Distill-RE: Few-Shot Relation Extraction via Teacher LLM Reasoning Distillation

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

1 Problem Statement

- **Sentence-level closed relation extraction (RE):** Given a sentence s with entity set E and predefined relation types R, predict triplets $\langle e_h, r, e_t \rangle$, where $e_h \neq e_t \in E$ and $r \in R$, including "no relation"
- **N-way k-shot:** Training and validation sets each contain k labeled examples per relation type ($k \in \{8, 16, 32\}$, $N = |R|$)

Marie Curie, born Maria Skłodowska in Warsaw, received the Nobel Prize in 1903.

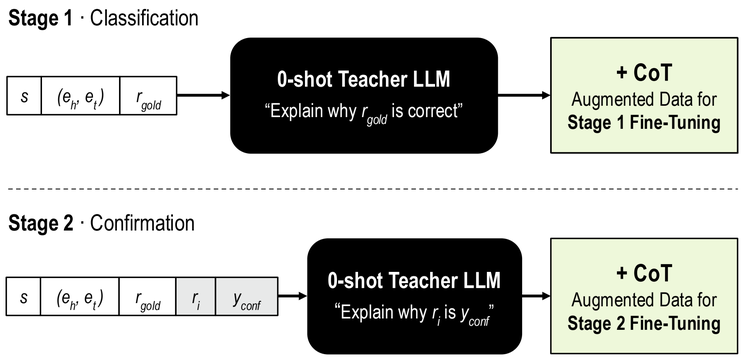
Marie Curie	alternate names →	Maria Skłodowska
Marie Curie	city of birth →	Warsaw
Warsaw	no relation →	1903

- Despite their scale, large language model (LLM) performance trails much smaller pre-trained language models (PLMs) on few-shot RE
- **Why aren't LLMs high performers in RE out of the box?** LLMs resist answering "no relation" and assign a relation to pairs that have none (i.e., the overprediction problem)
- This overprediction bias persists under in-context learning with frozen LLMs, even with careful prompting

~80% of instances in standard RE benchmarks are "no relation," so overprediction has an outsized cost

2 Training Methodology

- **Chain-of-Thought (CoT) supervision:** Each student model outputs its label followed by an explanation, learned from a frozen teacher (Mixtral-8x22B), that is given the answer
- **Sample efficiency:** Explicit reasoning lets each of the k examples carry more reasoning signal
- **Hard negatives:** Stage 2 also trains on two of the three most similar type-compatible relations per prediction, labeled FALSE
- **Teacher choice:** The teacher LLM model comes from a different model family than the students to limit shared inductive bias



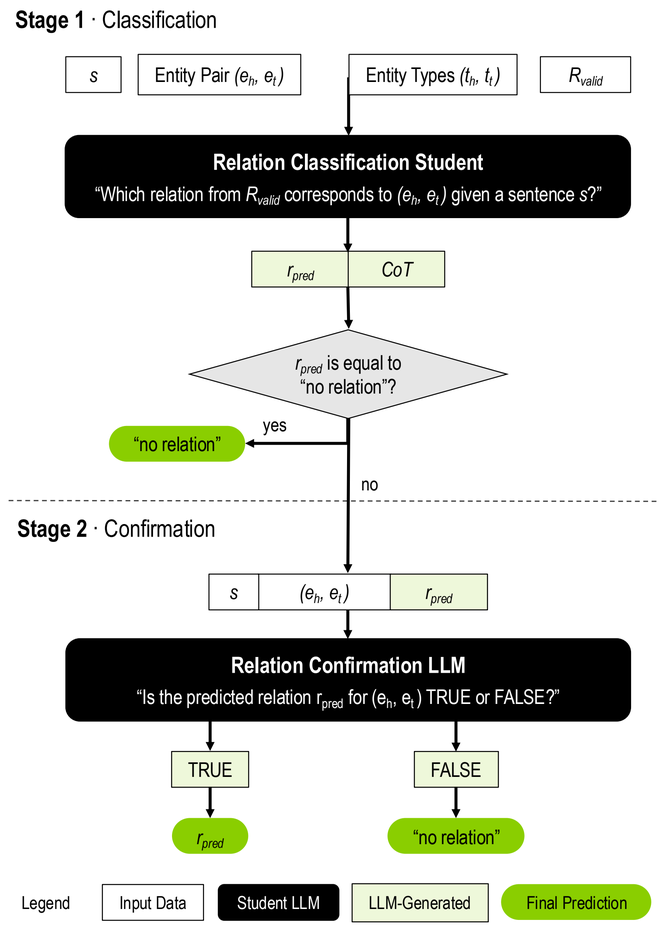
r_i = student 1's prediction or a sampled hard negative
y_{conf} = TRUE or FALSE if r_i is equal to y_{conf}

- **Objective (both students with LoRA):** causal language modeling loss over z = *label* + *CoT*, given prompt x

$$\mathcal{L} = -\sum_{t=1}^{|z|} \log P_\theta(z_t \mid z_{<t}, x)$$

3 Inference Methodology

- **Stage 1:** Classification student maximizes named-relation recall
- **Stage 2:** Confirmation student removes the false positives corresponding to "no relation"



Why two stages?

- Stage 1 targets recall by casting a wide net, so no true relation is missed
- Stage 2 targets precision by re-labeling false positives as "no relation"
- Decoupling these competing objectives sidesteps the overprediction bias of single-stage LLMs

Implementation Details

- **Type-aware filtering:** Prompts list only R_{valid}, the relations compatible with the pair's entity types, plus "no relation"
- **Answer parsing:** Regex on the leading label; Stage 1 outputs map to the nearest relation by BERT embedding similarity
- **Early exit:** Stage 2 runs only when $r_{pred} \neq$ "no relation"

Worked Example · Illustrative, sentence from Problem Statement section

Entity Pair	Stage 1 → r_{pred}	Stage 2	Final
(Marie Curie, Warsaw)	city of birth	TRUE	city of birth ✓
(Marie Curie, 1903)	date of birth ✗	FALSE	"no relation" ✓

4 State-of-the-Art Few-Shot RE

- **Data:** TACRED, TACREV, Re-TACRED with k = 8, 16, 32 with 5 different seeds
- **Teacher Model:** Mixtral-8x22B (frozen); writes CoT for training data only
- **Student Models:** Llama-3.2-3B, Mistral-7B, Llama-3-70B with LoRA (Low-Rank Adaptation) fine-tuning

- We achieved +31.5 micro-F1 over the strongest baseline on average (Llama-70B student)
- Comparing with the highest performing baselines in literature, we include PLM classifiers or hybrids where an LLM only synthesizes data or breaks ties

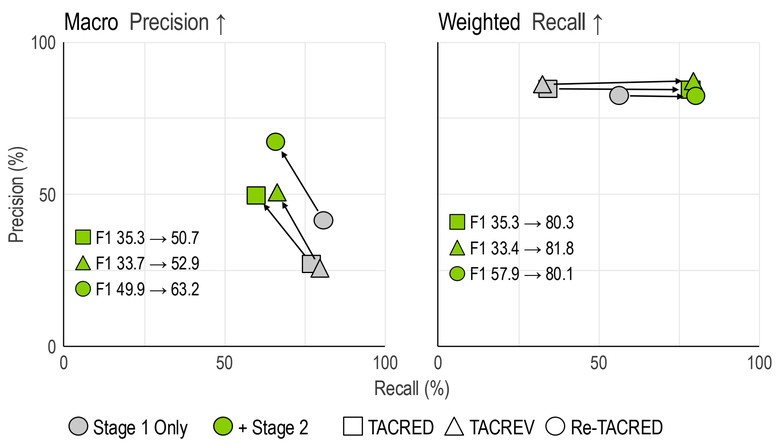
	TACRED			TACREV			Re-TACRED		
	8	16	32	8	16	32	8	16	32
LLM-based RE (ours)									
Distill-RE · Llama-70B	**81.0**	**82.8**	**83.1**	**82.7**	**87.0**	**87.1**	**83.3**	**83.5**	**84.4**
Distill-RE · Mistral-7B	**78.8**	**74.0**	**76.6**	**77.3**	**79.6**	**76.2**	**79.2**	**79.3**	**79.6**
Distill-RE · Llama-3B	**75.6**	**79.8**	**77.2**	**78.6**	**79.7**	**79.7**	**77.9**	**79.3**	**80.6**
Δ 70B vs. best baseline	***+37.2***	***+37.4***	***+37.2***	***+38.0***	***+40.4***	***+40.2***	***+22.0***	***+15.9***	***+15.5***
PLM-based RE									
TYP Marker	28.9	32.0	32.4	27.6	31.2	32.0	44.8	54.1	60.0
PTR	28.1	30.7	32.1	28.7	31.4	32.4	51.5	56.3	62.1
KnowPrompt	32.0	35.4	36.5	32.1	33.1	43.7	55.3	63.3	65.0
GenPT (BART)	32.6	35.6	37.1	31.1	34.3	36.6	52.9	58.1	63.9
GenPT (RoBERTa)	35.7	36.6	37.4	34.4	34.6	36.2	57.1	60.4	65.2
GenPT (T5)	31.8	33.3	36.0	31.3	32.6	34.6	54.7	58.7	62.5
RetrievalRE	32.0	35.2	37.1	28.9	32.3	37.3	33.7	55.8	61.8
PTRE	32.7	36.5	39.2	33.9	36.3	38.1	58.8	64.4	66.3
Hybrid LLM–PLM									
S²ynRE	36.3	37.2	38.7	35.2	36.1	36.9	59.4	63.9	67.5
DSARE	43.8	45.4	45.9	44.7	46.6	46.9	60.0	66.8	67.1
TKRE (TYP Marker)	36.7	38.3	39.1	37.2	38.2	39.0	58.7	63.8	65.5
TKRE (GenPT)	39.6	42.1	43.2	40.2	41.3	42.9	61.3	67.6	68.9

The reported scores are micro-F1 and the mean over five random k-shot splits on full test sets; baselines as reported in their papers.

- ODDA, the strongest baseline method, beats Distill-RE by only 0.4–3.2 micro-F1 (k = 8, 16) but adds *8 x k* synthetic samples per run, whereas Distill-RE uses only the k real ones
- Our few-shot RE approach reaches state-of-the-art RE performance in a computationally and data-efficient manner

5 Effect of Stage 2 Confirmation

- Below shows Stage 1 → Stage 2 at k = *8* with the same trends present at k = 16, 32



- **Stage 1 retrieves named relations effectively:** strong macro recall but low weighted highlights LLM overprediction
- **Stage 2 corrects overprediction:** +16.3 macro-F1 and +36.9 weighted-F1 from Stage 2, averaged over all 135 runs
- Without the Stage 2 confirmation, the fine-tuned LLM trails the best hybrid baseline on all three datasets

6 Effect of CoT Supervision

+9.1 Macro-F1 | **+7.6** Micro-F1 | **11.1%** of runs degrade

- 135 matched runs with and without teacher CoT shows the drops are rare and small (mean −2.7)
- Largest gains are seen on the cleanest labels (Re-TACRED) and with the Llama-70B student

7 Limitations and Future Work

- **Compute:** LLMs cost more than PLMs, but we observed that accuracy scales with size
- **Supervision:** Need a capable teacher and a clear relation schema
- **Limited scope:** We evaluated only on TACRED-family datasets
- **Future work:** LLM data augmentation, stronger teachers, and extending this approach for document-level RE

8 Conclusion

- **Overview:** Distill-RE is a two-stage distillation framework for few-shot RE that transfers teacher LLM chain-of-thought reasoning to student models

Key Takeaways

- **Decouple recall and precision:** Two specialized students tame overprediction
- **Distill reasoning, not just labels:** Teacher CoT makes few-shot learning more sample-efficient
- **Strong at every scale:** Even the 3B student beats every state-of-the-art baseline in all nine evaluation settings
- **Confirmation is the key step:** Stage 2 adds +36.9 micro-F1 on average across all 135 runs
- **Data-efficient:** Uses only the k real examples per relation, while the strongest baseline (ODDA) adds *8* × k synthetic samples per run

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