HeRE: A Novel Benchmark for Hebrew Relation Extraction via LLM-Guided Denoising of Knowledge Graph Alignments

Ron Keinan[1] **and Amir David Nissan Cohen**[2] **and Reut Tsarfaty**[1]
[1]Bar-Ilan University [2]Origin AI
{ronke21, amirdnc, reut.tsarfaty}@gmail.com

Abstract

While LLMs demonstrate strong capabilities in English Relation Extraction (RE) tasks, a significant gap remains in assessing their RE performance in low-resource languages, due to a lack of standardized benchmarks. Creating native, language-specific annotated RE benchmarks is expensive and requires human expertise, while relying on translation yields suboptimal, inaccurate datasets that introduce English-centric biases. To tackle this, we propose a scalable, language-agnostic methodology to create RE data by denoising naturally occurring yet noisy alignments of knowledge bases and corresponding texts. We define a proper alignment via a Natural Language Inference (NLI) task, where a relation triplet is entailed by an input text. We empirically assess five denoising strategies, and demonstrate that multilingual models outperform language-specific models in the few-shot LLM setting, while Hebrew-specific encoders remain competitive under NLI fine-tuning and are strongest in the label-free cross-training setting. We further find that prompting in the native language achieves performance on par with or slightly exceeding English prompts, confirming that the pipeline operates naturally in the target language. The best open-weights model, Gemma-4-31B-it (F1 = 0.915), matches the best API model, GPT-5.4 (0.921), and reaches F1 = 0.844 on a fully held-out, dual-annotated test set. Applying the optimized pipeline, we release HeRE: a novel, massive, 1.5-million-instance silver dataset for Hebrew RE, designed to support data-intensive experimentation and model fine-tuning. HeRE bridges a critical gap in standard Hebrew RE evaluation, while providing the NLP community with a reproducible, language-agnostic pipeline for RE dataset construction in other low-resource settings.

1 Introduction

Relation Extraction (RE) is a fundamental task in Natural Language Processing (NLP), which maps pairs of entity text spans to a predefined set of semantic relations (Dixit and Al-Onaizan, 2019), thereby bridging the gap between unstructured text and structured knowledge (Zhou et al., 2005). As a foundational component of Information Extraction, RE is essential for advanced downstream tasks, most notably Knowledge Base construction (Zhang et al., 2019; Niu et al., 2025), end-to-end information extraction pipelines (Keinan et al., 2025a), and Question Answering systems (Li et al., 2019; Swarup et al., 2025).

In high-resource languages such as English, pre-trained large language models (LLMs) now exhibit state-of-the-art proficiency in both extracting and classifying complex relational triplets (Zhao et al., 2024; Xue et al., 2024; Goyal and Singh, 2025). However, this progress remains heavily confined to high-resource languages. The NLP community critically lacks language-specific standardized benchmarks in low-resource languages (Pakray et al., 2025; Jinensibieke et al., 2024). Consequently, we lack the empirical resources necessary to evaluate whether the high performance of LLMs achieves true parity in low-resource settings (Huguet Cabot et al., 2023; Ali et al., 2025).

Overcoming this data bottleneck presents a dual challenge. First, generating natively annotated resources from scratch dictates a heavy reliance on specialized human domain and language expertise. This manual curation is not only financially prohibitive and time-consuming, but it also creates a severe scalability bottleneck, particularly for languages where qualified native annotators are scarce. Second, the common alternative of automatically machine-translating existing English benchmarks often introduces pervasive structural unnaturalness (Vanmassenhove et al., 2021; Zouhar et al., 2026). These synthetic pipelines may inject deep-seated linguistic biases, fundamentally failing to capture the authentic cultural context, richness, and semantic nuances inherent to the native target language

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5311–5329
November 6-10, 2026 ©2026 Association for Computational Linguistics

(Zahraei and Emami, 2025). To resolve this gap, we argue that the primary scientific challenge is not merely data generation, but rather effectively denoising naturally aligned, real-world information, without relying on exhaustive human annotation or machine translation.

We suggest a scalable, language-agnostic paradigm. Our approach capitalizes on the naturally occurring alignments between large-scale knowledge bases and their corresponding native text corpus, such as the Wikidata knowledge graph (Vrandečić and Krötzsch, 2014), and text sourced from Wikipedia abstracts. While this distant supervision paradigm offers a significant source of authentic relational data without relying on synthetic translation, a vanilla alignment between the text and the knowledge graph exhibits substantial noise. The co-occurrence of two entities in a sentence does not guarantee the expression of their underlying Wikidata relation, leading to a significant prevalence of false positives (Huguet Cabot and Navigli, 2021). Therefore, our core objective is to rigorously denoise these noisy natural alignments with a scalable, highly effective pipeline.

To determine the optimal architecture for automated data denoising, we conducted a comprehensive empirical study. We systematically evaluate a broad spectrum of modeling strategies, directly comparing transformer-encoder relation classifiers (Liu et al., 2022) and LLMs against encoders and LLMs fine-tuned for Natural Language Inference (NLI), whose entailment judgments are well studied as an annotation task (Kalouli et al., 2019). By framing text-relation alignment as an entailment task, where the native text acts as the premise and the relational triplet acts as the hypothesis, NLI provides a highly robust, logical mechanism to verify authentic semantic alignment and filter out noisy relations. By evaluating these strategies across both monolingual and multilingual paradigms of models and prompts, we execute a rigorous comparative analysis.

Ultimately, this empirical study demonstrates that multilingual LLMs outperform the competing architectures in the few-shot setting. This advantage comes with important exceptions: monolingual encoders remain competitive with multilingual encoders under NLI fine-tuning and outperform them when cross-trained on silver relation labels, and task-specific NLI fine-tuning narrows the multilingual advantage for decoder models.

Applying this methodology, we deliver **HeRE**, to our knowledge the first comprehensive **He**brew **R**elation **E**xtraction benchmark specifically engineered for Modern Hebrew.[1,2] We manually annotated a gold-standard validation set of 500 examples, used it to select the denoising method, and applied that method to the remaining noisy alignments to create a silver-standard dataset. HeRE provides a silver-standard dataset of 1.5 million instances, the scale needed to fine-tune Hebrew RE models. We also release two gold-standard sets of 500 rigorously human-annotated examples each: the validation set on which the pipeline was selected, and a held-out test set, never used for selection, that provides a reliable, culturally grounded benchmark for evaluating LLMs.

The best open-weights model (Gemma-4-31B-it) reaches F1 = 0.915 on the gold validation set, on which all model and prompt choices were made. Its performance is not significantly different from that of the top API model (GPT-5.4, F1 = 0.921) under paired bootstrap testing, supporting the feasibility of applying the pipeline to new languages without incurring API costs. Prompting models in Hebrew achieves performance on par with or above English, supporting the pipeline's native-language applicability to native, non-English texts. To confirm that the denoising quality generalizes beyond the gold benchmark, we measured how the selected configuration generalizes: on a second, fully held-out test set of 500 dual-annotated examples, on which nothing was tuned, the same model reaches F1 = 0.844 overall, and 0.776 on the 203 test examples whose predicates never appeared in the validation set (0.944 on predicates that did).

Our contributions are threefold. First, we introduce HeRE as a standardized benchmark for Hebrew RE, together with a quantified comparison to the filter it replaces (+0.131 F1 over CROCODILE's filter after tuning it and +0.509 over its published configuration, for the deployed Gemma-4-31B-it). Second, by open-sourcing our validated, language-agnostic dataset generation pipeline, we provide a reproducible blueprint for researchers to systematically generate localized RE benchmarks across the vast spectrum of other low-resource languages. Third, we demonstrate downstream utility by distilling HeRE into a provenance-grounded Hebrew knowledge graph, showing that, at matched graph size, consensus-denoised edges

[1]Code: https://github.com/Ronke21/HeRE

[2]Data, HeRE-KG and fine-tuned models: https://huggingface.co/datasets/ronke21/HeRE

support better link prediction than randomly selected ones (Sections 6 and 7).

2 The Challenge

Large-scale pre-trained models and, more recently, instruction-tuned LLMs have driven remarkable advancements in NLP throughout the last decade (Brown et al., 2020; Chang et al., 2024). These developments have accelerated progress on high-resource languages, particularly English (Qin et al., 2026), where LLMs now achieve near-human performance on a wide array of information extraction tasks, including RE (Wadhwa et al., 2023; Zhao et al., 2024). However, this progress has not transferred uniformly across languages. Low-resource languages continue to lag significantly behind their high-resource counterparts (Joshi et al., 2020; Hennig et al., 2023; Keinan and Hacohen-Kerner, 2023; Huguet Cabot et al., 2023; Jinensibieke et al., 2024; Keinan, 2024).

Despite these advances, the low-resource NLP landscape contains a critical gap: there is a lack of native RE benchmarks. We cannot reliably evaluate whether current LLMs, whether monolingual models or multilingual models, have acquired genuine RE competence in low-resource languages. The absence of a large training corpus for RE limits the ability to fine-tune models for Hebrew-specific relation patterns.

The two standard alternatives for creating benchmarks are both problematic at scale. Human annotation from scratch, while gold-standard, is prohibitively expensive at the volume needed for RE, requiring linguistically expert annotators to evaluate thousands of entity-relation-entity triples across diverse document types. Translating existing English benchmarks such as Re-TACRED (Stoica et al., 2021) or DocRED (Yao et al., 2019) introduces systematic biases: translated texts fail to reflect the cultural and encyclopedic context native to Hebrew Wikipedia, and morphological issues in machine-translated output can corrupt the entity spans that RE models must identify (Vanmassenhove et al., 2021).

We propose a third path that sidesteps both limitations.

3 The Proposal: Denoising Noisy Text-KG Alignments

Our starting point is the naturally occurring structural link between two complementary, large-scale resources: the Wikidata knowledge graph (Vrandečić and Krötzsch, 2014), which stores billions of structured factual triples of the form (s, p, o), and the Hebrew Wikipedia text corpus, where many of the same facts are expressed in natural language. Wikipedia articles contain hyperlinks to other pages, and each page corresponds to a unique Wikidata entity identifier, providing automatic entity grounding without a separate NER step. Exploiting this correspondence, we align every Hebrew Wikipedia passage in which two entities s and o co-occur to every Wikidata triple (s, p, o) in which they jointly participate, yielding our candidate corpus $\mathcal{C}$. This *distant supervision* paradigm (Mintz et al., 2009) provides a scalable, annotation-free source of native-language relational data. Its central drawback is noise: entity co-occurrence is only a weak signal that a passage actually expresses the target relation, so $\mathcal{C}$ contains a substantial fraction of false positives. Denoising $\mathcal{C}$ is therefore the core task we address.

Let $\mathcal{C} = \{(t_i, s_i, p_i, o_i)\}_{i=1}^{N}$ be a corpus of candidate tuples, where t_i is a text passage drawn from a native-language text corpus, s_i and o_i are entity mentions appearing in t_i, and $p_i \in \mathcal{P}$ is a relation type from a knowledge graph's predicate vocabulary $\mathcal{P}$.

The denoising task is to assign a binary label $y_i \in \{0, 1\}$ to each tuple: $y_i = 1$ if and only if t_i explicitly or clearly implicitly asserts that s_i stands in relation p_i to o_i, and $y_i = 0$ otherwise. Tuples in $\mathcal{C}$ are produced by distant supervision (any passage where s_i and o_i co-occur is aligned to every knowledge-graph triple (s_i, p_i, o_i) in which they participate), yielding a large but noisy set with an unknown fraction of false positives. Our goal is to learn a denoising function $f : \mathcal{C} \to \{0, 1\}^N$ that produces a reliably labeled dataset $\hat{\mathcal{C}} \subseteq \mathcal{C}$.

We build on CROCODILE (Huguet Cabot and Navigli, 2021) to construct the Hebrew candidate corpus (Appendix C) and replace its single NLI filter with a systematic comparison of five denoising strategies.

Rather than committing to a single denoising architecture, we adopt a **multi-strategy approach**: we benchmark five architecturally distinct denoising strategies. The pipeline proceeds in two stages (Figure 1): in the **data preparation stage**, we clean Wikipedia markup and generate three NLI hypothesis variants per triple; in the **denoising stage**, we apply and compare all five strategies on a gold-standard evaluation set.

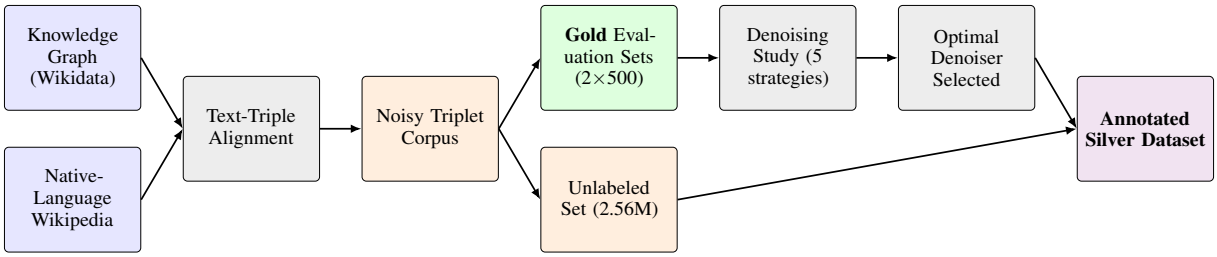


Figure 1: The HeRE construction pipeline.

4 Denoising Strategies

We define five strategies for binary KG-text alignment validation. Each takes a tuple (t, s, p, o) and a triple verbalization h (Section 5.2) as input and produces a binary prediction $\hat{y} \in \{0, 1\}$.

Strategy 1: NLI Encoders. A pre-trained encoder is fine-tuned on a Hebrew NLI corpus as a 3-class classifier (entailment / neutral / contradiction). At inference time, t is the NLI premise and h is the hypothesis. An entailment-based score ($P(e)$, $P(e)-P(c)$, or $P(e)/(P(e)+P(c))$) is thresholded to produce $\hat{y}$. This strategy is fast and GPU-efficient but requires an NLI training corpus transferable to the target language.

Strategy 2: Fine-tuned LLM NLI. A decoder LLM is fine-tuned on the same NLI corpus using LoRA (Hu et al., 2022). At inference time it is prompted in NLI format; the binary prediction is derived from first-token log-probabilities of the entailment, neutral, and contradiction label tokens. Fine-tuning adapts the model to the entailment verification framing without requiring gold RE labels.

Strategy 3: Cross-trained Predicate Classifiers. Encoder models are trained as multi-class predicate classifiers directly from the raw distant-supervision corpus using K-fold cross-training, with no gold labels or NLI data required. For each fold i, a model is fine-tuned on the remaining $K-1$ folds to predict which predicate p is expressed given (t, s, o) as input. A correct prediction on the held-out fold signals a true positive; an incorrect or uncertain prediction signals a false positive. The final label is majority vote across all K fold-models.

Strategy 4: Open-source LLMs (Few-shot). An open-source instruction-tuned LLM is prompted in a binary yes/no format: *does passage t assert relation p between s and o?* This is a binary textual entailment question, not a generative extraction task. No fine-tuning is required; prompt language (Hebrew or English) and triple verbalization style are evaluated as independent design choices, with in-context examples provided for calibration.

Strategy 5: Foundation LLMs (Few-shot). Commercial frontier models are prompted in the same binary format as Strategy 4, accessed via API. Only hard-label predictions are used; token probabilities are not. These models represent the practical performance ceiling but are cost-prohibitive for large-scale silver annotation.

5 Experimental Settings

We evaluate all five denoising strategies on a common gold-standard benchmark.

5.1 Data

Applying CROCODILE to an early-2025 Hebrew Wikipedia dump (`heb25`) and the matching Wikidata snapshot yields 3.1 million candidate (t, s, p, o) tuples spanning 1,137 distinct predicates after entity-field normalization. Each tuple pairs a section-level passage from Hebrew Wikipedia with a Wikidata triple where both s and o are mentioned; passage lengths vary considerably (mean 297 words after markup cleaning).

5.2 Hypothesis Generation

For each triple we generate three NLI hypothesis variants used across all strategies:

- **Basic**: A surface-form concatenation: *"{subject} {predicate} {object}"*. Straightforward but lacks the grammatical structure of a natural assertion.
- **Template**: A predicate-specific Hebrew sentence from hand-crafted templates covering all predicates in the set. For example, the *place of birth* template produces: *"{subject} is the place of birth of {object}"*. Templates

Statistic	Gold val.	Gold test	Silver
Examples	500	500	1,491,070
Unique documents	454	494	∼228K
Triples per document	1.1	1.0	∼6.5
Unique predicates	97	124	1,137
Predicates w/ ≥20 ex.	4	1	593
Positive rate	61.4%	52.0%	72.5%
Pos. rate std (pred.)	40.6%	40.9%	n/a
Mean text (words)	70	424	297
Mean text (chars)	428	2,522	1,852

Table 1: Dataset statistics.

are grammatically natural and semantically precise.

- **LLM Relation**: A fluent Hebrew sentence generated by Gemma-3-27B-it (Gemma Team et al., 2025) from the triple, prompted to express the relation in a single natural sentence.

5.3 Models

For **Strategy 1** we fine-tune six encoders on HebNLI[3]: XLM-RoBERTa-large/XL, mE5-large, mmBERT, NeoDictaBERT, and AlephBERT.

Strategy 2 fine-tunes nine decoder LLMs on HebNLI: Gemma-2-9B, Gemma3-12B-it, Gemma-3-27B-it, Gemma-4-26B-A4B, Gemma-4-31B, DictaLM-3.0-1.7B-Instruct, DictaLM-3.0-24B, Aya-Expanse-32b, and Qwen3.5-35B-A3B-Base. NLI fine-tuning details and HebNLI performance for all models are reported in Appendix D.

Strategy 3 cross-trains XLM-RoBERTa-large, mmBERT, NeoDictaBERT, and mE5-large with $K \in \{3, 5, 7, 10\}$ (predicates with <20 silver examples excluded).

In **Strategy 4** we evaluate DictaLM-3.0-24B-Base, Gemma-3-27B-it, Gemma-4-26B-A4B-it, Gemma-4-31B-it, and Aya-Expanse-32b at 0/2/5 shots in Hebrew and English with both hypothesis types.

In **Strategy 5** we evaluate GPT-5.4, Gemini 3 Flash, and Claude Haiku 4.5[4] at 0 and 5 shots. Full setup details are in Appendix B.

5.4 Evaluation

All strategies are evaluated as binary classifiers on the 500-example gold validation set (Section 7; soft scores are converted to hard predictions as described in Appendix B). Our primary metric is **binary F1 on the positive class** (relation present). F1 is the most informative metric here because the class distribution is imbalanced (61.4% positive) and both false positives and false negatives carry distinct costs in a denoising pipeline. Because $n = 500$ bounds the resolution of any comparison, every reported F1 carries a 95% confidence interval from a 10,000-resample percentile bootstrap, and model pairs are compared with a paired bootstrap on the same examples; the mean CI half-width is ± 0.022. Since the positive-class prior varies sharply across predicates (Table 12, Appendix G), we also report macro-F1 averaged over predicates alongside the micro-averaged headline.

5.5 Downstream Evaluation on HeRE-KG

To test whether consensus denoising yields measurably better downstream models than the raw alignments it filters, we train TransE (Bordes et al., 2013) on four variants of the HeRE knowledge graph (Section 7) and evaluate filtered link prediction on one shared held-out set of 20,000 triples removed from every training graph, scoring the ∼14.2K of them whose entities and relation occur in all four training graphs so that no variant is penalised for unseen vocabulary: **raw**, all unique distant-supervision triples; **denoised**, the `agree3` subset (unanimous positive vote of Gemma-4-31B-it, Gemma-3-27B-it and the HebNLI-tuned DictaLM-3.0-24B; Section 7); a **size-matched random control**, a uniform sample of raw with exactly as many triples as denoised; and **confidence-weighted**, the full raw graph with training triples sampled in proportion to mean signal confidence. The control separates the effect of selecting better edges from that of simply training on fewer. Held-out triples are sampled from the denoised graph and removed from every training graph; splitting is over unique triples, since a row-level split leaks other mentions of a test triple into training. We report mean and standard deviation over three seeds.

6 Results

Table 2 reports the best F1 per model on the 500-example gold validation set, the set on which all model and configuration choices were made (the held-out test set is reserved for the generalization analysis below); complete results appear in Tables 7–11 (Appendix F).

[3] https://github.com/NNLP-IL/HebNLI

[4] Accessed through OpenRouter in August 2026 as `openai/gpt-5.4`, `google/gemini-3-flash-preview`, and `anthropic/claude-haiku-4-5`.

Model	Type	F1	Macro
Baseline: CROCODILE NLI filter			
XLM-R-large-XNLI (as published)	multi	0.406	0.299
XLM-R-large-XNLI (best tuned)	multi	0.784	0.595
Strategy 1: Fine-tuned NLI Encoders			
AlephBERT	mono	0.618	0.560
mE5-large	multi	0.692	0.614
mmBERT	multi	0.732	0.614
XLM-R-large	multi	0.756	0.680
NeoDictaBERT	mono	0.771	0.668
XLM-R-XL	multi	**0.779**	0.753
Strategy 2: Fine-tuned LLM NLI			
DictaLM-3.0-1.7B-Instruct	mono	0.429	0.259
Gemma-4-26B-A4B	multi	0.465	0.392
Gemma3-12B-it	multi	0.623	0.449
Gemma-3-27B-it	multi	0.726	0.596
DictaLM-3.0-24B-Base	mono	0.803	0.676
Aya-Expanse-32b	multi	0.829	0.702
Gemma-2-9B	multi	0.829	0.682
Gemma-4-31B	multi	0.849	0.760
DictaLM-3.0-24B	mono	0.861	0.718
Qwen3.5-35B-A3B-Base	multi	**0.873**	0.734
Strategy 3: Cross-trained Predicate Classifiers			
mE5-large	multi	0.675	0.574
XLM-R-large	multi	0.729	0.613
mmBERT	multi	0.743	0.631
NeoDictaBERT	mono	**0.751**	0.659
Strategy 4: Open-source LLMs			
DictaLM-3.0-24B-Base	mono	0.835	0.717
Aya-Expanse-32b	multi	0.847	0.732
Gemma-3-27B-it	multi	0.877	0.726
Gemma-4-26B-A4B-it	multi	0.880	0.743
Gemma-4-31B-it	multi	**0.915**	0.771
Strategy 5: Foundation LLMs			
Claude Haiku 4.5	multi	0.890	0.765
Gemini 3 Flash	multi	0.917	**0.798**
GPT-5.4	multi	**0.921**	0.779

Table 2: Best micro-F1 and macro-F1 per model on the gold validation set (adjudicated labels; macro-F1 over predicates with at least 10 examples).

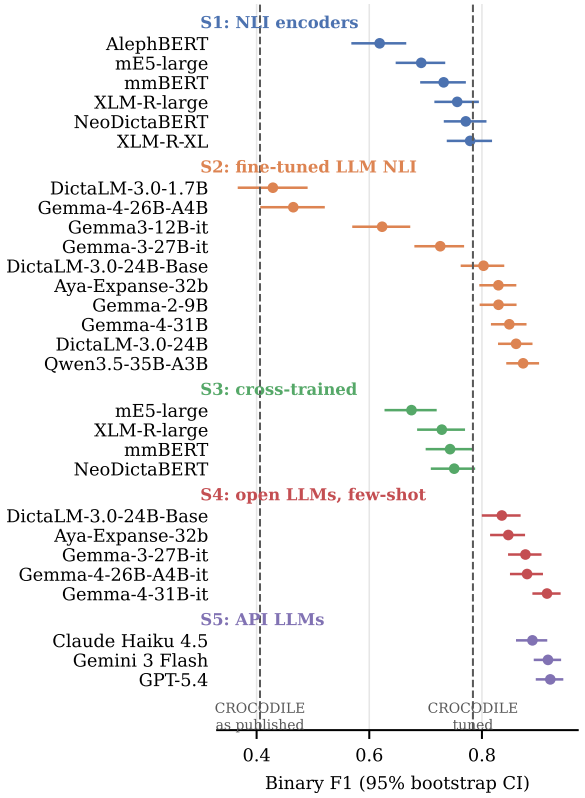


Figure 2: Best micro-F1 per model on the gold validation set with 95% bootstrap confidence intervals, by strategy. Dashed lines: the CROCODILE filter as published (0.406) and after tuning (0.784).

HeRE outperforms the CROCODILE baseline it replaces. CROCODILE's single XLM-R NLI filter, as published (P(entailment) >0.75 over a surface-concatenated hypothesis), has precision 0.830 but recall 0.269 on Hebrew: it discards two thirds of genuine relations. Tuned, it remains 0.131 below the deployed Gemma-4-31B-it and 0.137 below GPT-5.4, with non-overlapping confidence intervals. Changing only its hypothesis format helps more than retuning only its threshold (+0.32 vs. +0.16 F1), corroborating the hypothesis-format finding below; the tuned XNLI model even edges the HebNLI-fine-tuned XLM-R-large of Strategy 1, so for encoders, hypothesis format and threshold matter more than the NLI corpus.

Open-weights LLMs match the best API models. GPT-5.4, Gemini 3 Flash and the open-weights Gemma-4-31B-it form one statistical cluster (Figure 2): a paired bootstrap between the best configuration and the next five (the Hebrew and English Gemini 3 Flash prompts and three Gemma-4-31B-it prompts) finds no distinguishable difference (largest $\Delta = +0.007$, CI $[-0.012, +0.027]$). The best open-weights model is thus not measurably worse than the best commercial API, which is what makes the pipeline reproducible.

Macro-F1 reveals much weaker per-predicate performance. Across all configurations macro-F1 trails micro-F1 by 0.115 on average, and the ranking changes under macro averaging: Gemini 3 Flash is the most balanced system while Gemma-4-31B-it leads on micro only (Table 2).

Cross-trained classifiers. Cross-training needs neither gold labels nor NLI data but is the costliest strategy (∼480 GPU-hours for the 16-configuration grid, 8–28 per configuration on an A100) and trails Strategy 1's best encoder. NeoDictaBERT leads the multilingual encoders at $K \in \{3, 5, 7\}$ (by 0.01–

0.02, driven by recall) and ties mmBERT at $K=10$; F1 varies by at most ± 0.01 with K, so the number of folds is not a tuning axis (Table 9).

Multilingual LLMs win at few-shot prompting; Hebrew encoders win at cross-training. The multilingual advantage is task- and scale-dependent. In few-shot prompting (Strategy 4) multilingual models dominate: DictaLM-3.0-24B trails Gemma-4-31B-it by 0.08, though this comparison is neither size-controlled (24B vs. 31B) nor tuning-controlled (a base model against instruction-tuned ones; Appendix B). For encoders the pattern reverses or vanishes: XLM-R-XL and NeoDictaBERT are tied in Strategy 1, and NeoDictaBERT leads every multilingual encoder in Strategy 3. NLI fine-tuning narrows the gap for decoders: Qwen3.5-35B leads DictaLM-3.0-24B by a margin inside the confidence interval.

Structured hypotheses outperform plain triple concatenation. A structured hypothesis beats basic concatenation for every encoder, by up to 0.10 F1 for NeoDictaBERT (Table 7), and the template hypothesis is best for every fine-tuned decoder (Table 8). Few-shot and API models are less sensitive to hypothesis form (Table 10).

Prompting in Hebrew works as well as in English. The best Hebrew and English configurations are within 0.002 of each other for Gemma-4-31B-it and Gemini 3 Flash, and Hebrew leads by 0.013 for GPT-5.4 (Tables 10 and 11). Frontier LLMs, though trained predominantly on English, are proficient enough in Hebrew for native-language RE classification.

Denoising improves downstream link prediction. At matched training size the denoised graph outperforms the random control by 24% relative MRR and 29% relative Hits@10, consistently across seeds (Table 3): the consensus signal selects edges that are genuinely more learnable, not merely fewer. The full raw graph still scores highest in absolute terms, because link prediction rewards coverage; denoising trades coverage for precision. Confidence-weighted sampling keeps raw's coverage but does not improve on uniform sampling (MRR 0.150 vs. 0.161), so for link prediction the per-signal scores are best used to choose an operating point, not to reweight training.

Generalization to the held-out test set. Validation scores are optimistic, since configurations were selected on that set. On the held-out test set (Table 4) Gemma-4-31B-it loses 0.07 F1 overall and 0.14 on the 203 unseen-predicate examples, while on seen predicates it does not drop at all: the loss is predicate transfer, not overfitting to the annotation scheme. Since the test set is a stress test rather than a sample of the silver distribution (40% unseen predicates, passages 40% longer), single-model quality on silver lies between the unseen-predicate and the silver-drawn subset figures. Across every signal deployed on the silver corpus (Table 5) the top of the ranking mirrors validation, the seven-signal majority vote is the best rule that labels every row, and the published `agree3` label is the most accurate on the two thirds it covers (Table 16), and the gap to the encoder and cross-trained signals widens on these longer passages.

Training graph	Triples	MRR	Hits@10
Raw (all alignments)	1,479,292	0.161 ± 0.001	0.307 ± 0.003
Confidence-weighted	1,479,292	0.150 ± 0.003	0.289 ± 0.005
Denoised (agree3)	840,991	$\mathbf{0.093} \pm 0.002$	$\mathbf{0.189} \pm 0.002$
Size-matched control	840,991	0.075 ± 0.001	0.147 ± 0.002

Table 3: Link prediction on HeRE-KG. Triples are unique (s, p, o): HeRE-KG's 1,080,605 evidence-linked edges are mentions of 860,990 unique triples, of which 20,000 were drawn for the held-out set and removed from every training graph; the 14,163–14,213 scored per seed are those whose entities and relation appear in every training graph.

Evaluation set	n	F1
Gold validation (adjudicated)	500	0.915
Gold test (held out)	**500**	**0.844**
seen predicates (depth)	97	0.944
unseen predicates	**203**	**0.776**
silver-validation subset	200	0.885

Table 4: Gemma-4-31B-it: validation set vs. held-out test set and its subsets.

Error analysis. Errors fall into three patterns (Appendix E): predicates with no positive gold examples that many configurations still predict, low-prior geographic predicates such as *borders*, and a small set of annotation-ambiguous cases.

7 HeRE: The Hebrew Relation Extraction Benchmark

Drawing on the empirical study above, we release HeRE, the first comprehensive Hebrew RE benchmark, consisting of three complementary resources.

Model	Str.	F1	95% CI	Macro
Gemma-4-31B-it	S4	0.844	[0.808, 0.877]	0.619
Gemma-3-27B-it	S4	0.829	[0.794, 0.862]	0.543
DictaLM-3.0-24B-Base	S4	0.725	[0.682, 0.766]	0.426
DictaLM-3.0-24B	S2	0.745	[0.703, 0.785]	0.454
Gemma-2-9B‡	S2	0.332	[0.271, 0.390]	0.113
Aya-Expanse-32b‡	S2	0.159	[0.103, 0.217]	0.132
NeoDictaBERT	S1	0.691	[0.644, 0.734]	0.419
XLM-R-large	S1	0.423	[0.360, 0.484]	0.290
NeoDictaBERT (K=3)	S3	0.652	[0.600, 0.699]	0.401
mmBERT (K=5)	S3	0.544	[0.486, 0.600]	0.372

Table 5: Held-out test-set F1 of each signal deployed on the silver corpus. ‡Deployed at score threshold 0.9. Ensembles: Table 16.

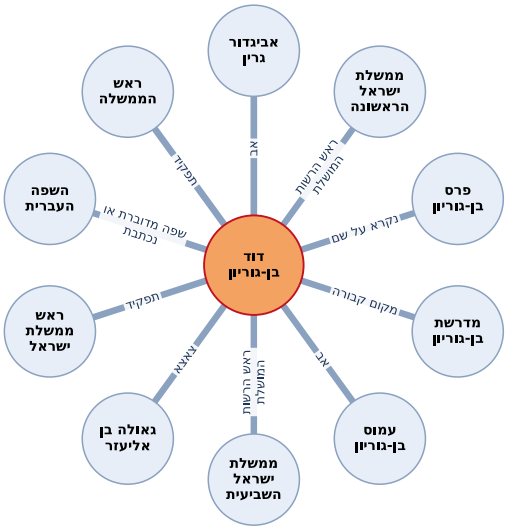


Figure 3: HeRE-KG neighbourhood example: ten highest-confidence edges (width = mean confidence).

Gold datasets. We sampled 500 examples from the vanilla-alignment corpus, stratified so that each predicate's share reflects its frequency in the full corpus, and, to keep annotation tractable, drawn from shorter passages (mean 70 words vs. 297 in the corpus; Table 1), an effect measured in Section 6. Each example was labelled for *relation_present* (does the passage genuinely express the target Wikidata relation?) independently by two native Hebrew speakers with NLP expertise following the guidelines in Appendix A; inter-annotator agreement is Cohen's $\kappa = 0.809$, disagreements were partially adjudicated, and all results are reported against the adjudicated labels. This *validation* set, on which every model and configuration choice was made, covers 97 predicates with 307 positive examples (61.4%); the standard deviation of per-predicate positive rates is 40.6% (per-predicate rates in Table 12, Appendix G), underscoring the need for denoising. A second, fully held-out *test* set of 500 examples was dual-annotated in the same way ($\kappa = 0.757$) and structured to probe generalization: 203 examples carry predicates absent from the validation set, 97 add depth on seen predicates, and 200 are drawn from the silver distribution. Both sets are released as a culturally grounded evaluation resource for Hebrew RE systems.

Silver dataset. After removing the 500 validation examples and exact duplicates, 2,564,534 candidate tuples remain as the raw silver corpus. Because API models cannot be applied to 2.5 million examples at acceptable cost, we deploy ten open-weights signals from Strategies 1 to 4 on the full silver corpus (Table 15, Appendix G). To construct the published silver dataset, we require unanimous agreement among the three strongest open-weights models that were applied to the full corpus (Gemma-4-31B-it and Gemma-3-27B-it from Strategy 4 and DictaLM-3.0-24B from Strategy 2; the Qwen3.5-35B and Gemma-4-26B-A4B-it runs were evaluated on gold only), hereafter `agree3`. This yields 1,080,605 positive examples and 410,465 negative examples (1,491,070 total; 58.1% of the silver corpus, after withholding the unanimous rows that belong to the held-out test set), with a positive rate of 72.5% within the published set. Per-signal positive rates are given in Table 15 (Appendix G). Their positive rates range from 54.0% (Gemma-4-31B-it) to 70.4% (Gemma-3-27B-it), reflecting calibration differences. Requiring unanimous agreement across all three models (`agree3`) retains 58.1% of silver at a substantially higher confidence threshold. The release is not limited to this binary label: every example ships with the raw score and hard prediction of each of the ten deployed signals and with four consensus rules (Table 16), so users can choose their own precision/recall operating point or build alternative consensus rules. All 500 test-set rows, which were drawn from the silver corpus, are withheld from every released artifact. The published `agree3` label reaches precision 0.899, recall 0.934 and F1 = 0.916 (Table 16, Appendix G).

HeRE-KG. We additionally release the positively labelled portion of HeRE as a knowledge graph in which every edge carries the Hebrew Wikipedia passage that expresses it and the per-

signal confidence scores behind its label, a textual grounding Wikidata does not provide for Hebrew. HeRE-KG holds 1,080,605 evidence-linked edges over 424,216 entities and 917 relation types (306 with at least 100 edges) from 212,332 documents; 95.6% of entities lie in one connected component (mean degree 5.1). Figure 3 shows one entity's neighbourhood.

8 Related Work and Future Directions

Relation Extraction Datasets. English RE benchmarking spans sentence-level corpora such as ACE (Doddington et al., 2004) and Re-TACRED (Stoica et al., 2021), document-level DocRED (Yao et al., 2019), and the large-scale REBEL (Huguet Cabot and Navigli, 2021) with its multilingual successor RED$^{\text{FM}}$ (Huguet Cabot et al., 2023), which use the same Wikidata-Wikipedia alignment paradigm we adopt. Despite this progress, native-language RE benchmarks for low-resource languages remain scarce.

Distant Supervision and Denoising. Distant supervision for RE was formalized by Mintz et al. (2009) and has underpinned large-scale corpus construction ever since. A central challenge is noise: entity co-occurrence does not guarantee relational expression. Prior denoising approaches include attention-based instance selection (Lin et al., 2016) and entailment-based relation verification (Sainz et al., 2021). CROCODILE (Huguet Cabot and Navigli, 2021) applies a single lightweight XLM-RoBERTa NLI filter. Building directly upon their pipeline, we expand this denoising phase into a systematic comparative study of five architectures, establishing multilingual LLMs as the most effective choice in the few-shot setting, while showing that monolingual encoders with NLI fine-tuning remain competitive for low-resource RE benchmark construction.

Multilingual Task Benchmarks. Cross-lingual benchmarks exist for dependency parsing (Nivre et al., 2016), morphology (Batsuren et al., 2022), question answering (Lewis et al., 2020; Clark et al., 2020), and summarization (Hasan et al., 2021). RE remains among the least-resourced multilingual tasks; to our knowledge, HeRE is the first native Hebrew RE benchmark.

Hebrew NLP Resources. Hebrew-specific encoders include AlephBERT (Seker et al., 2022), DictaBERT (Shmidman et al., 2023), and NeoDictaBERT (Shmidman et al., 2025); the DictaLM family (Shmidman et al., 2024) provides native Hebrew generative LLMs. Task benchmarks cover NER (Bareket and Tsarfaty, 2021), morphological parsing (Sade et al., 2018), summarization (Paz-Argaman et al., 2024), QA (Cohen et al., 2023), coreference resolution (Greenfeld and Tsarfaty, 2026), and various classification tasks (Keinan et al., 2025b, 2026). To our knowledge, Hebrew RE has had no native benchmark prior to this work.

Future Directions. The HeRE silver dataset enables end-to-end Hebrew RE model fine-tuning. The language-agnostic pipeline is directly applicable to other languages with sufficient Wikidata-Wikipedia coverage: its only data prerequisite is a Wikipedia edition linked to Wikidata, which exists for more than 300 languages, most of them low-resource in research attention and LLM pretraining exposure; extending to Arabic, Persian, and other Semitic languages is a natural next step. Because HeRE-KG edges are passage-grounded (Section 7), they can also be proposed as candidate Wikidata statements for the sparse Hebrew coverage there, each with its supporting passage as the reference Wikidata requires.

9 Conclusions

We have presented HeRE, the first Hebrew Relation Extraction benchmark, together with a language-agnostic pipeline for constructing RE benchmarks via Wikidata-Wikipedia distant supervision and empirically optimized denoising. Across five denoising strategies, multilingual LLMs outperform Hebrew-specific models in the few-shot open-source setting, though monolingual encoders remain competitive, and in the label-free cross-training setting the Hebrew-specific encoder is the strongest classifier. The best open-weights model (Gemma-4-31B-it, F1 = 0.915) is statistically indistinguishable from the top commercial API (GPT-5.4, F1 = 0.921). Hebrew prompting matches English, and on a fully held-out, dual-annotated test set denoising reaches F1 = 0.844 (0.776 on unseen predicates). As a knowledge graph, the corpus yields over a million provenance-grounded edges, and consensus denoising improves link prediction over a size-matched control. The released datasets, per-signal scores and pipeline are a reproducible blueprint for RE benchmarks in any language with a Wikidata-linked Wikipedia.

10 Limitations

Domain scope. HeRE is built from Hebrew Wikipedia passages aligned to Wikidata triples, so its text is encyclopedic; models trained or evaluated on it may not transfer to social media, clinical, news or legal Hebrew.

Predicate coverage. The gold validation and test sets cover 97 and 124 of the 1,137 predicates, few with 20 or more examples: a deliberate consequence of the long-tail setting, which no annotation budget can cover densely. We quantify rather than assume its cost (0.944 on seen vs. 0.776 on unseen predicates, §6), but F1 on rare predicates remains indicative.

Passage length. Gold passages are shorter than silver ones (mean 70 vs. 297 words). Stratifying the test set by length, F1 declines from 0.909 (100–250 words) to 0.835 (250–500) and 0.818 (500+; Table 14), a consistent but small trend whose bucket confidence intervals overlap, so length alone does not explain the validation-to-test gap.

Language scope. The pipeline is language-agnostic by design but evaluated only on Modern Hebrew; it presumes adequate Wikidata coverage and capable encoders or LLMs, and transfer to a second language, the most direct test of its generality, is left to future work.

11 Ethical Considerations

Data Licensing. HeRE is derived from Hebrew Wikipedia text and Wikidata, both of which are freely licensed under Creative Commons and open-data licenses respectively. We do not introduce any private or personally sensitive data; all entity and relational information originates from publicly available knowledge bases.

Annotation. Both gold sets were annotated independently by two paid expert annotators compensated at a fair academic rate, with disagreements adjudicated; inter-annotator agreement is reported in Section 7. Annotation guidelines were reviewed and refined iteratively to minimize ambiguity (see Appendix A).

Acknowledgments

This research has been funded by the following funding agencies: a grant from the Israeli Science Foundation (ISF grant 670/23), a KAMIN grant by the Israeli Innovation Authority (IIA), and a VATAT grant by the Israeli Planning and Budgeting Committee (PBC), for which we are grateful.

References

Manzoor Ali, René Speck, Hamada M. Zahera, Muhammad Saleem, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. Multilingual relation extraction: A survey. *IEEE Access*, 13:151907–151933.

Dan Bareket and Reut Tsarfaty. 2021. Neural modeling for named entities and morphology (NEMO^2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, and 1 others. 2022. UniMorph 4.0: Universal morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Amir DN Cohen, Hilla Merhav, Yoav Goldberg, and Reut Tsarfaty. 2023. HeQ: a large and diverse Hebrew reading comprehension benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13693–13705, Singapore. Association for Computational Linguistics.

Kalpit Dixit and Yaser Al-Onaizan. 2019. Span-level model for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5308–5314, Florence, Italy. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840, Lisbon, Portugal. European Language Resources Association (ELRA).

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nandita Goyal and Navdeep Singh. 2025. Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, 618:129171.

Refael Shaked Greenfeld and Reut Tsarfaty. 2026. Beyond word boundaries: A Hebrew coreference benchmark and an evaluation protocol for morphologically complex text. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10669–10683, San Diego, California, United States. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Leonhard Hennig, Philippe Thomas, and Sebastian Möller. 2023. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED^{fm}: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

Dawulie Jinensibieke, Mieradilijiang Maimaiti, Wentao Xiao, Yuanhang Zheng, and Xiaobo Wang. 2024. How good are LLMs at relation extraction under low-resource scenario? Comprehensive evaluation. *arXiv preprint arXiv:2406.11162*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Martha Palmer, and Valeria de Paiva. 2019. Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics.

Ron Keinan. 2024. Text mining at SemEval-2024 task 1: Evaluating semantic textual relatedness in low-resource languages using various embedding methods and machine learning regression models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 420–431, Mexico City, Mexico. Association for Computational Linguistics.

Ron Keinan, Amir David Nissan Cohen, and Reut Tsarfaty. 2025a. From named entities to relations: End-to-end biomedical information extraction. In *CLEF (Working Notes)*, pages 359–372.

Ron Keinan and Yaakov Hacohen-Kerner. 2023. JCT at SemEval-2023 tasks 12 a and 12B: Sentiment analysis for tweets written in low-resource African languages using various machine learning and deep learning methods, resampling, and HyperParameter tuning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 365–378, Toronto, Canada. Association for Computational Linguistics.

Ron Keinan, Efraim Margalit, and Dan Bouhnik. 2025b. Emotional analysis in a morphologically rich language: Enhancing machine learning with psychological feature lexicons. *Electronics*, 14(15):3067.

Ron Keinan, Efraim Margalit, and Dan Bouhnik. 2026. Depression detection with machine learning in a low-resource language: A case study in Hebrew digital health communities. *Array*, 31:101157.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Wanlong Liu, Li Zhou, Dingyi Zeng, and Hong Qu. 2022. Document-level relation extraction with structure enhanced transformer encoder. In *2022 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Songjie Niu, Kaisen Yang, Rui Zhao, Yichao Liu, Zonglin Li, Hongning Wang, and Wenguang Chen. 2025. Tree-KG: An expandable knowledge graph construction framework for knowledge-intensive domains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18516–18529, Vienna, Austria. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Tzuf Paz-Argaman, Itai Mondshine, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. HeSum: a novel dataset for abstractive text summarization in Hebrew. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6378–6388, Bangkok, Thailand. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2026. Large language models meet NLP: A survey. *Frontiers of Computer Science*, 20(11):2011361.

Shoval Sade, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting LLMs to Hebrew: Unveiling DictaLM 2.0 with enhanced vocabulary and instruction capabilities. *arXiv preprint arXiv:2407.07080*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. DictaBERT: A state-of-the-art BERT suite for modern Hebrew. *arXiv preprint arXiv:2308.16687*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2025. NeoDictaBERT: Pushing the frontier of BERT models for Hebrew. *arXiv preprint arXiv:2510.20386*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-TACRED: Addressing shortcomings of the TACRED dataset. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 13843–13850.

Anushka Swarup, Tianyu Pan, Ronald Wilson, Avanti Bhandarkar, and Damon Woodard. 2025. LLM4RE: A data-centric feasibility study for relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6670–6691, Abu Dhabi, UAE. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. AutoRE: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Pardis Sadat Zahraei and Ali Emami. 2025. Translate with care: Addressing gender bias, neutrality, and reasoning in large language model translations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 476–501, Vienna, Austria. Association for Computational Linguistics.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3016–3025, Minneapolis, Minnesota. Association for Computational Linguistics.

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

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.

Vilém Zouhar, Niyati Bafna, Mukund Choudhary, Maike Züfle, Sara Rajaee, Pinzhen Chen, Jannis Vamvas, Sara Papi, Ona de Gibert, Bhavitvya Malik, and 1 others. 2026. Last translation benchmark. *arXiv preprint arXiv:2609.04173*.

A Annotation Guidelines

The following guidelines were used by both annotators to assign the binary label *relation_present* to each example in the gold sets.

Task Definition. Given a Hebrew text passage, a subject entity, a Wikidata predicate, and an object entity, the annotator must determine: *Does this passage explicitly or clearly implicitly express the stated relation between the subject and the object?*

Positive (relation present). The label is **1** (relation present) when:

- The passage contains a sentence or clause that directly states the relation (e.g., "X was born in Y" for a *place of birth* triple).
- The relation can be unambiguously inferred from the passage within a single sentence or adjacent sentences, without requiring external knowledge.

Negative (relation absent). The label is **0** (relation absent) when:

- The passage mentions both entities but does not express the stated relation between them.
- The passage expresses a related but distinct relation (e.g., expressing that X *lives in* Y rather than *was born in* Y for a *place of birth* triple).
- The relation can only be inferred by combining information across multiple distant passages or requires external knowledge.
- The passage is ambiguous as to which of two possible objects the predicate applies.

Borderline Cases. Annotators were instructed to default to **0** (relation absent) in genuinely ambiguous cases, and to flag examples where the predicate definition itself was unclear for adjudication.

B Experimental Setup

Hardware. Experiments were run on NVIDIA A100-80GB GPUs, with the largest decoders (24B–35B parameters) on B200 nodes and the final Strategy-3 grid cells on V100-32GB nodes. Encoder NLI models and the cross-trained predicate classifiers were run on a single GPU where memory allowed; the complete cross-training grid (four encoders, $K \in \{3, 5, 7, 10\}$, matched training volume of 2.05M examples per fold) consumed approximately 480 GPU-hours in total (part of the grid ran on slower V100 GPUs), or 8–28 GPU-hours per configuration on an A100 depending on K. Fine-tuned LLM NLI inference (DictaLM-3.0-24B, Aya-Expanse-32b, Gemma-2-9B) used 1–2 A100s via vLLM with LoRA adapter loading. Few-shot open-source LLM experiments (Gemma-3-27B-it, Gemma-4-31B-it) used tensor parallelism across 2 GPUs.

NLI Encoder Fine-tuning. Encoder models were fine-tuned on HebNLI for 3-class NLI classification (entailment / neutral / contradiction) with the hyperparameters given in Appendix D, selecting the checkpoint with the best validation macro-F1. The checkpoints used in our experiments are: XLM-RoBERTa-large (checkpoint-8000), XLM-RoBERTa-XL (checkpoint-8000), mE5-large (checkpoint-6000), mmBERT (checkpoint-5800), NeoDictaBERT (checkpoint-5600), AlephBERT (checkpoint-4600).

LLM NLI Fine-tuning. LoRA adapters for all nine Strategy-2 decoders were trained on HebNLI with rank-16 LoRA over all linear projection matrices (hyperparameters in Appendix D); the best checkpoint of each model (step 5,280 or 7,920) was used in all NLI inference experiments.

API models. GPT-5.4, Gemini 3 Flash and Claude Haiku 4.5 were accessed through OpenRouter in August 2026 as `openai/gpt-5.4`, `google/gemini-3-flash-preview`, and `anthropic/claude-haiku-4-5`; request parameters are recorded in the released code.

Strategy 4 model choice. DictaLM-3.0-24B has no instruction-tuned release; we prompt its base model, since its Thinking variant is too costly for corpus-scale prompting. The few-shot comparison between Hebrew-specific and multilingual models is therefore between a base model and instruction-tuned ones.

Released models. Alongside the data, the HeRE collection on the Hugging Face Hub (footnote in Section 1) releases the fine-tuned denoisers needed to regenerate the published signals, together with the per-strategy winners: the HebNLI LoRA adapter of DictaLM-3.0-24B (Strategy 2; one of the three `agree3` models), the HebNLI-fine-tuned NeoDictaBERT and XLM-R-large encoders (Strategy 1; deployed silver signals), the Strategy-1 winner XLM-R-XL, AlephBERT, and the Strategy-2 adapters of Qwen3.5-35B-A3B-Base and Gemma-

4-31B. Each model card records the retained checkpoint, HebNLI development and test macro-F1, and the model's HeRE validation F1. The Strategy-4 signals come from prompting public instruction-tuned models with the prompts in the code release; the Strategy-3 fold models are not released.

Few-shot Prompts. The demonstration examples are hand-written and disjoint from both gold sets. Hebrew prompts are literal translations of the English ones below and are included in the released code.

English triplet prompt (5-shot):

> Instruction: Read the passage and determine whether it expresses the given relation. Answer **yes** if the relation is explicitly stated in the passage, and **no** otherwise.
>
> [5 demonstration examples]
>
> Passage: {text}
> Relation: {subject} – {predicate} – {object}
> Answer:

English template prompt (2-shot):

> Instruction: Read the passage and determine whether the following sentence is true according to it. Answer **yes** or **no** only.
>
> [2 demonstration examples]
>
> Passage: {text}
> Sentence: {template_relation}
> Answer:

Threshold Conversion. Our evaluation metrics assume a binary entailed/not-entailed decision per strategy, but most strategies produce soft scores rather than hard labels. We define the conversions as follows. For **Strategy 1**, we sweep fixed thresholds $\{0.6, 0.7, 0.8, 0.9\}$ across all scoring formulas and hypothesis types and report the best hard-label F1. (CV threshold optimization tends toward degenerate low-threshold solutions given the 61.4% positive rate.) For **Strategies 2 and 4**, we sweep soft-score thresholds in $\{0.3, 0.4, \ldots, 0.9\}$ and report at the F1-maximizing threshold. For **Strategy 5**, only hard-label metrics are reported (token probabilities were not used). For **Strategy 3**, the final prediction is majority vote across K fold-models at threshold 0.5. In all cases, Table 2 reports the best F1 over all hypothesis types and prompt configurations per model.

C CROCODILE Data Construction Pipeline

The CROCODILE pipeline (Huguet Cabot and Navigli, 2021) constructs a distant-supervision RE corpus for any language with a sufficiently large Wikipedia. The pipeline, on which we base our Hebrew data construction, proceeds as follows.

Step 1: Wikipedia text extraction. A Wikipedia dump for the target language is processed using a modified version of WikiExtractor that preserves internal hyperlinks. Each article's text is split into section-level passages (or, optionally, the full article body). Entity surface forms and their Wikipedia page identifiers are identified via the hyperlinks present in the raw Wikipedia markup, providing high-precision entity mention detection without requiring a separate NER system.

Step 2: Wikidata triple retrieval. For each Wikipedia entity (subject) identified in Step 1, Wikidata is queried for all relational triples in which it participates as subject. The Wikidata entity identifier is recovered from the Wikipedia page identifier via a Wikipedia-Wikidata mapping. This yields a set of (subject, predicate, object) triples grounded in the Wikidata knowledge graph.

Step 3: Passage-triple alignment. For each triple (subject, predicate, object), passages from Step 1 that mention both the subject and the object entity (identified by their Wikipedia hyperlink identifiers) are retrieved and aligned to the triple. The result is a candidate tuple (text, subject, predicate, object) in which the text is the passage where both entities co-occur. Co-occurrence is treated as a weak positive signal that the passage expresses the target Wikidata relation.

Step 4: Initial NLI denoising (CROCODILE default). The original CROCODILE pipeline applies an XLM-RoBERTa NLI filter as a lightweight denoising step, discarding candidate tuples for which the passage does not entail a natural-language verbalization of the triple. In our work, we replace and substantially extend this denoising step with the comprehensive empirical denoising study described in Section 4.

Hebrew-specific notes. For the Hebrew heb25 slice, the Wikipedia dump is drawn from Hebrew Wikipedia (early 2025). Entity detection relies on Hebrew Wikipedia's internal hyperlink structure, which is generally reliable for named entities (persons, locations, organizations) but sparser for abstract or relational concepts. Date and numeric value entities are handled by a separate regex-based parser. The resulting heb25 corpus contains 3.1

million candidate tuples spanning 1,257 distinct Wikidata predicate strings before entity-field normalisation (1,137 after; Section 5.1).

D HebNLI Fine-tuning

Strategies 1 and 2 of our denoising pipeline require models pre-trained for Hebrew NLI before being applied to RE. We fine-tune all encoder and decoder models on **HebNLI**, a Hebrew NLI dataset constructed by translating and adapting MultiNLI and SNLI into Hebrew. HebNLI contains 300,067 training examples, 1,999 validation examples, and 884 test examples across three Hebrew labels, transliterated here as *hisek* (entailment), *stira* (contradiction), and *neitrali* (neutral). The models fine-tuned here are subsequently used as the NLI backbone for our RE denoising experiments; they are not evaluated on RE directly but provide the entailment/contradiction probability estimates that underpin our scoring functions.

Encoder fine-tuning. Each encoder is adapted as a 3-class sequence classifier using a linear classification head over the `[CLS]` representation. Training uses the AdamW optimizer with a learning rate of 2×10^{-5}, batch size 64, linear warmup over 10% of steps, weight decay 0.01, maximum sequence length 250, and a total of 8,000 training steps. Models are evaluated on the validation set every 200 steps; the checkpoint with the best validation macro F1 is retained.

Decoder fine-tuning. Decoder models are adapted using LoRA (rank 16, targeting all linear projection matrices, bfloat16 precision) with a learning rate of 5×10^{-5}, batch size 16, and up to 8,000 training steps; the checkpoint with the best validation macro-F1 among the periodic evaluations was retained (step 5,280 or 7,920 for every retained model). At inference time, models are prompted with the template *"mishpat 1: {premise} mishpat 2: {hypothesis}"* (Hebrew for "sentence 1" / "sentence 2") and the label is extracted from the first generated token.

Results. Table 6 reports macro F1 on the HebNLI validation and test sets for all fine-tuned models used in the HeRE denoising experiments. All hardware details are provided in Appendix B.

Model	Type	Dev F1	Test F1
Encoder models			
AlephBERT	mono	0.763	0.781
mmBERT	multi	0.819	0.845
mE5-large	multi	0.845	0.850
NeoDictaBERT	mono	0.861	0.878
XLM-RoBERTa-large	multi	0.849	0.876
XLM-RoBERTa-XL	multi	0.853	0.882
Decoder models			
DictaLM-3.0-1.7B-Instruct	mono	0.822	0.845
Gemma3-12B-it	multi	0.851	0.886
Qwen3.5-35B-A3B-Base	multi	0.869	0.904
Gemma-3-27B-it	multi	0.857	0.894
Gemma-4-26B-A4B	multi	0.853	0.890
Gemma-4-31B	multi	0.870	0.905
Gemma-2-9B	multi	0.873	0.899
Aya-Expanse-32b	multi	0.753	0.920
DictaLM-3.0-24B-Base	mono	0.884	0.921
DictaLM-3.0-24B	mono	0.876	0.902

Table 6: HebNLI macro-F1 of all fine-tuned models.

E Error Analysis: Annotated Examples

Three error categories emerge, and the per-predicate breakdown (Appendix G) exposes the mechanism behind the first two. First, *official language*, *language in use* and *opposite of* have no positive gold examples under adjudication, because the passages mention both entities while discussing something else, so their F1 is undefined and reported as 0. The best configurations predict at most two false positives on them, but roughly half of all configurations predict the relation at least once. Second, geographically adjacent predicates, *borders* (positive rate 0.19) and *diplomatic relations* (0.13), show the same low-prior signature: a football tournament or a species listing can mention two neighbouring countries without asserting a border. The best configurations handle these (F1 0.857 and 1.000) but they remain the main false-positive source for mid-tier models and the origin of the micro-macro gap. Third, a small residual set is misclassified by every model; several cases involve annotation ambiguity inherent to distant supervision, where a structurally sound triple happens not to be expressed by the aligned passage.

We present three representative error cases drawn from the gold set, each illustrating a distinct failure mode of the denoising models.

Example 1: False positive from ambiguous co-occurrence (*borders*). Subject: Brazil **Predicate:** borders **Object:** Venezuela **Gold label:** 0 (absent)

Passage (translated from Hebrew):

> The 2000 Copa Libertadores season was the 41st edition of the South American club football competition, organised by CONMEBOL. The first match took place on 15 February 2000 and the competition ended on 21 June 2000…

Analysis. Both Brazil and Venezuela appear in the passage as participant nations in a football tournament. The text says nothing about a shared border, yet the geographic plausibility of the relation (Brazil and Venezuela do share a border in Wikidata) causes most models to predict *yes*. This is a prototypical distant-supervision false positive: entity co-occurrence is a weak positive signal that fires even when the passage is semantically unrelated to the target predicate.

Example 2: Annotation ambiguity from incidental evidence (*official language*). Subject: Serbia **Predicate:** official language **Object:** Serbian **Gold label:** 0 (absent)

Passage (translated from Hebrew):

> The Great War Island (in Serbian: *Veliko ratno ostrvo*) is an island in the Danube River, in the Belgrade area, in Serbia. The island was used in many wars, hence its name. The island is located at the confluence of the Sava with the Danube…

Analysis. The only hint of the relation is the parenthetical "in Serbian:" before the island's local name. The original single-annotator pass labelled this example positive by inferring Serbia's official language from that parenthetical; under the guidelines (Appendix A), a relation that needs external knowledge is negative, and both annotators of the dual-annotation round labelled it 0. All 12 API prompt configurations also answer *no*, while 16 of the 72 open-weights configurations answer *yes*. The case shows how bilingual naming conventions make *official language* look expressed when it is not, which is why this predicate has no positive examples in the adjudicated validation set and why a model that predicts it at all only produces false positives.

Example 3: Universal false positive from predicate granularity (*jurisdiction area*, all models wrong). Subject: Doubs (district) **Predicate:** jurisdiction area **Object:** France **Gold label:** 0 (absent)

Passage (translated from Hebrew):

> Doubs (French: Doubs) is a district in France in the Franche-Comté region, named after the Doubs River. The district was established on 4 March 1790 and is one of the 83 original districts created after the French Revolution…

Analysis. The passage explicitly states that Doubs is "a district in France", a fact that every model reads as confirming the *jurisdiction area* relation. The annotator, however, labelled this 0: the Wikidata property *applies to jurisdiction* (P1001) is a more specific administrative relation than simple geographic containment, and the passage does not assert it in the formal sense required. This example highlights a broader tension in distant supervision: when Wikidata predicates are semantically finer-grained than the natural-language expressions found in Wikipedia, even accurate models trained on surface co-occurrence will systematically overpredict.

F Full Results by Strategy

Tables 7–11 report complete results for every model and configuration evaluated in this work. The best F1 per model is **bolded**; these bolded values are the entries reported in Table 2.

Model	Type	Hypothesis	F1
AlephBERT	mono	basic	0.578
		template	0.605
		llm	**0.618**
mE5-large	multi	basic	0.642
		template	**0.692**
		llm	0.680
mmBERT	multi	basic	0.712
		template	**0.732**
		llm	0.723
XLM-R-large	multi	basic	0.685
		template	**0.756**
		llm	0.752
XLM-R-XL	multi	basic	0.747
		template	**0.779**
		llm	0.740
NeoDictaBERT	mono	basic	0.672
		template	**0.771**
		llm	0.736

Table 7: Strategy 1: F1 by model and hypothesis, best over scoring formula and fixed threshold.

Model	Type	Hypothesis	F1
DictaLM-3.0-1.7B-Instruct	mono	basic	0.365
		template	**0.429**
Gemma-2-9B	multi	basic	0.794
		template	**0.829**
Gemma3-12B-it	multi	basic	0.606
		template	**0.623**
Gemma-3-27B-it	multi	basic	0.639
		template	**0.726**
Gemma-4-26B-A4B	multi	basic	0.440
		template	**0.465**
Gemma-4-31B	multi	basic	0.811
		template	**0.849**
DictaLM-3.0-24B-Base	mono	basic	0.566
		template	**0.803**
DictaLM-3.0-24B	mono	basic	0.789
		template	**0.861**
Aya-Expanse-32b	multi	basic	0.793
		template	**0.829**
Qwen3.5-35B-A3B-Base	multi	basic	0.825
		template	**0.873**

Table 8: Strategy 2: best soft-score F1 by model and hypothesis (constrained-prompt reruns for Gemma-4-31B, Qwen3.5-35B, Aya-Expanse-32b and DictaLM-3.0-24B).

Model	Type	K	F1	Macro	Silver pos.%
XLM-R-large	multi	3	**0.729**	0.613	53.5%
		5	0.713	0.622	56.8%
		7	0.718	0.622	57.4%
		10	0.714	0.621	57.9%
mmBERT	multi	3	0.732	0.614	52.1%
		5	0.733	0.619	54.9%
		7	0.737	0.616	56.5%
		10	**0.743**	0.631	59.5%
NeoDictaBERT	mono	3	**0.751**	0.659	64.3%
		5	0.750	0.649	65.7%
		7	0.749	0.647	66.0%
		10	0.742	0.645	67.5%
mE5-large	multi	3	**0.675**	0.568	47.8%
		5	0.674	0.576	48.4%
		7	**0.675**	0.574	49.6%
		10	0.669	0.578	50.3%

Table 9: Strategy 3: ensemble F1 and macro-F1 by model and K; Silver pos.% is the ensemble's positive rate on the silver corpus.

Model	Lang	Shots	F1
GPT-5.4	He	0	0.893
		5	**0.921**
	En	0	0.887
		5	0.908
Gemini 3 Flash	He	0	0.903
		5	0.916
	En	0	0.894
		5	**0.917**
Claude Haiku 4.5	He	0	0.863
		5	**0.890**
	En	0	0.850
		5	0.889

Table 11: Strategy 5: F1 by model, prompt language and shots (triplet hypothesis). **Bold**: best per model.

G Per-predicate, Length and Silver-Signal Analyses

	Validation		Test	
Predicate	n	% pos.	n	% pos.
Country	79	81.0	16	75.0
Diplomatic relations	15	13.3	29	6.9
Instance of	29	93.1	9	88.9
Official language	19	0.0	17	5.9
Occupation	17	88.2	17	94.1
Borders	21	19.0	11	0.0
Administrative unit	25	84.0	1	100.0
Place of birth	7	85.7	19	73.7
Place of origin	8	12.5	16	0.0
Citizenship	16	56.2	7	85.7
Language in use	17	0.0	5	0.0
Capital of	8	25.0	14	14.3
Admin. subdivision	18	83.3	3	33.3
Part of	11	81.8	10	70.0
Siblings	9	66.7	12	66.7
Subclass of	18	61.1	3	33.3
Performer	11	72.7	9	100.0
Opposite of	9	0.0	11	0.0
Predecessor	2	0.0	7	42.9

Table 12: Positive rate of the most frequent predicates in the gold validation and test sets.

Model	Shots	Hebrew F1		English F1	
		triplet	template	triplet	template
DictaLM-3.0-24B-Base	0	0.643	0.658	0.635	0.802
	2	0.822	0.818	0.723	0.811
	5	**0.835**	0.822	0.595	0.821
Aya-Expanse-32b	0	0.820	0.821	0.839	0.833
	2	0.823	**0.847**	0.824	0.835
	5	0.830	0.832	0.828	0.836
Gemma-3-27B-it	0	0.860	0.875	0.844	0.860
	2	0.861	0.867	0.871	0.858
	5	0.858	**0.877**	0.868	0.862
Gemma-4-26B-A4B-it	0	0.788	0.861	0.850	0.860
	2	0.789	0.869	0.877	**0.880**
	5	0.739	0.871	0.870	0.876
Gemma-4-31B-it	0	0.868	0.915	0.851	0.905
	2	0.905	**0.915**	0.913	0.910
	5	0.899	0.914	0.907	0.913

Table 10: Strategy 4: F1 by model, hypothesis, shots and prompt language. **Bold**: best per model.

Predicate	n	F1
Country	79	0.947
Instance of	29	0.962
Administrative unit	25	0.955
Borders	21	0.857
Official language	19	n/a*
Admin. subdivision	18	0.903
Subclass of	18	0.900
Language in use	17	n/a*
Occupation	17	0.966
Citizenship	16	0.800
Diplomatic relations	15	1.000
Part of	11	0.889
Performer	11	0.947

Table 13: Per-predicate F1 of the best configuration (GPT-5.4, Hebrew, triplet, 5-shot) on the validation set, $n \geq 10$. *No positive gold examples, so F1 is undefined; the model predicts 0 and 2 false positives, respectively.

Passage length (words)	n	F1	95% CI
100–250	96	0.909	[0.840, 0.966]
250–500	256	0.835	[0.784, 0.880]
500+	133	0.818	[0.740, 0.882]

Table 14: Gemma-4-31B-it test-set F1 by passage length (15 passages under 100 words omitted).

Per-signal positive rates and consensus rules. The two Strategy-3 signals (NeoDictaBERT $K{=}3$, mmBERT $K{=}5$) predate the grid re-analysis (Table 9), which later showed K to be immaterial; they abstain on the 0.4% of rows whose predicate has fewer than 20 silver examples. Aya-Expanse-32b and Gemma-2-9B were deployed at score threshold 0.9.

Model	Strategy	Pos.	Neg.
Gemma-4-31B-it	S4: Open-source LLM	54.0%	46.0%
Gemma-3-27B-it	S4: Open-source LLM	70.4%	29.6%
DictaLM-3.0-24B-Base	S4: Open-source LLM	66.9%	33.1%
DictaLM-3.0-24B	S2: Fine-tuned NLI	67.8%	32.2%
Gemma-2-9B	S2: Fine-tuned NLI	23.2%	76.8%
Aya-Expanse-32b	S2: Fine-tuned NLI	18.7%	81.3%
NeoDictaBERT	S1: Encoder NLI	55.7%	44.3%
XLM-R-large	S1: Encoder NLI	35.4%	64.6%
NeoDictaBERT ($K{=}3$)	S3: Cross-trained	63.1%	36.6%
mmBERT ($K{=}5$)	S3: Cross-trained	54.2%	45.4%
`agree3` (published)	3-model unanimous	72.5%	27.5%

Table 15: Positive rates of the ten deployed signals on the silver corpus; `agree3` rates are within the published set.

Ensemble	F1	95% CI
`agree3` (published label)‡	0.916	[0.886, 0.944]
`vote7` (majority of 7)	0.801	[0.763, 0.837]
`agree2`	0.770	[0.728, 0.809]
`vote_all`	0.680	[0.628, 0.729]
`agree4`	0.590	[0.531, 0.646]

Table 16: Held-out test-set F1 of consensus rules over the silver-corpus signals: `agree3` is the unanimous vote of Gemma-4-31B-it, Gemma-3-27B-it and DictaLM-3.0-24B (S2); `agree2` the unanimous vote of Gemma-3-27B-it and DictaLM-3.0-24B (S2); `agree4` adds NeoDictaBERT (S1) and NeoDictaBERT $K{=}3$ (S3) to `agree2`; `vote7` the majority of the three S4 models, DictaLM-3.0-24B (S2), NeoDictaBERT (S1) and both S3 signals; `vote_all` the majority of all ten. Unanimity rules abstain on split votes (abstentions counted as negatives). ‡; precision 0.899, recall 0.934.