicitly stated in the context and meets the definition of the relationship or if the numerical rating is higher than 5. The rating is an integer ranging from 0 to 10 where 0 implies no relevance and 10 implies the statement is highly relevant to the context text.

Fill in the answer below:
context: c
statement: e_h and e_t are somehow connected through the relationship of 'r_i'.
relationship definition: d_i

LLM Output Example: True, 10. The statement is supported in the provided context because...

Figure 5: Generation prompt template not including output formatting instructions for brevity.

A.2 Parsing Scripts

All models were evaluated using standardized Python scripts that parse LLM outputs to extract binary predictions and relevance scores. While most models adhered to the specified output format placing answers at the beginning of responses, Llama-4-Scout exhibited poor formatting behavior, frequently positioning final answers at the end of its responses despite explicit instructions. To ensure fair comparison, our parsing scripts were adapted to handle this variation by searching the entire response for the required information. Figures 6 and 7 contain the script for all LLMs (excluding Llama-4-Scout) and then the one for Llama-4-Scout, respectively.

For relevance elicitation scoring included in Figure 8, when a numeric score could not be extracted from the model output, we conservatively retained the relation as a potential candidate with a default score of 10 (the maximum). This default assignment strategy may contribute to both false positives through over-inclusion of uncertain relations (reducing precision) and missed correct relations when multiple candidates receive the default score and random selection chooses incorrectly (reducing recall). Despite these parsing challenges, the evaluation methodology remained consistent across all models, ensuring that performance differences reflect genuine model capabilities rather than evaluation artifacts.

A.3 Re-TACRED and TACREV Relation-Specific Results

Figures 9 and 10 reveal a striking pattern: all evaluated models exhibit consistently low performance on approximately the same subset of relations across both Re-TACRED and TACREV datasets. This shared weakness pattern persists regardless of overall model accuracy differences or architectural variations, indicating that certain relations pose fundamental challenges beyond what model selection or prompting strategies can address. The consistency of this pattern across two independently revised versions of TACRED (which differ in their annotation corrections and label distributions) further confirms that these difficulties are robust to dataset-specific artifacts. Thus, while BRE's relevance elicitation improves overall performance and NoTA handling, it cannot fully compensate for relations that lack clear distinguishing features or require deeper contextual reasoning beyond what current LLMs can reliably provide in zero-shot settings.

```python
def parse_binary_output(llm_output):
    llm_output = llm_output.replace('\n', ' ')
    parts = re.split(r'[^\w\s]', llm_output)
    for part in parts: # Assuming output starts response with 'true' or 'false'
        if 'true' in part.lower() or 'false' in part.lower():
            return 'true' in part.lower() and 'not true' not in part.lower()
    return False # No 'true' or 'false' in the output
```

Figure 6: Python script to parse outputs for all LLM models except Llama-4-Scout-17B-16E.

```python
def parse_binary_output_scout(llm_output):
    parts = llm_output.split('\n\n')

    # If the output starts response as requested (with 'true' or 'false')
    if 'true' in parts[0].lower() or 'false' in parts[0].lower():
        return 'true' in llm_output[0].lower()

    # Identify where the final answer is located
    phrases = ['correct answer', 'final answer']
    for part in parts:
        for phrase in phrases:
            if phrase.lower() in part.lower():
                sp = ' '.join(part.split('\n')[-2:])
                return 'true' in sp.lower()
    return False
```

Figure 7: Python script to parse outputs for Llama-4-Scout-17B-16E.

```python
def parse_relevance_elicitation_score(llm_output):
    parts = re.findall(r'[0-9]+', llm_output)
    if parts:
        result = parts[0]
    else:
        result = 10 # No confidence score found or we miss it; keep the result
    return result
```

Figure 8: Python script to parse outputs relevance elicitation scores for all LLM models.

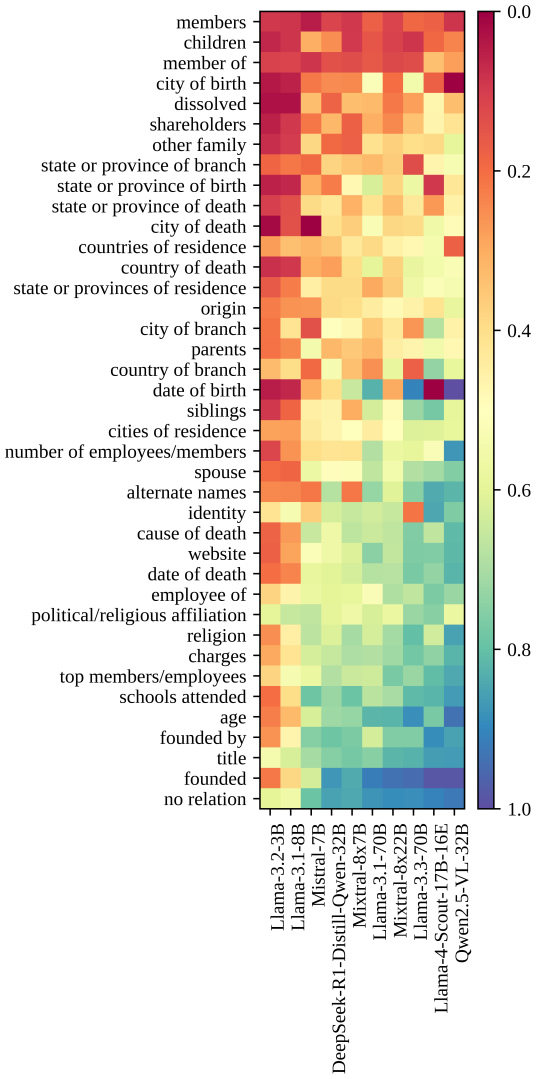

Figure 9: F_1 scores for the varying relation types in the Re-TACRED dataset. The relations are sorted based on their F_1 and the models are sorted based on their macro-F_1 scores.

Figure 10: F_1 scores for the varying relation types in the TACREV dataset. The relations are sorted based on their F_1 and the models are sorted based on their macro-F_1 scores.