# Zero-Shot Relation Classification with LLMs

## Binary Prompting and Relevance Elicitation

Sheila Alemany, Danielle Sullivan-Pao, Olga Simek · MIT Lincoln Laboratory, Lexington, MA, USA

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

## Motivation

### Why zero-shot relation classification?

- Relation classification (RC) turns unstructured text into structured knowledge: triplets that build knowledge graphs for tasks including search and question answering
- Text volumes are growing too fast for manual knowledge extraction, and supervised models need large, consistently labeled data that rarely exists for new domains

### Task: Closed RC

- Given a sentence $c$, an entity pair $(e_h, e_t)$, and predefined relation types $R$, predict $(e_h, r_{pred}, e_t)$ where $r_{pred} \in R \cup \{\text{“no relation”}\}$

Steve Jobs, who was born in San Francisco, co-founded Apple in 1976.

| | | |
|---|---|---|
| Apple | founded by → | Steve Jobs |
| Steve Jobs | city of birth → | San Francisco |
| San Francisco | no relation → | 1976 |

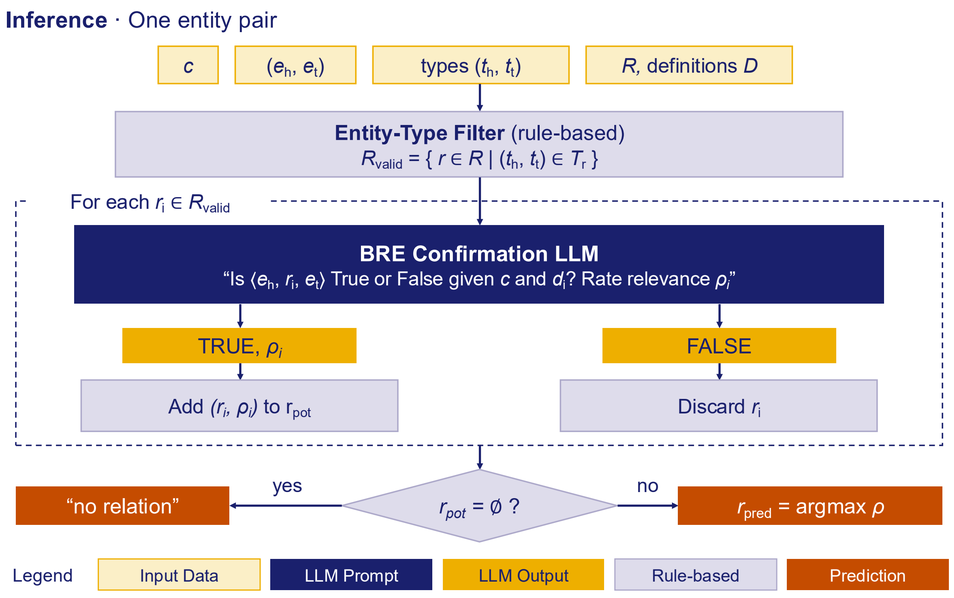
- **Zero-shot:** Users only name the relations of interest so there are no annotated data, no fine-tuning, and no examples in the prompt

### Problem Statement

- **Gap:** Large language models (LLMs) overpredict named relations instead of “no relation” (none-of-the-above, NoTA), and existing prompts rely on manual, domain-specific templates
- **Binary Relevance Elicitation (BRE):** we provide true/false questions per relation, then rank the “true” relations by an LLM-elicited relevance score

## BRE Methodology

- For each type-compatible relation $r_i$, the LLM answers true/false and rates relevance $\rho_i \in [0, 10]$ such that the highest-$\rho$ “true” relation is the prediction




- **Entity-type aware querying:** Dataset-defined type constraints prune implausible candidates (e.g., “founded by” between two locations)
- **Relation definitions:** Guideline definitions $d_i$ disambiguate overlapping relations (e.g., “employed by” vs. “affiliated with”)

## Relation Classification Results

- **Data:** Full test sets of TACRED (42 relation types) and the relabeled versions TACREV and Re-TACRED
- **LLMs:** 10 open-source instruction-tuned models (Llama, Mistral, Qwen, DeepSeek; 3B–176B) with the same prompt
- **Runs:** vLLM on up to 4 NVIDIA H100s; mean of 5 inference runs

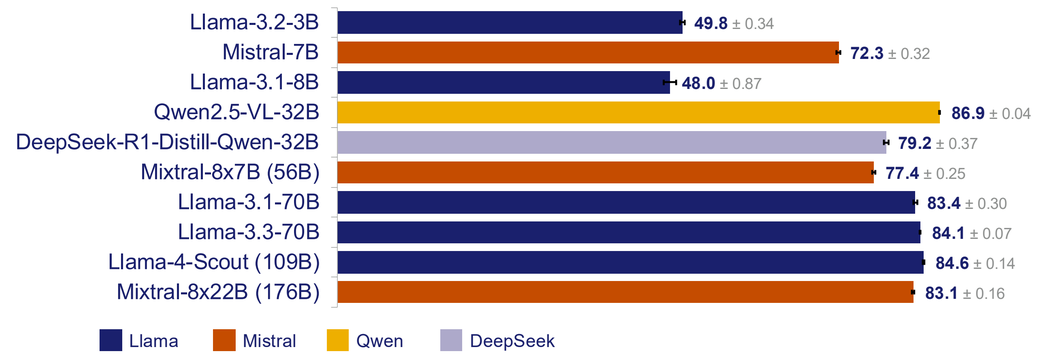
- BRE beats every zero-shot baseline on all three datasets
- Same Mixtral-8x7B backbone: +17% relative accuracy over REA (TACRED with NoTA-excluded)

| Micro-F1 (%) | NoTA-included | | | NoTA-excluded | | |
|---|---|---|---|---|---|---|
| | Re-TAC | TACREV | TACRED | Re-TAC | TACREV | TACRED |
| **BRE (ours) · Zero-shot** | | | | | | |
| Qwen2.5-VL-32B | **85.6** | **90.0** | **86.8** | **74.7** | **82.4** | **73.9** |
| Llama-4-Scout-17B-16E | 83.2 | **86.7** | **84.4** | **78.8** | **86.3** | **78.8** |
| Llama-3.3-70B | 79.5 | 86.4 | 83.9 | 65.6 | 78.9 | **73.9** |
| Mixtral-8x22B | 79.4 | 85.9 | 83.0 | 69.0 | 79.4 | 70.7 |
| Llama-3.1-70B | 79.1 | 85.8 | 83.4 | 62.1 | 68.3 | 62.7 |
| DeepSeek-R1-Distill-Qwen-32B | 74.5 | 80.8 | 79.3 | 66.7 | 75.7 | 68.5 |
| Mixtral-8x7B | 72.4 | 78.7 | 77.7 | 65.9 | 71.3 | 68.0 |
| **Zero-shot Baselines** | | | | | | |
| SUMASK · GPT-3.5-Turbo | 73.8 | 75.1 | 79.6 | – | – | 57.7 |
| 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 |
| **Few-shot Baselines (8-shot)** | | | | | | |
| ODDA · TYP Marker | 76.8 | – | 75.8 | – | – | – |
| ODDA · KnowPrompt | **84.1** | – | 81.0 | – | – | – |
| ConsistRE · TYP Marker | 71.8 | – | 72.6 | – | – | – |
| ConsistRE · KnowPrompt | 75.8 | – | 80.1 | – | – | – |
| **Δ Best BRE vs. Best zero-shot** | **+11.8** | **+14.9** | **+7.2** | **+11.5** | **+26.9** | **+15.3** |

**Bold orange** = best, **bold** = second best per column. BRE scores are means over 5 runs; baseline scores as reported in their papers (– = not reported).

## What Drives Performance?

TACRED micro-F1 (NoTA-included, mean ± std), sorted by parameter count



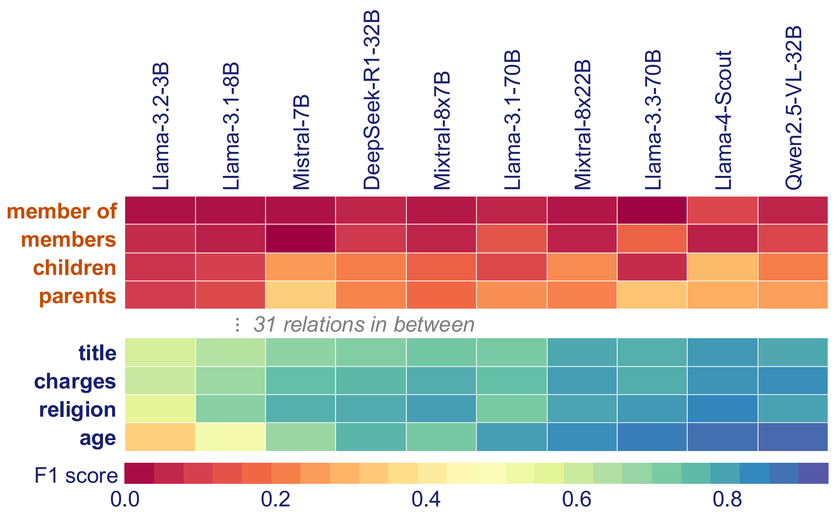

**p = 0.006**
Parameter Size: Significant

**p = 0.667**
Release Date: Not Significant

- Regression controlling for both factors (also holds with NoTA excluded): gains over time come from scale, not newer releases
- Qwen2.5-VL-32B’s lead is driven by strong NoTA classification; Llama-4-Scout is more balanced and best NoTA-excluded (shown in the table results)
- 5 runs per model to account for stochastic variation: standard deviations stay ≤ 0.9 pts on TACRED, so model rankings are stable

## Relation-Specific Results

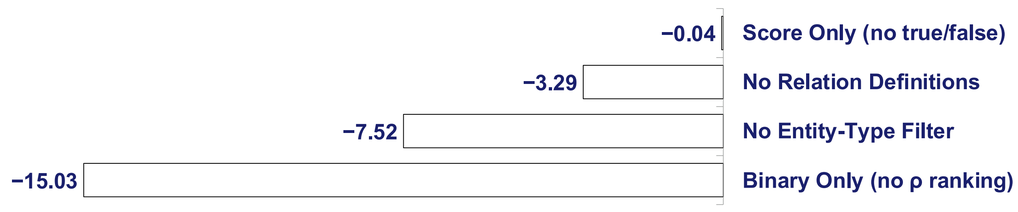
- All 10 LLMs struggle on the same relations (TACRED per-relation F1)




- **Hard relations are look-alike competitors:** Inverse pairs (“member of” ↔ “members”) or directionality describe the same link and earn similar relevance scores
- **Easy relations are distinctive targets:** An age, title, charge, or religion is its own entity type, so few candidates survive type filtering

## Ablation Study

Average Δ micro-F1 (pts) for all BRE runs, across datasets and models




- **Relevance elicitation drives the gains:** Binary-only raises macro recall (+15.3), but precision falls as the LLM affirms too many relations
- **Type filtering and definitions matter:** Without filtering, macro-F1 drops 47.5 pts; removing definitions costs 4% relative accuracy ($p = 0.01$)

## Limitations and Future Work

- **Schema dependence:** High performance needs well-documented relation definitions and, more importantly, entity types despite the zero-shot assumption
- **Compute:** One query per valid relation and entity pair, $O(|C| \times |E|^2 \times |R|)$ if no entity-type filtering is provided
- **Future work:** Resolve relevance ties (8 to 54% of instances), extend effort to document-level relation classifications, and evaluate on newer LLM models

## Key Takeaways

- **Rank, don’t just verify:** Relevance scores curb LLM overprediction by eliminating irrelevant or overly ambiguous relations
- **No labels needed:** State-of-the-art zero-shot RC that rivals state-of-the-art 8-shot methods
- **Language interpretation is hard:** Ambiguity in language remains an open problem
- **Scale over recency:** Parameter count, not release date, predicts accuracy

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Combatant Commands under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Combatant Commands.
© 2026 Massachusetts Institute of Technology. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY