DistilledGemma: Balanced Efficiency-Accuracy for Person-Place Relation Extraction from Multilingual Historical Articles

Youssef Aboelwafa[1,†], Ahmed Samir[1,†], Nagwa Elmakky[1] and Marwan Torki[1]

[1]*Alexandria University, Egypt*

Abstract

We present **DistilledGemma**, an efficient and accurate system for the **HIPE-2026** shared task on person-place relation extraction from multilingual historical newspaper articles in English, German, and French. Our approach adopts a three-stage knowledge distillation pipeline designed to balance classification accuracy with computational efficiency. In the first stage, we systematically explored prompt engineering strategies across eight large language models to identify the most effective reasoning architecture for this challenging task. In the second stage, we applied supervised fine-tuning (SFT) via QLoRA to a **Gemma 4 26B A4B** teacher model, leveraging its strong multilingual capabilities to generate silver-standard chain-of-thought traces across the training corpus. In the final stage, we performed response-level distillation to transfer these learned reasoning patterns into a compact **Gemma 4 E2B** student model. To address the class imbalance between positive and negative relation pairs and to enforce semantic consistency, we incorporated rule-based post-processing on all generated outputs. Specifically, this step enforced entity-type compatibility, relation-direction constraints, contextual trigger patterns, and the logical constraint. In the official evaluation, our team **WHEREAMI** ranked **3rd** on the standard test set with an accuracy profile mean score of **0.688**, and **2nd** on the binary test set with a mean score of **0.8156**. Notably, by distilling knowledge from the 26B teacher to the 2.3B student, we preserved strong reasoning capabilities while reducing the deployed model size to approximately **2.3B effective parameters**; the LoRA adapters used during training were merged into the student for inference. This configuration ranked **2nd** in the balanced efficiency-accuracy profile across both the standard and binary test sets. These results demonstrate that knowledge distillation provides a practical and scalable solution for historical document processing, achieving competitive performance without excessive computational cost.

1. Introduction

Historical newspapers constitute invaluable resources for studying past societies, migration patterns, and geopolitical events. A critical step in unlocking this information is **person-place relation extraction**: determining whether a person mentioned in an article bears a geographic connection to a place also mentioned therein.

The HIPE-2026 shared task [1, 2, 3] formalizes this problem as a classification task over (person, place) pairs extracted from multilingual newspaper articles in English, German, and French. Each pair must be labelled for two relations: **at** (whether the person has *ever* been associated with the place; TRUE/PROBABLE/FALSE) and **isAt** (whether the person is at the place *within the immediate temporal horizon of the article*; TRUE/FALSE) as shown in Figure 1 and 2.

This task presents several challenges:

- **Noisy OCR text**: Historical documents contain frequent optical character recognition errors that degrade entity mention quality.
- **Class imbalance**: The FALSE label dominates both relations (∼55% for *at*, ∼78% for *isAt*), biasing models toward negative predictions.
- **Multilingual reasoning**: Models must perform consistently across three typologically diverse languages with varying training data availability.

The Conference and Labs of the Evaluation Forum (CLEF) 2026

[†] Equal contribution.

es-YoussefHossamElDin2025@alexu.edu.eg (Y. Aboelwafa); es-AhmedAbdelMaksoud2025@alexu.edu.eg (A. Samir); nagwamakky@alexu.edu.eg (N. Elmakky); mtorki@alexu.edu.eg (M. Torki)

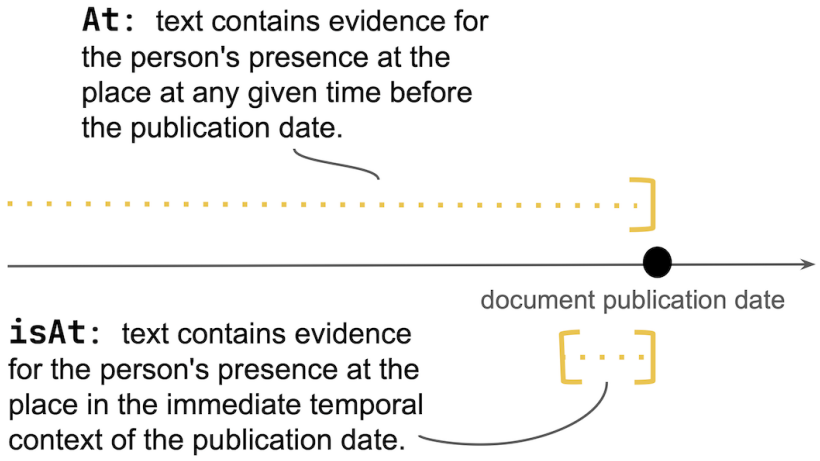

Figure 1: Illustration of the distinction between *at* and *isAt* relations.

Figure 2: Illustration of the distinction between *at* and *isAt* relations in HIPE-2026. The *at* relation captures historical or general associations, while *isAt* denotes the person's location at the time described in the article.

- **Implicit evidence**: Many person-place connections are implied rather than explicitly stated, requiring world knowledge and contextual inference.

Large language models (LLMs) have demonstrated impressive zero-shot and few-shot capabilities for relation extraction [4, 5], but models with sufficient capacity (e.g., 26B+ parameters) are impractical for deployment in digital humanities pipelines that must process millions of documents. This motivates our central research question: *Can we distill the relation extraction capabilities of a large, fine-tuned LLM into a model small enough for practical deployment without catastrophic performance loss?*

Our contributions are as follows:

1. A **comprehensive evaluation** of eight LLM configurations spanning prompt engineering, zero-shot inference, and supervised fine-tuning for person-place relation extraction across three languages.
2. A **three-stage distillation pipeline** that fine-tunes a Gemma 4 26B A4B teacher via QLoRA, generates silver-standard training data with chain-of-thought (CoT) reasoning, and trains a Gemma 4 E2B student on the teacher's outputs.
3. **Empirical evidence** that knowledge distillation recovers ∼88% of teacher performance, establishing a practical efficiency-accuracy trade-off for historical NLP applications.

2. Related Work

Encoder-based relation extraction. Before the recent shift to generative LLMs, relation extraction was commonly framed as discriminative classification over marked entity pairs, using either hand-engineered features [6] or contextual encoders [7]. These lightweight baselines remain important for low-resource and imbalanced settings because they condition directly on entity spans and require only a small task head. Entity-aware representation learning, including task-agnostic relation representations from entity-linked text [8], typed entity-marker models [9], and PL-Marker span-pair packing [10], provides strong baselines for sentence-level RE. Cross-lingual encoders such as XLM-R [11] and entity-aware multilingual encoders such as mLUKE [12] are especially relevant for multilingual HIPE-style data. For longer article-level evidence, document-level RE benchmarks such as DocRED highlight the need to synthesize evidence across sentences and entity mentions [13]. Recent multilingual and digital-humanities RE work has also used guided distant supervision to build German biographical relation data and evaluate multilingual transfer [14]. Low-resource RE benchmarks show that class balancing, data augmentation, and self-training are useful alternatives, but their gains can be inconsistent under long-tailed label distributions [15]. These approaches are therefore strong lightweight baselines and complementary alternatives to our LLM distillation pipeline.

Relation extraction with LLMs. Recent work has demonstrated that LLMs can perform competitive relation extraction through in-context learning [5] and chain-of-thought prompting [16], particularly when explicit reasoning steps are elicited before label prediction. Wadhwa et al. [4] showed that GPT-class models match or exceed supervised baselines on standard benchmarks, albeit at significantly higher computational cost.

Historical NLP. The HIPE shared task series [17, 18] has driven substantial progress in named entity recognition and linking for historical texts, establishing standardized benchmarks across multiple languages and time periods. These tasks highlight unique challenges including OCR noise, archaic language usage, and domain-specific entity types that are absent from modern NLP benchmarks. The HIPE-2026 edition [2] extends the series to person-place relation extraction, introducing new evaluation profiles that jointly consider accuracy and computational efficiency.

Knowledge distillation and teacher alignment. Hinton et al. [19] introduced knowledge distillation (KD) as a model compression technique in which a compact student network learns to mimic a larger teacher's output distribution. For generative models, sequence-level KD trains students on complete teacher outputs rather than only token-level distributions [20]. This paradigm has been widely adopted for language models: DistilBERT [21] compresses BERT while retaining 97% of its performance, and MiniLLM [22] extends KD to autoregressive LLMs with a policy-level reverse KL objective. On-policy approaches such as GKD [23] further improve upon standard KD by training on the student's own generations and allowing alternative KL divergences. Closer to our setting, rationale- and explanation-trace distillation methods show that small models can benefit from LLM-generated reasoning supervision [24, 25]. Preference-based alignment methods such as DPO [26] provide another route by optimizing pairwise teacher or human preferences rather than directly imitating a single teacher

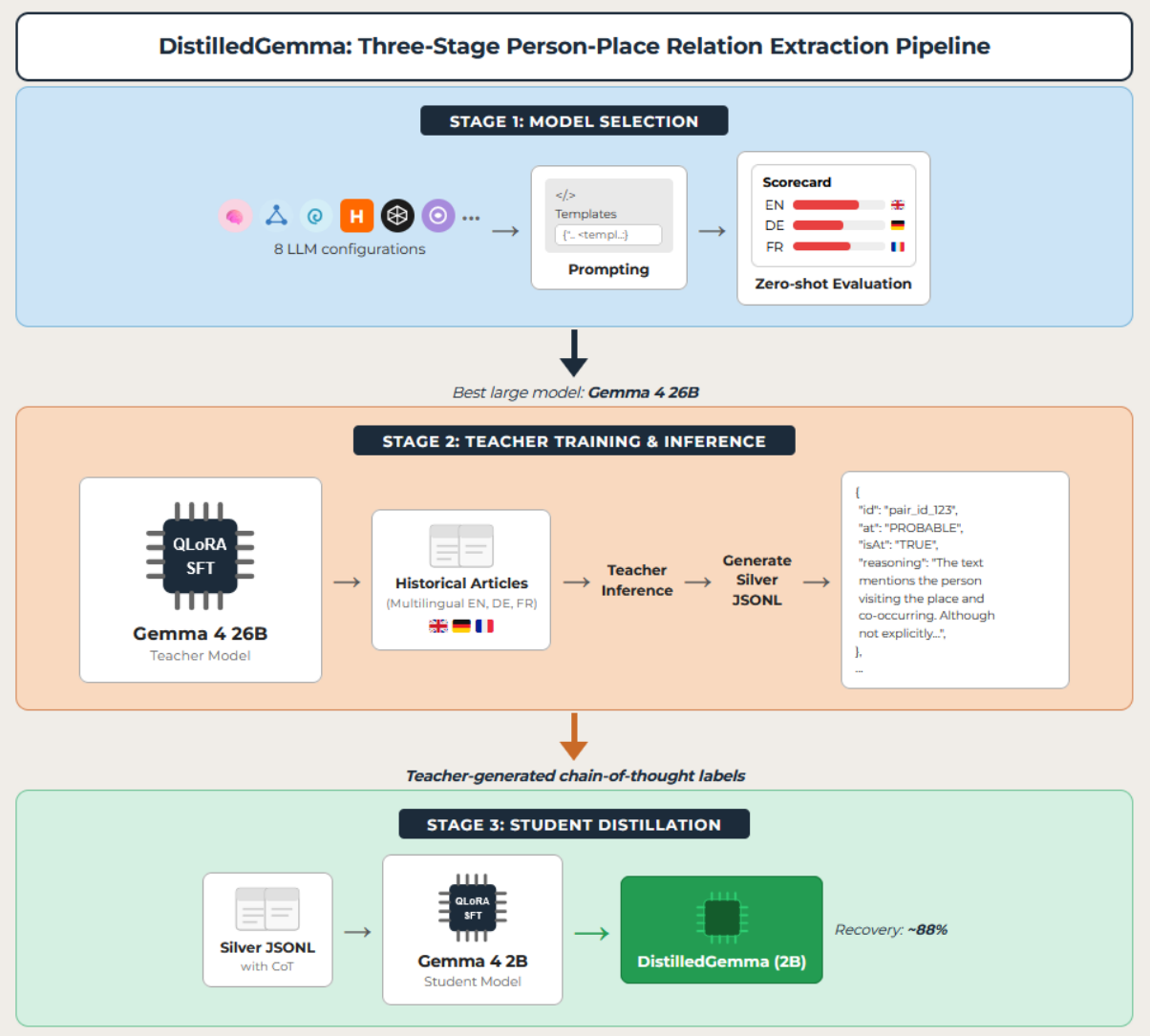

Figure 3: Overview of the DistilledGemma three-stage pipeline. First, multiple LLM configurations and prompting strategies are evaluated to select the strongest teacher model. The teacher is then fine-tuned and used to generate chain-of-thought silver annotations, which are subsequently used to distill knowledge into a compact student model.

response. In low-resource RE, self-training and confidence calibration are also plausible alternatives; for example, PRiSM calibrates document-level RE logits with relation-aware scores when only a small amount of labeled data is available [27]. In our work, we adopt a *response-level* distillation strategy in which the teacher's evidence-grounded explanations and label outputs serve as silver training data for the student, combining structured reasoning transfer with the simplicity of standard SFT and avoiding the extra preference construction or confidence-threshold tuning required by these alternatives.

Parameter-efficient fine-tuning. LoRA [28] and its quantized variant QLoRA [29] enable fine-tuning of billion-parameter models on consumer hardware by training only low-rank adapter matrices while keeping the base model weights frozen. We leverage QLoRA throughout our pipeline for both teacher and student training, making the entire workflow feasible on a single GPU.

3. Approach

Our approach follows a three-stage pipeline, illustrated in Figure 3: (1) model selection through systematic experimentation, (2) teacher fine-tuning and silver data generation, and (3) student distillation.

3.1. Stage 1: Model Selection and Prompt Engineering

We evaluated a diverse set of LLM families in multiple configurations:

Prompt engineering. We designed five prompt variants ranging from a minimal zero-shot baseline to structured chain-of-thought templates that guide the model through explicit evidence assessment before emitting labels. The best-performing prompt (chain-of-thought), shown in Listing 1, was used as the inference prompt for all LLM candidates to predict the *at* and *isAt* relations during base-model selection. It frames the model as a historian, defines the two relation labels, emphasizes OCR robustness and conservative inference, enforces the logical dependency between labels, and requires concise evidence-grounded JSON outputs.

```
You are a historian working on person-location relations in multilingual historical
European newspapers.

Read carefully, expect OCR noise, and base every judgment only on the document text and
metadata given here.

Task:
- Classify relation "at" for the person and place. Allowed labels: TRUE, PROBABLE, FALSE
- Classify relation "isAt" for the person and place. Allowed labels: TRUE, FALSE

Relation meanings:
- "at" = the article gives textual evidence that the person was at the location at some
  point before publication.
- "isAt" = the article supports that the person was at the location in the article's
  immediate temporal horizon, meaning current, ongoing, or very recent relative to the
  publication date.

Decision guidance:
- Use TRUE for "at" only when the text gives clear evidence of presence.
- Use PROBABLE for "at" only when the text gives indirect, partial, or weak evidence that
  still points toward presence.
- Use FALSE for "at" when the article does not support presence.
- Use TRUE for "isAt" only when the article clearly places the person there in a current,
  ongoing, or very recent frame.
- Never use PROBABLE for "isAt". It must be TRUE or FALSE.
- If "at" is FALSE, then "isAt" must also be FALSE.
- If "isAt" is TRUE, then "at" must also be TRUE.

Rules:
- Do not use external knowledge.
- Do not infer more than the text warrants.
- Prefer FALSE when evidence is missing or too uncertain.
- Return JSON only.
- Use the exact person and place strings given for the current pair.
- Be robust to OCR noise and line-break artifacts in the text and in the entity mentions.
- Treat clearly equivalent mention surfaces as the same entity even if they differ by escaped line
    breaks, hyphenation, or minor OCR spelling noise.
- First write `at_explanation`, then decide `at`.
- Then write `isAt_explanation`, then decide `isAt`.
- Keep each explanation concise, at most 100 words.
- Each explanation must mention only evidence from the article, not your reasoning process.

Document language: {language}
Publication date: {publication_date}

Article text:
{article_text}

{pair_context}
```

Listing 1: Prompt template used for LLM inference, teacher generation, and student fine-tuning.

Model families. We evaluated models from four families: **Gemma** [30] (E2B, E4B, and 26B A4B), **Qwen** [31] (2B to 9B), **Mistral** [32] (3B), and **HY-MT** [33] (1.8B). Models were served via vLLM [34] for efficient batch inference.

Training strategies. Beyond zero-shot prompting, we explored supervised fine-tuning (SFT) using LoRA [28] on the HIPE training data. The SFT data was prepared by converting each labeled (person, place) pair into a chat-format record with the prompt as the user message and the gold JSON as the assistant response.

3.2. Stage 2: Teacher Fine-Tuning

We evaluated **Gemma 4 26B A4B** [35] using the same prompt as in the selection stage. We selected it as the teacher model because it achieved the strongest zero-shot macro-averaged recall (0.7255 across languages) and belongs to the same model family as the target student.

The teacher was fine-tuned using QLoRA [29] with the following configuration:

- **LoRA**: rank r=32, α=64, dropout 0.1, applied to all attention and MLP projections (q/k/v/o/-gate/up/down).
- **Training**: 3 epochs, batch size 1 with gradient accumulation 32 (effective batch size 32), cosine learning rate schedule with η=2×10^{-4} and 5% warmup.
- **Data**: All labeled pairs from the newspaper and sandbox train splits across EN/DE/FR, with a 90/10 stratified train/dev split.

After training, the LoRA adapter was merged into the base model for efficient inference. The merged teacher was then used to generate silver-standard labels over the entire training corpus, producing chain-of-thought explanations alongside *at* and *isAt* predictions for each pair.

3.3. Stage 3: Student Distillation

The student model (**Gemma 4 E2B**) was trained on the teacher-generated silver data using QLoRA with a lighter configuration:

- **LoRA**: rank r=16, α=32, dropout 0.05.
- **Training**: 6 epochs (longer training compensates for the smaller model capacity), effective batch size 16.
- **Data**: Teacher-generated distilled JSONL, where the assistant messages contain the teacher's chain-of-thought with its predicted outputs rather than gold labels.

This *response-level distillation* approach transfers not only the teacher's label decisions but also its reasoning patterns, enabling the student to learn structured inference over noisy historical text. The student learns to emulate the teacher's chain-of-thought process, which we hypothesize improves generalization compared to training on gold labels alone, as the latter lack detailed explanatory reasoning for many relation pairs.

For parameter accounting, we count only the deployable student model: the 26B teacher is used offline to create silver data, and the LoRA matrices are training-time adapters that are merged into the 2.3B base model for inference. Consequently, the efficiency profiles and compression ratios report the student as a 2.3B-parameter deployed model rather than summing teacher and student parameters or treating the adapters as a separate multi-billion-parameter component.

3.4. Post-Processing

To address the imbalance between positive and negative relation pairs and ensure semantic consistency in model outputs, we applied rule-based post-processing to all generated predictions. Specifically, this step enforced: (1) entity-type compatibility between the predicted relation and the entity types involved,

(2) relation-direction constraints ensuring the correct mapping of person-to-place, (3) contextual trigger patterns that leverage surface-level cues in the article, and (4) the logical constraint that isAt=TRUE ⇒ at=TRUE.

4. Dataset

The HIPE-2026 training corpus contains multilingual historical newspaper articles in German, English, and French with varying document lengths and entity distributions. Table 1 summarizes corpus-level statistics, while Tables 2 and 3 report relation and entity characteristics.

Table 1
Dataset overview by language.

Language	Documents	Avg. Words	Avg. Characters
German	34	744.1	5000.9
English	35	325.8	1802.7
French	35	585.9	3635.0
Total	104	–	–

Table 2
Relation label distribution across languages with overall statistics and class imbalance percentages. The label distribution exhibits significant class imbalance: for *at*, FALSE accounts for ∼55%, PROBABLE for ∼10%, and TRUE for ∼35%; for *isAt*, FALSE dominates at ∼78% versus TRUE at ∼22%.

Label	German	English	French	Overall (percentage)
at relation				
TRUE	135	125	181	441 (35.25%)
FALSE	269	152	269	690 (55.16%)
PROBABLE	62	30	28	120 (9.59%)
isAt relation				
TRUE	89	83	107	279 (22.30%)
FALSE	377	224	371	972 (77.70%)

Table 3
Unique entity statistics.

Language	Unique Persons	Unique Locations
German	174	217
English	120	127
French	209	237

5. Experiments

5.1. Experimental Setup

All models were served using vLLM [34] with temperature 0.0 and seed 42 for output determinism. We report macro- and micro-averaged recall following standard classification evaluation practice [36]. The logical constraint (isAt=TRUE ⇒ at=TRUE) was enforced as a post-processing step for all configurations.

5.2. Results

Table 4 presents the results across all eight model configurations.

Model	EN	DE	FR	Macro Avg	Micro Avg
Gemma 4 26B A4B	**0.7516**	**0.7239**	**0.7009**	**0.7255**	**0.8417**
Qwen3-4B	0.6452	0.6443	0.6493	0.6463	0.7626
Gemma 4 E2B	0.6696	0.6336	0.6154	0.6395	0.7714
Qwen3.5 9B	0.5641	0.6279	0.6268	0.6063	0.7912
Mistral3 3B	0.5846	0.5852	0.5461	0.5720	0.6986
Gemma 4 E4B	0.5458	0.5839	0.5830	0.5709	0.7804
Qwen3.5 2B	0.5042	0.4886	0.4658	0.4862	0.5855
HY-MT1.5 1.8B	0.4043	0.4219	0.3906	0.4056	0.3649

Table 4
Macro-averaged recall per language and overall metrics on the HIPE-2026 dataset. Models are sorted by macro-averaged recall. **Bold** indicates best in column.

Several key findings emerge from these results:

Large models dominate. Gemma 4 26B A4B achieves the highest zero-shot macro-averaged recall (0.7255), confirming that model scale provides substantial advantages for this task. The model's strong multilingual pretraining contributes to consistent performance across all three languages.

Model scale is not the only factor. Interestingly, smaller models with superior architecture can outperform larger but less capable ones. For instance, Qwen3-4B (macro 0.6463) and the base Gemma 4 E2B (macro 0.6395) both outperform the larger Qwen3.5 9B (macro 0.6063) and Gemma 4 E4B (macro 0.5709) in macro-averaged recall, suggesting that model architecture and pretraining data composition play a significant role alongside parameter count.

Small models are viable with distillation. As shown in Section 6.1, the distilled Gemma 4 E2B student achieves competitive performance that surpasses several larger zero-shot models, validating our distillation approach.

6. Analysis

6.1. Distillation Effectiveness

Table 5 quantifies the efficiency-accuracy trade-off achieved through distillation.

	Deployed Params	Macro Avg	Recovery Rate	Compression
Gemma 4 26B A4B (teacher)	26B	0.7015	—	—
Gemma 4 E2B (distilled)	2.3B	0.6171	∼88%	∼11×
Gemma 4 E2B (zero-shot)	2.3B	0.5329	∼76%	∼11×

Table 5
Comparison of teacher vs. distilled student performance on sandbox dev set. Recovery rate is the ratio of student to teacher macro-averaged recall. Parameter counts report deployed base-model size; LoRA adapters are merged for inference.

The distilled student achieves ∼88% of the teacher's macro-averaged recall while being 11× smaller in deployed parameter count, confirming that teacher-generated reasoning patterns transfer effectively to the smaller model. These sandbox dev results are not directly comparable to the full test metrics reported in Table 4.

6.2. Official Evaluation Results

In the official HIPE-2026 evaluation, our system (team12, "whereami") achieved competitive rankings across multiple profiles:

- **Standard accuracy profile**: Ranked **3rd** among all teams with a mean profile score of **0.688**.
- **Binary accuracy profile**: Ranked **2nd** among all teams with a mean profile score of **0.8156**.
- **Balanced efficiency-accuracy profile**: Ranked **2nd** among all teams on both the standard and binary test sets, demonstrating an effective trade-off between model size and classification performance.

These results demonstrate that DistilledGemma provides a strong balance of accuracy and efficiency in several profiles while using substantially fewer parameters.

6.3. Cross-Lingual Analysis

Performance varies across languages, with English generally achieving the highest scores and French exhibiting the greatest variance. German performance is notably consistent across model sizes, possibly because German historical newspapers in the HIPE dataset tend to feature cleaner OCR output and more explicit geographic references.

The distilled student shows the largest improvement on German, suggesting that the teacher's German reasoning patterns are particularly amenable to distillation. French performance proves more challenging to distill, likely due to the more complex syntactic structures characteristic of French historical prose.

6.4. Error Analysis

We identify three dominant error categories:

1. **PROBABLE confusion**: The three-way *at* classification proves challenging, with models frequently conflating PROBABLE and TRUE. This distinction requires nuanced assessment of evidence strength that smaller models struggle to capture.
2. **False negatives on isAt**: The extreme class imbalance (77.7% FALSE) biases all models toward negative predictions, particularly degrading recall on the minority TRUE class.
3. **OCR-induced entity confusion**: In documents with severe OCR degradation, entity mentions are corrupted, leading to misidentification of both persons and places, which cascades into downstream relation classification errors.

6.5. Efficiency Considerations

The Gemma 4 E2B student model can be served on a single consumer GPU with $<$8 GB VRAM, processing approximately 50 pairs per second via vLLM. This deployment count excludes the 26B teacher, which is used only offline for silver-data generation. In contrast, the 26B teacher requires a high-memory GPU. For large-scale historical document processing involving millions of articles, the 2.3B student provides a practical solution with acceptable performance degradation, while the 26B teacher remains preferable when accuracy is paramount and computational resources are available.

7. Conclusion

We presented DistilledGemma, a knowledge distillation pipeline for person-place relation extraction from multilingual historical newspaper texts. Through systematic evaluation of eight LLM configurations, we demonstrated that supervised fine-tuning of large models yields the strongest performance,

and that response-level distillation effectively transfers chain-of-thought reasoning patterns to compact models.

Our key findings are: (1) QLoRA fine-tuning enables the selected teacher to generate task-specific silver data, and response-level distillation transfers much of that performance to the student; (2) knowledge distillation from a 26B teacher to a 2.3B student recovers ∼88% of teacher performance at $1/11$ the parameter cost; and (3) chain-of-thought distillation, where the student learns the teacher's reasoning process rather than labels alone, combined with rule-based post-processing, provides an effective pathway to competitive performance with limited computational resources.

In the official HIPE-2026 evaluation, DistilledGemma ranked 3rd in the standard accuracy profile and 2nd in both the binary accuracy and balanced efficiency-accuracy profiles, demonstrating that strong performance on multilingual historical relation extraction can be achieved without excessive computational expenditure.

Future work includes exploring on-policy self-distillation methods [23] that could further narrow the teacher-student performance gap, investigating encoder-based discriminative models as an alternative lightweight approach, and extending the framework to additional historical document domains.

References

[1] J. Opitz, C. Raclé, E. Boros, A. Michail, M. Romanello, M. Ehrmann, S. Clematide, Clef hipe-2026: Evaluating accurate and efficient person–place relation extraction from multilingual historical texts, in: European Conference on Information Retrieval, Springer, 2026, pp. 354–363.

[2] J. Opitz, C. Raclé, A. Michail, M. Romanello, M. Ehrmann, S. Clematide, Overview of HIPE-2026: Person–Place Relation Extraction from Multilingual Historical Texts, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026.

[3] J. Opitz, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, M. Ehrmann, S. Clematide, Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS, 2026. doi:10.5281/zenodo.20344461.

[4] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, in: Annual Meeting of the Association for Computational Linguistics, 2023.

[5] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, Gpt-re: In-context learning for relation extraction using large language models, in: EMNLP, 2023.

[6] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware attention and supervised data improve slot filling, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 35–45. URL: https://aclanthology.org/D17-1004/. doi:10.18653/v1/D17-1004.

[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[8] L. Baldini Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 2895–2905. URL: https://aclanthology.org/P19-1279/. doi:10.18653/v1/P19-1279.

[9] W. Zhou, M. Chen, An improved baseline for sentence-level relation extraction, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), 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), Association for Computational Linguistics, Online only, 2022, pp. 161–168. URL: https://aclanthology.org/2022.aacl-short.21/. doi:10.18653/v1/2022.aacl-short.21.

[10] D. Ye, Y. Lin, P. Li, M. Sun, Packed levitated marker for entity and relation extraction, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4904–4917. URL: https://aclanthology.org/2022.acl-long.337/. doi:10.18653/v1/2022.acl-long.337.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[12] R. Ri, I. Yamada, Y. Tsuruoka, mLUKE: The power of entity representations in multilingual pretrained language models, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7316–7330. URL: https://aclanthology.org/2022.acl-long.505/. doi:10.18653/v1/2022.acl-long.505.

[13] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A large-scale document-level relation extraction dataset, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074/. doi:10.18653/v1/P19-1074.

[14] A. Plum, T. Ranasinghe, C. Purschke, Guided distant supervision for multilingual relation extraction data: Adapting to a new language, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 7982–7992. URL: https://aclanthology.org/2024.lrec-main.703/.

[15] X. Xu, X. Chen, N. Zhang, X. Xie, X. Chen, H. Chen, Towards realistic low-resource relation extraction: A benchmark with empirical baseline study, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 413–427. URL: https://aclanthology.org/2022.findings-emnlp.29/. doi:10.18653/v1/2022.findings-emnlp.29.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[17] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, HIPE-2022: Evaluation of named entity processing and entity linking in historical newspapers and classical commentaries, in: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), European Language Resources Association, 2022.

[18] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, 2023.

[19] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015. URL: https://arxiv.org/abs/1503.02531. arXiv:1503.02531.

[20] Y. Kim, A. M. Rush, Sequence-level knowledge distillation, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1317–1327. URL: https://aclanthology.org/D16-1139/. doi:10.18653/v1/D16-1139.

[21] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL: https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[22] Y. Gu, L. Dong, F. Wei, M. Huang, Minillm: Knowledge distillation of large language models, in: International Conference on Learning Representations, volume 2024, 2024, pp. 32694–32717.

[23] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. Ramos Garea, M. Geist, O. Bachem, On-policy distillation of language models: Learning from self-generated mistakes, in: International Conference on Learning Representations, volume 2024, 2024, pp. 21246–21263.

[24] C.-Y. Hsieh, C.-L. Li, C.-k. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, T. Pfister, Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 8003–8017. URL: https://aclanthology.org/2023.findings-acl.507/. doi:10.18653/v1/2023.findings-acl.507.

[25] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of GPT-4, arXiv preprint arXiv:2306.02707 (2023). URL: https://arxiv.org/abs/2306.02707. doi:10.48550/arXiv.2306.02707.

[26] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference op-

timization: Your language model is secretly a reward model, in: Advances in Neural Information Processing Systems, volume 36, 2023. URL: https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

[27] M. Choi, H. Lim, J. Choo, PRiSM: Enhancing low-resource document-level relation extraction with relation-aware score calibration, in: Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings), Association for Computational Linguistics, 2023, pp. 39–47. URL: https://aclanthology.org/2023.findings-ijcnlp.4/. doi:10.18653/v1/2023.findings-ijcnlp.4.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022.

[29] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[30] Google, gemma-4-e2b-it: Instruction-tuned 2.3b parameter model, https://huggingface.co/google/gemma-4-E2B-it, 2026.

[31] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[32] M. A. Team, Ministral 3: A family of efficient small language models, arXiv preprint arXiv:2601.08584 (2026). URL: https://arxiv.org/abs/2601.08584.

[33] M. Zheng, Z. Li, T. Chen, M. Song, D. Wang, Hy-mt1.5 technical report, 2025. URL: https://arxiv.org/abs/2512.24092. arXiv:2512.24092.

[34] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: Proceedings of the 29th symposium on operating systems principles, 2023, pp. 611–626.

[35] Google DeepMind, Gemma 4: Open models for edge devices, 2026. URL: https://ai.google.dev/gemma/docs/core/model_card_4, accessed: May 27, 2026.

[36] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (2009) 427–437.