Combining Distantly Supervised Models with In Context Learning for Monolingual and Cross-Lingual Relation Extraction

Vipul Kumar Rathore Malik Hammad Faisal Parag Singla Mausam
Indian Institute of Technology
New Delhi, India
{rathorevipul28, faisalmalikhammad}@gmail.com
{parags, mausam}@cse.iitd.ac.in

Abstract

Distantly Supervised Relation Extraction (DSRE) remains a long-standing challenge in NLP, where models must learn from noisy bag-level annotations while making sentence-level predictions. While existing state-of-the-art (SoTA) DSRE models rely on task-specific training, their integration with in-context learning (ICL) using large language models (LLMs) remains underexplored. A key challenge is that the LLM may not learn relation semantics correctly, due to noisy annotation.

In response, we propose HYDRE – **HY**brid **D**istantly Supervised **R**elation **E**xtraction framework. It first uses a trained DSRE model to identify the top-k candidate relations for a given test sentence, then uses a novel dynamic exemplar retrieval strategy that extracts reliable, sentence-level exemplars from training data, which are then provided in LLM prompt for outputting the final relation(s). We further extend HYDRE to cross-lingual settings for RE in low-resource languages. Using available English DSRE training data, we evaluate all methods on English as well as a newly curated benchmark covering four diverse low-resource Indic languages – Oriya, Santali, Manipuri, and Tulu. HYDRE achieves up to 20 F1 point gains in English and, on average, 17 F1 points on Indic languages over prior SoTA DSRE models and naive prompting baselines. Detailed ablations exhibit HYDRE's efficacy compared to other prompting strategies.

1 Introduction

Relation Extraction (RE) is a core task in Information Extraction (IE) that aims to identify semantic relations between entity mentions in text. Given a sentence s containing a head entity e_1 and a tail entity e_2, the goal is to predict the set of relations $(\hat{r} \subset \mathcal{R})$ expressed between them, where $\mathcal{R}$ is a predefined ontology of relations. RE plays a crucial role in downstream applications such as knowledge base construction, and question answering. Supervised RE methods depend on sentence-level annotations, which are costly and time-consuming to obtain at scale (Zhang et al., 2017).

Distantly Supervised Relation Extraction (DSRE) (Mintz et al., 2009) alleviates this challenge by aligning text with knowledge base (KB) triples to generate weakly labeled training data. Specifically, DSRE groups all sentences mentioning an entity pair (e_1, e_2) into a bag $B(e_1, e_2)$, which is labeled with all relations $R(e_1, e_2)$ known between (e_1, e_2) in the KB. Although the supervision is weak and bag-level, inference is typically performed at the sentence level (micro-reading (Mitchell et al., 2009)). This mismatch between training and inference granularity – coupled with noisy training labels – often causes state-of-the-art (SoTA) DSRE models to confuse fine-grained relation types, such as *Nationality* vs. *Place_of_Birth* or *Founder* vs. *CEO*, thereby limiting their overall performance.

Existing DSRE methods (Chen et al., 2021; Rathore et al., 2022; Jian et al., 2024) primarily rely on task-specific fine-tuning of moderate-sized language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). However, the potential of Large Language Models (LLMs) for this task remains largely unexplored. Modern LLMs excel at in-context learning (ICL), where the model performs reasoning by conditioning on a few task-specific exemplars. Yet, directly applying LLMs to DSRE is non-trivial: noisy distant supervision degrades exemplar quality, and the lack of clean, sentence-level exemplars undermines effective ICL. Consequently, prior works in DSRE either ignore LLMs altogether or fail to exploit their reasoning capabilities effectively (Zhao et al., 2024).

In this work, we propose HYDRE: a **HY**brid **DSRE** framework that combines the high-recall candidate label selection of fine-tuned DSRE models with the reasoning capabilities of LLMs. Given a query sentence, a fine-tuned DSRE model first

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45483–45509

identifies a candidate relation set by filtering out the obvious negatives. These candidates are then passed to an LLM for disambiguation, guided by carefully selected in-context exemplars. To construct these exemplars, we retrieve relevant bags from the DSRE training corpus using a joint scoring function that blends model confidence with semantic similarity to the query. From each selected bag, we extract the most representative sentence to form a dynamic, relation-specific prompt, guiding the LLM to accurately select the correct relation(s).

We further extend HYDRE to the cross-lingual setting, focusing on low-resource languages — a largely underexplored area in DSRE. To facilitate this, we construct a new multilingual benchmark covering four low-resource Indic languages: Oriya, Santali, Manipuri, and Tulu. Given the limited representation of these languages in LLM pretraining corpora (Singh et al., 2024; Nag et al., 2025), they pose an interesting challenge for evaluating cross-lingual RE in the context of latest LLMs.

We evaluate HYDRE under three transfer settings - (1) *English-only*, where no target language data is used; (2) *Translate-train*, where English DSRE training data is translated to target language; and (3) *Translate-test*, where test queries in the target language are translated to English.

Experiments with both open-source and proprietary LLMs show that HYDRE consistently outperforms prior DSRE baselines and naive LLM prompting strategies. Our exemplar retrieval strategy proves robust across both monolingual and cross-lingual setups. Ablation analyses further reveal that removing retrieval components can degrade performance by up to 7 F1 points in English and 12 points in cross-lingual transfer.

In summary, our key contributions are as follows. (1) We present the first systematic integration of LLMs via In-Context Learning (ICL) into DSRE inference, achieving significant gains over both fine-tuned and prompting-only baselines. (2) We propose a novel retrieval strategy that combines model confidence and semantic similarity to select high-quality ICL exemplars for LLM-based relation disambiguation. (3) We curate and release gold-standard evaluation datasets for relation extraction for four typologically diverse low-resource Indic languages. (4) We propose effective cross-lingual strategies for adapting HYDRE to low-resource languages, demonstrating robustness across multiple transfer scenarios.

We release our code[1] to facilitate further research in multilingual DSRE.

2 Related Work

Distantly Supervised Relation Extraction: DSRE (Mintz et al., 2009) aligns KB triples with text to create bag-level supervision, where labels apply at the bag rather than sentence level. Neural DSRE models typically adopt the multi-instance multi-label (MIML) framework (Surdeanu et al., 2012). Earlier works encoded sentences in a bag using piecewise CNNs (Zeng et al., 2015) or graph CNNs (Vashishth et al., 2018), while recent models employ pre-trained transformers. PARE (Rathore et al., 2022) encodes a bag by treating all bag sentences as a passage, whereas CIL (Chen et al., 2021) uses intra-bag attention and contrastive learning. HiCLRE (Li et al., 2022) introduces hierarchical contrastive learning, and HFMRE (Li et al., 2023b) employs Huffman-tree structures to denoise bags. These advances improve bag-level reasoning but still struggle with fine-grained sentence-level inference.

Multilingual DSRE: Research in multilingual RE has progressed (Ni et al., 2020; Nag et al., 2021), but multilingual DSRE is scarce. DisRex (Bhartiya et al., 2022) introduced a dataset across four European languages, though without manual sentence-level evaluation. No prior work targets typologically diverse low-resource Indic languages, which motivates our benchmark contribution.

LLMs for Relation Extraction: Large language models (LLMs) have recently been explored for RE primarily in supervised settings. Wadhwa et al. (2023) demonstrated that few-shot prompting with GPT-3 can rival fully supervised baselines, particularly when enhanced with chain-of-thought (CoT) reasoning. Other works employ zero-shot prompting with relation label definitions (Zhou et al., 2024) or leverage LLMs to denoise distantly supervised training data (Li et al., 2023a; Jian et al., 2024). However, these approaches either assume access to clean, labeled supervision or utilize LLMs solely for data relabeling purposes. In contrast, HYDRE takes a complementary view by employing LLMs directly at test time, leveraging them as reasoning engines rather than annotation tools.

Diversity-aware exemplar selection: The standard in-context exemplar selection strategy is to

[1] https://github.com/dair-iitd/HYDRE

retrieve the top-K semantically most similar exemplars to the query, but this often results in redundancy amongst the selected exemplars (Gupta et al., 2023). To address this, Kapuriya et al. (2025) propose maximal marginal relevance (MMR), which balances similarity with diversity, while Wang et al. (2024) employ determinantal point processes (DPPs) to encourage diversity through submodular modeling objectives. These methods highlight the importance of complementing similarity with diversity in exemplar selection, a principle that we extend within our hybrid DSRE-LLM framework.

LLM-as-Judge and Hybrid Models: Recent work explores LLMs as evaluators ("judges") for ranking candidate outputs (Zheng et al., 2023; Bavaresco et al., 2024). Our approach could be viewed as an instance of this paradigm, where a DSRE model provides high-recall candidate relations and their exemplars, and the LLM acts as a judge to select the correct candidate. Related hybrid paradigms exist for other NLP tasks (Rathore et al., 2024), such as sequence labeling, but not for DSRE.

3 HYDRE: HYbrid Distantly supervised Relation Extraction

DSRE models are trained on bags $B(e_1, e_2)$, each comprising multiple sentences mentioning an entity pair (e_1, e_2) and annotated with a set of relations $R(e_1, e_2)$. As a result, such models often generalize poorly for sentence-level queries (Gao et al., 2021; Chen et al., 2021). Nevertheless, our preliminary analysis (Appendix A.7) shows that strong SOTA DSRE models such as *PARE* achieve high sentence-level Recall@k – approximately 85% for English and 70% for Indic languages – even for small values of k (E.g. k=5). This makes models such as *PARE* well-suited as a high-recall candidate generator, with an LLM acting as a precision-oriented selector over the filtered candidate set. The core challenge lies in constructing reliable in-context exemplars from noisy bags. To solve this, we propose a three-stage exemplar selection pipeline (Fig. 1) that extracts clean, representative sentences from distantly supervised data.

Stage 1: Candidate Relation Selection. A trained DSRE model (e.g., *PARE*) assigns confidence scores $f_{\text{PARE}}(q, r)$ to each relation r for a query sentence q. The top-k relations with the highest scores form the candidate relation set $\mathcal{R}'$.

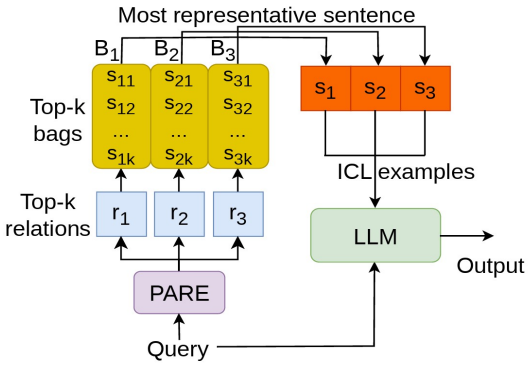


Figure 1: **HYDRE pipeline** ($K = 3$). Given a query, *PARE* produces top-K candidate relations (r_1, r_2, r_3). For each candidate, the most relevant bag (B_1, B_2, B_3) is retrieved from the DSRE corpus. The most representative sentence from each bag is selected as an ICL exemplar and passed to the LLM alongside the query.

Stage 2: Bag Selection. For each candidate relation $r \in \mathcal{R}'$, we identify the most relevant bag B_r from the subset $\mathcal{B}_r$ of bags annotated with relation r as one of it's relations. Each bag $B_j \in \mathcal{B}_r$ is scored using a weighted combination of semantic similarity and DSRE confidence:

$$\text{score}(B_j, r) = (1-\lambda)\,\text{sim}(q, B_j) + \lambda\, f_{\text{PARE}}(B_j, r)$$

Here, $\text{sim}(q, B_j) = \max_{s \in B_j} \text{sim}(q, s)$ measures semantic similarity of bag B_j to the query, and $f_{\text{PARE}}(B_j, r) = \max_{s \in B_j} f_{\text{PARE}}(s, r)$ represents B_j's confidence score for relation r. This selection favours bags whose sentences are strongly indicative of the candidate relation, while also being semantically aligned with the query. The trade-off parameter λ controls the relative importance of semantic similarity versus label confidence and is tuned on a development set separately for each target language.

Stage 3: Sentence Selection. From each selected bag B_r, we extract a single sentence s_r that best captures the relation(s) expressed in the bag. For every sentence $s \in B_r$, we compute a coverage score $c(s)$, defined as the number of bag relations whose confidence exceeds a threshold τ:

$$c(s) = \sum_{r_a \in \text{labelset}(B_r)} \mathbb{I}[f_{\text{PARE}}(s, r_a) > \tau]$$

Among sentences with maximum coverage, we select s_r as the one with the highest aggregate confidence $\sum_{r_a} f_{\text{PARE}}(s, r_a)$. This encourages selection of sentences that express the largest number of valid bag relations with strong model con-

fidence. In the prompt we place the selected sentences s_r along with their corresponding bag labels labelset(B_r).

Our stage 3 offers three key benefits. First, it aids the *multi-label* nature of the problem, since co-occuring relations (which may be present in B_r, but not in $\mathcal{R}'$) are also added to the prompt. This may not have been possible if the method tried to, instead, search for sentences that express only the specific relation. Second, selecting a sentence that has high label confidence over many relations promotes diversity (coverage) of relevant labels in the prompt – this may not occur if only semantic similarity with test sentence were used for selection. Finally, scoring sentences using aggregate confidence over all bag labels helps surface sentences with stronger overall evidence, not just noisy single-label confidence. We justify Stage 3 design via systematic ablations in §A.5.

Once selected, exemplars are ordered by ascending $f_{\text{PARE}}(q, r)$, placing more relevant candidates closer to the query in the prompt. The full selection process is described in Algo. 1 (Appendix).

Prompting and Parsing. A prompt comprises (i) task instruction, (ii) candidate relations along with their definitions, (iii) ICL exemplars, and (iv) the query (see Appendix A.2). The output is parsed using partial string matching heuristics to obtain predicted labels; if no valid match is found, the query is labeled as "NA".[2]

3.1 Dataset Curation

To evaluate HYDRE in low-resource settings, we construct gold-standard test sets for four Indic languages: Oriya (Odia), Santali, Manipuri, and Tulu. We deliberately selected these languages to ensure broad diversity across linguistic families, scripts, and resource availability – key factors for rigorously evaluating cross-lingual generalization. Notably, they span four distinct language families and distinct scripts: Indo-Aryan (Oriya written in Oriya script), Austro-Asiatic (Santali in Ol Chiki), Tibeto-Burman (Manipuri in Meitei), and Dravidian (Tulu in Kannada), allowing us to test cross-lingual transfer across typologically diverse settings.

Although Oriya, Santali, and Manipuri are officially Scheduled languages of India with massive speaker bases (ranging from ∼2M to ∼40M), they remain severely underrepresented in NLP, possessing limited web presence (only 3K–20K Wikipedia articles).

We begin with a stratified subset of the English NYT-10m test split (Gao et al., 2021), having 538 multi-label queries with a total of 722 labels (incl. 30 "NA"s), ensuring balanced relation coverage. Sentences are translated into each target language using CODEC (Le et al., 2024) with IndicTrans2 (Gala et al., 2023), which jointly performs translation and entity projection to preserve head and tail entity spans. For Tulu, which lacks IndicTrans2 support, we use Google Translate API.

All translations are manually verified by native speakers for both sentence quality and entity alignment. Across languages, over 70% of translations required no correction, and inter-annotator agreement exceeded 90%, indicating high annotation quality. Detailed dataset statistics, annotation procedures, and quality assessment are provided in Appendix A.11.

4 Experiments

We conduct experiments in both monolingual (English) and cross-lingual settings across four low-resource Indic languages. For cross-lingual evaluation, we follow three standard settings commonly used in cross-lingual NLP. Let X_{train} denote the DSRE training corpus translated from English into a target language X. We denote by *PARE-X* and *CIL-X* the target-language counterparts of *PARE* and *CIL*, respectively, fine-tuned upon X_{train}.

1. **English-only:** Models are trained or prompted exclusively using English DSRE data. For English test queries, HYDRE uses the PARE confidence scores for $f_{\text{PARE}}(q, r)$, and the off-the-shelf encoder *e5-large-v2* (Wang et al., 2022) for semantic similarity. For Indic test queries, since PARE is trained only on English, it cannot produce reliable confidence scores $f_{\text{PARE}}(\cdot, \cdot)$ for Indic scripts. Consequently, f_{PARE} is omitted completely, and HYDRE relies solely on the multilingual sentence encoder *BGE-M3* (Chen et al., 2024) to retrieve semantically similar English exemplars for cross-lingual ICL.
2. **Translate-train:** HYDRE employs *PARE-X* for confidence estimation – given its higher Recall@5 compared to *CIL-X* (Table 13) – and *CIL-X* encoder for semantic similarity.

[2] In practice, we observed that Llama-based models occasionally generate truncated or slightly modified relation names (e.g., producing `Birthplace` instead of `place_of_birth`); other models typically adhered to the exact string forms.

We choose *CIL-X* for the latter because its contrastive training objective yields task-specific representations, enabling more effective retrieval in the target language than off-the-shelf sentence embeddings.

3. **Translate-test:** Test queries in the target language X are translated into English (X_{test}), after which the standard English HYDRE pipeline is applied without modification.

Setting	Similarity Model	Confidence Model
English-only (En)	e5-large-v2	PARE
English-only (Indic)	BGE-M3	–
Translate-train	CIL-X	PARE-X
Translate-test	e5-large-v2	PARE

Table 1: Models used for semantic similarity and confidence estimation across various evaluation settings.

Table 1 summarizes the models used for semantic similarity and confidence estimation under each evaluation setting.

Baselines. We compare HYDRE against the following categories of baselines:

- **Supervised DSRE models:** We include *PARE* and *CIL* for English, along with their target-language counterparts *PARE-X* and *CIL-X* under the *Translate-train* setting.
- **0-shot prompting LLM:** We evaluate both open-source and proprietary LLMs: *Qwen3-235B-A22B*, *GPT-4o*, *Llama3.1-8B*, and its fine-tuned variant *Llama3.1-8B-FT*. The fine-tuned model (*Llama3.1-8B-FT*) is trained on English DSRE data for English experiments, and on X_{train} under the *Translate-train* setting. In all cases, inference is performed without in-context exemplars. We consider two prompting variants:
  - **Direct:** The LLM is provided only with the query sentence and the list of candidate relation labels.
  - **Definition-based:** The LLM additionally receives natural language definitions of candidate relations.
- **Few-shot prompting with exemplar retrieval:** We compare against few-shot baselines that differ in how exemplars are selected:
  - *Random-K:* Randomly samples K exemplars for each query.
  - *TopK-sim:* Selects the top-K most semantically similar exemplars to the query.
  - *LM-MMR:* Employs a diversity-aware Maximal Marginal Relevance (MMR) objective to balance relevance and diversity (implementation details in Appendix A.1.2).

Evaluation Datasets. Our primary results are reported on NYT-10m (Gao et al., 2021), including its translated Indic variants used for cross-lingual evaluation. We additionally evaluate on Wiki-20m to assess generalization, with results and analysis deferred to Appendix 5.5 due to space constraints.

Implementation details. For obtaining X_{train} and X_{test} in our experiments, we use EasyProject (Chen et al., 2023), a more lightweight joint translation and entity projection tool as compared to CODEC. We use LoRA fine-tuning for LLaMA-3.1, updating only adapter and embedding weights while freezing other parameters (App. A.1.3). For PARE-X training, we continually pretrain mBERT on X_{train} before adapter-based fine-tuning.

Hyperparameters For HYDRE, we set the number of exemplars to $k = 5$, following the sensitivity analysis in §5.3. The confidence threshold is fixed to $\tau = 0.5$, consistent with the original PARE implementation. We tune the trade-off parameter λ using Llama 3.1 via grid search over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the dev set. The optimal value is $\lambda = 0.5$ for English, and λ = 0.1 for the Indic translate-train setting.

All LLMs are run with temperature = 0.0, maximum seq. length = 2048 tokens, and maximum generation length = 256 tokens.

Evaluation Metrics. We report both micro-F1 and macro-F1 scores. Area-under-the-curve (AUC) is omitted, as LLM-based methods do not produce calibrated confidence scores needed for threshold-based evaluation. Statistical significance is assessed using McNemar's test (McNemar, 1947), which is applicable for micro-F1 comparisons.

5 Results & Analysis

Table 2 reports results on English and four Indic languages across the three evaluation settings.

English results. Among supervised DSRE models, *CIL* (43 micro-F1) and *PARE* (42 micro-F1) perform best. Zero-shot GPT-4o already surpasses

Model	English	Indic Languages		
		English-only	Translate-train	Translate-test
Supervised				
HFMRE (Li et al., 2023b)	33/18	–	22/13	27/16
HiCLRE (Li et al., 2022)	31/18	–	20/13	25/14
CIL (Chen et al., 2021)	43/32	–	26/18	34/24
PARE (Rathore et al., 2022)	42/31	–	30/20	33/23
0-shot (direct)				
Qwen3-235B-A22B	47/39	25/21	25/21	40/34
Llama3.1-8b	24/21	16/12	16/12	22/20
Llama3.1-8B-FT	55/37	26/17	26/17	47/31
GPT-4o	56/55	31/29	31/29	51/49
0-shot (Definition-based)				
Qwen3-235B-A22B	49/40	25/21	25/21	44/35
Llama3.1-8b	31/17	22/16	22/16	29/25
Llama3.1-8B-FT	60/44	24/14	38/23	49/34
GPT-4o	56/57	33/31	32/30	54/51
few-shot (Random)				
Random-K(Qwen3-235B-A22B)	55/55	21/19	29/25	43/35
Random-K(Llama3.1-8B)	30/24	19/11	11/7	27/21
Random-K(Llama3.1-8B-FT)	55/40	24/14	25/12	46/32
Random-K(GPT-4o)	56/52	31/30	38/31	48/42
few-shot (Similarity-based)				
TopK-sim(Qwen3-235B-A22B)	52/49	22/18	32/26	45/38
TopK-sim(Llama3.1-8B)	33/27	22/14	19/9	30/22
TopK-sim(Llama3.1-8B-FT)	55/40	27/15	29/14	46/30
TopK-sim(GPT-4o)	58/53	29/26	41/32	48/41
few-shot (Diversity-based)				
LM-MMR(Qwen3-235B-A22B)	50/47	24/20	33/28	44/36
LM-MMR(Llama3.1-8B)	34/26	21/12	17/8	30/22
LM-MMR(Llama3.1-8B-FT)	56/40	26/16	28/15	47/32
LM-MMR(GPT-4o)	56/52	28/25	41/33	50/43
*few-shot (*HYDRE*)*				
HYDRE(QWEN3-235B-A22B)	**63*/62**	29/26	37/31	54/46
HYDRE(LLAMA3.1-8B)	52/47	30/19	31/20	44/36
HYDRE(LLAMA3.1-8B-FT)	61/45	35/21	**45***/28	51/37
HYDRE(GPT-4O)	**63***/60	**36*/33**	39/**35**	**56*/54**

Table 2: Results for English and Indic languages under three evaluation settings. In each entry, we report micro and macro F1 scores. The Indic results are averaged over 4 languages (language-wise results shown in Appendix A.4). * McNemar's p-value $< 10^{-5}$ (valid for micro-F1 comparison).

them (56 F1), and HYDRE's few-shot prompting further improves GPT-4o to 63 F1. Smaller open LLMs (Llama 3.1, Qwen 3) benefit even more – gaining up to +14 F1 – demonstrating that HYDRE's exemplar-guided prompting is especially effective for weaker LLMs.

Cross-lingual transfer (English-only setting). Since English-trained DSRE models (*PARE*, *CIL*) lack coverage for Indic scripts, their results are omitted. Zero-shot prompting – even with GPT-4o – yields low scores (22-33 F1), reflecting limited Indic-language coverage in LLM pretraining. Nevertheless, cross-lingual prompting with English exemplars via HYDRE still yields consistent gains (+3 to +11 F1). For example, HYDRE(*Llama-3.1-8B-FT*) improves from 24/14 to 35/21, showing that carefully selected English exemplars can meaningfully support reasoning in low-resource languages when aligned through cross-lingual retrieval.

Translate-train setting. Supervised translate-train baselines such as *PARE-X* (30/20 F1) outperform zero-shot LLM prompting (except for GPT-4o), yet HYDRE exceeds the supervised baselines by +8 avg. F1 points, showing that careful exemplar-guided prompting can outperform fine-tuned DSRE models. For fine-tuned Llama (*Llama-3.1-8B-FT*), the gains are more pronounced - exceeding *PARE-X* by +15 micro-F1 and over zero-shot *Llama-3.1-8B-FT* by +7 points. Fine-tuned Llama achieves the best overall performance (45/28 F1), with GPT-4o close behind (39/35 F1).

Analysis shows script-dependent trends: GPT-4o

excels on medium-resource scripts (Oriya, Tulu), while fine-tuned Llama dominates rarer scripts (Santali, Manipuri). This is consistent with *Llama-3.1-8B-FT* learning script-specific cues that proprietary models lack.

Overall, this setting suggests a favorable cost-performance tradeoff: locally fine-tuned open-source LLMs can outperform GPT-4o on very low-resource languages, making HYDRE potentially useful for cost-sensitive deployment scenarios.

Translate-test setting. Translating test queries into English introduces a mild degradation relative to native English evaluation, primarily due to translation noise. Nevertheless, HYDRE continues to deliver strong improvements across models: Llama3.1 and Qwen3 gain +15 and +10 micro-F1 points over their zero-shot counterparts, and GPT-4o reaches 56/54 F1—the best overall performance in this setting.

Overall, HYDRE demonstrates strong robustness across LLMs, scripts, and transfer settings, preserving high performance even when translation introduces mild noise.

HYDRE vs. other few-shot prompting methods. HYDRE consistently surpasses all few-shot baselines. Compared to the diversity-based LM-MMR, HYDRE(GPT-4o) achieves average gains of +7 micro-F1 (English) and +4 avg. micro-F1 on Indic languages. These improvements stem from HYDRE's hybrid scoring mechanism, which integrates DSRE confidence with semantic similarity. In contrast, methods such as MMR or TopK-sim rely solely on embedding-based similarity or diversity heuristics and do not leverage label-aware signals, limiting their ability to select informative and discriminative exemplars.

5.1 Ablations

To quantify the contribution of each component in HYDRE, we conduct systematic ablations of its three-stage design. We evaluate the following variants:

- *w/o candidate selection (Stage 1):* considering all relations instead of top-k.
- *w/o bag selection (Stage 2):* retrieving sentences directly without bag-level filtering.
- *w/o sentence selection (Stage 3):* providing full bags to the LLM without sentence filtering.
- *w/o semantic similarity:* using only *PARE* confidence for retrieval.
- *w/o* PARE *confidence:* using only semantic similarity for retrieval.
- *w/o ICL:* predicting from top-k candidate relations directly without their exemplars.

Table 3 summarizes ablation results for English and cross-lingual settings. Across most models and settings, HYDRE outperforms its ablations, confirming that its components contribute complementary gains.

Semantic similarity vs. DSRE confidence. For English, semantic similarity plays a dominant role: removing it degrades Llama 3.1 by up to 7 F1, while ablating *PARE* confidence leads to minor drops (2-3 F1) across LLMs.

In contrast, under the Indic translate-train setting, DSRE confidence becomes more critical – removing f_{PARE} consistently degrades performance, particularly for Llama models, while removing semantic similarity causes little or no loss. This underscores the limited effectiveness of embedding-based retrieval in extremely low-resource regimes.

Candidate relation selection. Providing exemplars for all 25 relations yields performance comparable to HYDRE, but at nearly $5\times$ higher prompt cost. While Qwen3 and GPT-4o remain relatively stable, LLaMA degrades noticeably, indicating that excessive distractor relations disproportionately affect weaker models. This confirms that candidate relation selection is essential not only for efficiency but also for stable and reliable reasoning.

Importance of bag selection. Omitting Stage 2 (Bag Selection) and retrieving sentences directly leads to notable performance drops across both English (up to 7 points) and Indic translate-train (up to 3 points) settings. By grounding sentence selection within bags relevant to candidate relations, HYDRE filters noisy supervision and preserves relation-consistent evidence for downstream reasoning.

Importance of sentence selection. Providing full bags instead of selecting representative sentences degrades performance by upto 6 micro F1 and 10 macro F1, as LLMs struggle with longer and noisier contexts. This highlights HYDRE's ability to exploit a key property of distant supervision:

Ablation	English			Indic (Translate-train)			
	L3.1	Qw3	G4o	L3.1	L3.1-FT	Qw3	G4o
HYDRE	**52/47**	**63/62**	**63**/60	31/**20**	**45/28**	37/**31**	**39/35**
w/o candidate selection (Stage 1)	39/31	**63**/61	62/**64**	–	–	–	–
w/o bag selection (Stage 2)	48/41	62/59	56/51	**33/20**	43/25	34/30	38/**35**
w/o sentence selection (Stage 3)	46/37	60/59	59/58	–	–	–	–
w/o f_{PARE} ($\lambda = 0.0$)	49/45	61/60	61/59	26/18	44/27	37/**31**	**39**/34
w/o sem. sim. ($\lambda = 1.0$)	45/39	60/59	60/62	32/**20**	**45/28**	**38/31**	38/34
w/o ICL	45/37	58/50	59/49	29/19	33/18	33/20	32/21

Table 3: Ablation results for HYDRE on English and Indic (translate-train) settings. We report micro/macro F1 scores. "–" indicates ablations that are not applicable in a given setting. For Indic languages (translate-train), Stage 1 and Stage 3 ablations are omitted due to excessive prompt length caused by high token fertility.

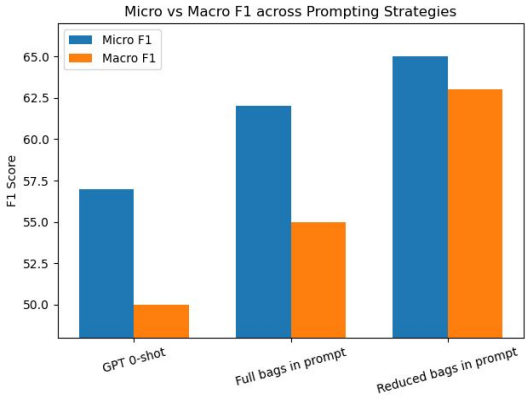


Figure 2: Zero-shot prompting vs full bag ICL vs reduced bag ICL for GPT-4o on English NYT-10m

only a small subset of sentences is required for precise relation inference.

Importance of in-context exemplars. Removing in-context exemplars reduces performance by up to 7 F1 on English and 12 F1 in the translate-train setting, confirming that HYDRE's gains stem from exemplar-guided reasoning rather than exposure to candidate labels alone.

5.2 Robustness to bag-level evaluation

Beyond sentence-level inference, we evaluate HYDRE under *bag-level* settings, as commonly studied in DSRE (Gao et al., 2021), where evidence for a relation is distributed across multiple sentences (e.g., news articles for KB population).

For bag-level inference, we follow Algorithm 1 up to bag selection, but instead of selecting a single sentence per bag, we retain one representative sentence *per relation*, chosen using the highest *PARE* confidence. This forms a *reduced-bag* representation that preserves relation-specific evidence while removing redundant context. As shown in Fig. 2, reduced-bag prompting improves performance by ≈ 3 micro-F1 and 7 macro-F1 over full-bag prompting, while reducing prompt length by nearly $2\times$ (2451 $\rightarrow$ 1242 tokens). These results show that effective DSRE requires only a small, carefully selected subset of sentences, yielding both better accuracy and lower inference cost.

5.3 HYDRE Sensitivity to k

We study the effect of the number of candidate relations (k) selected in Stage 1. As shown in Figure 15b, micro-F1 increases with k, peaks around $k = 5$-10, and then plateaus or slightly declines. Meanwhile, the Stage 1 DSRE model (PARE) already achieves high Recall@k at small values ($\approx$ 85% at $k = 5$) and continues to increase recall as k grows.

This gap highlights a trade-off: although larger k improves coverage, it introduces more distractor relations and longer prompts, which hinder LLM disambiguation. Moderate values of k thus offer the best balance between coverage, accuracy, and inference cost. Accordingly, we fix $k = 5$ in all experiments.

5.4 Qualitative and Error Analysis

We conduct a confusion analysis comparing HYDRE against *PARE* and zero-shot GPT-4o on the English NYT-10m dataset (Table 9). We further identify 12 semantically overlapping relation pairs that frequently exhibit confusion (Figures 13-14). Figures 3 – 12 (§A.6) illustrate HYDRE's reasoning behavior through working and failure cases.

Overall, HYDRE demonstrates a superior ability to resolve closely related relations that baselines conflate. For instance, it successfully disambiguates `Majorshareholders` from `Founders` (Fig. 3), correctly predicts `place_of_burial` instead of defaulting to the superficially similar `place_lived` (Fig. 5), and accurately

identifies fine-grained corporate relations like `company/advisors` (Fig. 10).

Crucially, HYDRE exhibits *strong multi-label support*. By conditioning the LLM on representative exemplars that maximize bag-level label coverage, HYDRE effectively recovers exhaustive label sets for a single entity pair where baselines only capture one (Fig. 6; Table 10).

Moreoever, an analysis of failure cases reveals three primary error categories:

1. **Position Bias:** HYDRE suffers from recency bias (Zheng et al., 2023), selecting *PARE*'s top-ranked candidate in 34% of errors. Placing the top-ranked exemplar closest to the query causes HYDRE to occasionally fixate on it. While HYDRE's overall multi-label coverage is superior, this hyper-focus can lead to partial multi-label prediction for complex queries (Fig. 7, 12).

2. **Candidate Bottleneck:** Performance drops when the gold relation is absent from *PARE*'s top-5 candidates ($\sim$14% of queries), trailing zero-shot prompting (Table 12). Here, the LLM is misled by grounding its reasoning on noisy or irrelevant exemplars.

3. **Persistent Semantic Overlap:** Some relations remain intrinsically ambiguous, often mirroring noisy gold annotations. For example, both HYDRE and zero-shot predict *nationality* for the demonym "American" where the gold label is *ethnicity* (Fig. 9), or misclassify *business/location* as *place_founded* even when given the cues such as "based in" (Fig. 8).

5.5 Additional Evaluation on Wiki-20m

We additionally test the HYDRE's generalizability on Wiki-20m (Gao et al., 2021), which features a significantly larger ontology of 81 relations. As shown in Table 4, HYDRE consistently outperforms zero-shot baselines across all models, with micro-F1 gains of up to 36 points for Llama-3.1 and Qwen3. The *w/o ICL* variant – which restricts the LLM to *PARE*'s top-5 candidate relations without providing exemplars –already shows a substantial improvement over the standard zero-shot baseline. This underscores the efficacy of using a specialized DSRE model to narrow the candidate space from 80 relations down to 5, significantly reducing task complexity for the LLM. Furthermore, the *w/o Stage 3* ablation (entire bag retrieval) performs comparably to the full HYDRE pipeline. This is due to the dataset's low density, with an average of 1.8 sentences per training bag, which renders the additional noise-filtering and sentence-selection mechanisms of Stage 3 largely redundant compared to denser datasets like NYT-10m (5.3 sentences/bag).

Ablation	Llama3.1	Qwen3	GPT-4o
0-shot	13/13	33/33	55/48
HYDRE	**49/47**	**69/64**	67/63
w/o semantic similarity	27/33	42/45	54/58
w/o PARE confidence	25/19	52/50	58/51
w/o Stage 2	41/40	59/56	65/62
w/o Stage 3	48/**47**	68/**64**	**69/64**
w/o ICL	31/34	56/56	60/56

Table 4: F1 scores (micro/macro) on Wiki-20m test set.

6 Conclusions and Future Work

We propose HYDRE, a novel hybrid framework that leverages SoTA DSRE models for guiding modern day LLMs for the task via efficient In-Context Learning. We show the efficacy of HYDRE across both sentence-level and bag-level evaluation settings. We propose effective strategies to adapt HYDRE to four low-resource Indic languages under three cross-lingual transfer settings. Ablations show the efficacy of each component in HYDRE in both monolingual and cross-lingual settings.

HYDRE lays the foundation for further research in DSRE in context of the latest LLMs. Possible future works involve a more complex agentic workflow wherein the agents interact iteratively until convergence to arrive at the correct answer. Applying HYDRE to newly added entries in Wikipedia or to local news articles for regional languages would be an interesting and useful future direction.

Acknowledgements

Mausam is supported by grants from Google, Microsoft, Huawei, Verisk, Tower Research and Nick McKeown Professorship. Parag is supported by Shanthi and K Ananth Krishnan Young Faculty Chair Professorship. Parag and Mausam are supported by the IBM AI Horizon Networks (AIHN) grant. We thank IIT Delhi HPC compute facility, Microsoft AFMR program and Google Gemini Credits grant. Any opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views or official policies of the funding agencies.

Limitations

One potential limitation of our work is the high cost or latency of retrieval from large bags in the training corpus. This problem is escalated further for low-resource languages due to their high token fertility even w.r.t. to latest LLMs.

Further our framework is not tested for low-resource domains such as biomedical or finance domains involving more complicated semantic relationships between the evolving entities. Similarly, we have only performed experiments on four Indic languages, and have not been able to perform experiments on other low-resource language families due to unavailability of relation extraction data.

Evaluating whether the core principles of HYDRE's label-aware exemplar selection and three-stage filtering extend to fully supervised or human-annotated relation extraction settings remains an important direction for future work.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, E. Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks.

Abhyuday Bhartiya, Kartikeya Badola, and Mausam. 2022. DiS-ReX: A multilingual dataset for distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6191–6200, Online. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Tianyu Gao, Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1306–1318, Online. Association for Computational Linguistics.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Zhaorui Jian, Shengquan Liu, Wei Gao, and Jianming Cheng. 2024. Distantly supervised relation extraction based on non-taxonomic relation and self-optimization. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. 2025. Exploring the role of diversity in example selection for in-context learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2962–2966, New York, NY, USA. Association for Computing Machinery.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*.

Dongyang Li, Taolin Zhang, Nan Hu, Chengyu Wang, and Xiaofeng He. 2022. Hiclre: A hierarchical contrastive learning framework for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2567–2578.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023a. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large

language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Min Li, Cong Shao, Gang Li, and Mingle Zhou. 2023b. Hfmre: Constructing huffman tree in bags to find excellent instances for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12820–12832.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Tom M. Mitchell, Justin Betteridge, Andrew Carlson, Estevam R. Hruschka Jr., and Richard C. Wang. 2009. Populating the semantic web by macro-reading internet text. In *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, volume 5823 of *Lecture Notes in Computer Science*, pages 998–1002. Springer.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient continual pretraining of LLMs for low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2021. A data bootstrapping recipe for low-resource multilingual relation classification. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587.

Jian Ni, Taesun Moon, Parul Awasthy, and Radu Florian. 2020. Cross-lingual relation extraction with transformers. *arXiv preprint arXiv:2010.08652*.

Vipul Rathore, Kartikeya Badola, Parag Singla, et al. 2022. Pare: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–354.

Vipul Kumar Rathore, Aniruddha Deb, Ankish Kumar Chandresh, Parag Singla, and Mausam . 2024. SSP: Self-supervised prompting for cross-lingual transfer to low-resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15081–15102, Miami, Florida, USA. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. 2024. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1280, Miami, Florida, USA. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware

attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Sizhe Zhou, Yu Meng, Bowen Jin, and Jiawei Han. 2024. Grasping the essentials: Tailoring large language models for zero-shot relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13462–13486, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

A.1.1 LLMs

We use (1) unsloth/Meta-Llama-3.1-8B-Instruct-unsloth-bnb-4bit version of Llama 3.1 for local inference as well as for fine-tuning, (2) TogetherAI's unsloth/gemma-3-4b-it-unsloth-bnb-4bit for Gemma3, (3) TogetherAI's Qwen3 235B A22B Instruct 2507 FP8 Throughput for Qwen3 and (4) OpenAI's gpt-4o-2024-05-13 for GPT-4o.

For local inference and fine-tuning, we use a single NVIDIA A100 40GB GPU node.

A.1.2 Baselines

For few-shot prompting baselines, we flatten bags to sentences (assign all bag labels to each of it's sentences) and then perform sentence selection using (a) Random, (b) Top-k similarity-based, and (c) MMR-based retrieval.

Following (Kapuriya et al., 2025), we implement MMR using iterative selection as follows:

$MMR(q, s, S_{t-1}) = \alpha \cdot Sim(q, s) - (1 - \alpha) \cdot max_{s' \in S_{t-1}} Sim(s, s')$,

where S_{t-1} denotes the candidate set selected so far at time step t. Here, α trades-off the relevance with diversity and is tuned on English dev set. The optimal value of α is found to be 0.3 and is used across all experiments.

A.1.3 Fine-tuning details

Translate-Train: PARE Adaptation: In the translate-train setting, we first adapt mBERT to each target language X by pretraining it on monolingual corpora derived from *X_train*. This results in a language-specific variant denoted $mBERT_X$. For pretraining, we extend the vocabulary with up to 10,000 randomly initialized new tokens for language X. Subsequently, we fine-tune the PARE model on *X_train* using $mBERT_X$ as the encoder using a task-specific adapter to obtain *PARE-X*.

LLaMA 3.1 Fine-Tuning: We use Low-rank adaptation (LoRA) with $lora\_alpha$ = 64, $lora\_r$ = 16, $lora\_dropout$ = 0.0, Learning rate scheduler as Cosine and warmup ratio as 10%. For training, we use $per\_device\_train\_batch\_size$ = 8, $gradient\_accumulation\_steps$ = 4 and maximum training steps of 5000 on single NVIDIA A100 40GB GPU.

The NYT-10m training data consists of 41624 bags and so the number of effective training epochs is = 8*4*5000/41624 $\approx$ 4 epochs.

Semantic Retriever Training: As we do not have an off-the-shelf retriever supporting our target languages, we seek to use a task-specific retriever. Specifically, we leverage the sentence-encoder of *CIL-X* for embedding queries and examples with cosine similarity as their similarity scores.

A.2 Prompt details

Task Description: Choose all applicable relations between head and tail entities from the set below. Print each relation in a new line. If none of the relations are applicable, output 'NA'.
/people/person/nationality : head entity is a person and tail entity is a country
/time/event/locations : head entity is an event and tail entity is a location
/people/person/children : head entity is a person and tail entity is another person (child)
/business/company/advisors : head entity is a company and tail entity is a person (advisor)
/business/location : head entity is a business and tail entity is a location
/business/company/majorshareholders : head entity is a company and tail entity is a person or organization (major shareholder)
/people/person/place_lived : head entity is a person and tail entity is a location
/business/company/place_founded : head entity is a company and tail entity is a location
/location/neighborhood/neighborhood_of : head entity is a neighborhood and tail entity is a larger location (city, town)
/people/deceasedperson/place_of_death : head entity is a deceased person and tail entity is a location
/film/film/featured_film_locations : head entity is a film and tail entity is a location
/location/region/capital: head entity is a region and tail entity is a city (capital)
/business/company/founders : head entity is a company and tail entity is a person (founder)
/people/ethnicity/geographic_distribution : head entity is an ethnicity and tail entity is a location where the ethnicity is commonly found
/location/country/administrative_divisions : head entity is a country and tail entity is a subdivision (state, province)
/people/deceasedperson/place_of_burial : head entity is a deceased person and tail entity is a burial site
/location/country/capital : head entity is a country and tail entity is a city (capital)
/business/person/company : head entity is a person

and tail entity is a company they are associated with
/location/location/contains : head entity is a larger location and tail entity is a smaller location within it
/location/administrative_division/country : head entity is an administrative division (state, province) and tail entity is a country
/location/us_county/county_seat: head entity is a U.S. county and tail entity is the county seat
/people/person/religion: head entity is a person and tail entity is a religion
/people/person/place_of_birth : head entity is a person and tail entity is a location (birthplace)
/people/person/ethnicity: head entity is a person and tail entity is an ethnicity
NA : no relation from the set exists between the given entity pair
Input format:
Input: {sentence}
Output format:
Output: {relation}
Verbalizer:
Extract the relation based on exact string match
A sample format of input and output for exemplars is shown in Figure 5.

A.3 HYDRE Algorithm

Please refer to Algorithm 1.

A.4 Detailed Language-wise results

Please refer to tables 5, 6 and 7.

A.5 Ablations to Assess Stage 3 Design

The final stage of HYDRE selects representative sentences by aggregating confidence scores across all bag-level labels, and assigns the full relation set to each chosen exemplar in the prompt. This strategy favors sentences with strong multi-label evidence over those with potentially noisy, single-label signals.

To evaluate this design, we consider two ablations:

1. **Candidate-only scoring:** Sentences are selected based solely on confidence for the target candidate relation ($r \in \mathcal{R}'$), ignoring other bag labels. The chosen exemplars still retain their full original bag relation set in the prompt.

2. **Single-label prompting:** Building on the candidate-only selection above, exemplars are presented in the prompt with *only* the target candidate label, discarding the rest of the bag's relation set.

As shown in Table 8, candidate-only scoring remains competitive for GPT-4o, but aggregate scoring substantially improves smaller models such as Qwen3 and Llama3.1. For Llama3.1, macro-F1 increases from 35 to 47, indicating that label-aware selection stabilizes performance in weaker architectures. Additionally, restricting exemplars to a single label noticeably degrades GPT-4o's performance (Micro-F1: 63 $\rightarrow$ 59), demonstrating the importance of preserving the full bag relations in the prompt.

A.6 Additional Qualitative Analysis

We present the confusion matrices for English NYT-10m across all relations in Table 9 and those for 12 manually identified, semantically confusing relation pairs in Figures 13 and 14. A few working and error examples of HYDRE are presented in Figures 5 - 12.

HYDRE's multi-label coverage. To test HYDRE's multi-label efficacy, we measure recall specifically on multi-label queries (Table 10). HYDRE improves micro- and macro-recall over zero-shot prompting by +9 and +11 points, respectively, indicating it actively recovers multiple relations.

We attribute this broad coverage to HYDRE's Stage 3 design, which retains all bag relations in the prompt. Comparing HYDRE against the single-label prompting ablation (§A.5) confirms this behavior (Table 11). While single-label prompting yields a marginal 1-point micro-F1 gain on single-label queries, it triggers a severe 12-point drop on multi-label queries.

Given that relation extraction is inherently a multi-label problem, HYDRE's ability to preserve multi-label signals in the prompt is critical for robust overall performance.

The Missing Candidate Bottleneck. Finally, we explicitly evaluate HYDRE's resilience when the gold relation is absent from *PARE*'s top-5 candidates, breaking down performance by single- and multi-label queries (Table 12).

Unsurprisingly, HYDRE consistently outperforms both baselines when the gold label is present in the top-5. However, even when the gold label is missing, HYDRE maintains meaningful performance. While it trails zero-shot prompting on

Algorithm 1 Exemplar Selection for HYDRE

Input: Query sentence q; bags of sentences $\mathcal{B} = \{B_j\}$; trained DSRE model scores $f_{PARE}(s, r)$ for sentence s and relation r; semantic similarity function $\text{sim}(q, B_j)$; confidence threshold τ; trade-off parameter λ; number of exemplars k.
Output: Ordered list of selected exemplar sentences.

Compute $f_{PARE}(q, r)$ for all relations r in the ontology.
$\mathcal{R}' \leftarrow$ top-k relations ranked by $f_{PARE}(q, r)$ ▷ Candidate relation selection
Initialize exemplar list $\mathcal{E} \leftarrow [\,]$.
for each $r \in \mathcal{R}'$ **do**
  $\mathcal{B}_r \leftarrow \{B_j \in \mathcal{B} \mid r \in \text{labelset}(B_j)\}$.
  $B_r \leftarrow \arg\max_{B_j \in \mathcal{B}_r} [(1-\lambda)\text{sim}(q, B_j) + \lambda f_{PARE}(B_j, r)]$. ▷ Select most relevant bag for r
  For each $s \in B_r$, compute $c(s) = \sum_{r_a \in \text{labelset}(B_r)} \mathbb{I}[f_{PARE}(s, r_a) > \tau]$. ▷ Compute coverage for each sentence in bag
  Let $v_{\max} \leftarrow \max_{s \in B_r} c(s)$. ▷ max. no. of relations with confidence > threshold
  $\mathcal{S}' \leftarrow \{s \in B_r \mid c(s) = v_{\max}\}$. ▷ candidate sentences with max. label coverage
  $s^* \leftarrow \arg\max_{s \in \mathcal{S}'} \sum_{r_a \in \text{labelset}(B_r)} f_{PARE}(s, r_a)$. ▷ Pick sentence with highest aggregate confidence
  Add $(s^*, \text{labelset}(B_r))$ to $\mathcal{E}$.
end for
Sort $\mathcal{E}$ in ascending order of $f_{PARE}(q, r)$. ▷ keep most informative examples at the last (closer to query)
return Ordered list of exemplar sentences $\{s^*\}$ from $\mathcal{E}$.

Model	English	Orya	Santali	Manipuri	Tulu	Avg. (Micro)†	Avg. (Macro)†
Supervised							
PARE (Rathore et al., 2022)	42/31	–	–	–	–	–	–
CIL (Chen et al., 2021)	43/32	–	–	–	–	–	–
zero-shot							
Qwen3-235B-A22B	49/40	46/40	6/5	2/1	45/39	25	21
Llama3.1-8b	31/17	26/21	21/14	10/7	29/21	22	16
Llama3.1-8B-FT$_{En}$	60/44	33/23	15/8	11/4	36/21	24	14
GPT-4o	56/57	56/56	10/5	11/7	55/**55**	33	31
5-shot							
HYDRE(QWEN3-235B-A22B)	**63*/62**	52/46	9/7	2/2	53/48	29	26
HYDRE(LLAMA3.1-8B)	52/47	37/24	**24*/16**	**21*/11**	38/25	30	19
HYDRE(LLAMA3.1-8B-FT)$_{En}$	61/45	47/32	**24***/13	**21***/8	49/31	35	21
HYDRE(GPT-4O)	**63***/60	**57/57**	18/10	11/8	**56/55**	**36***	**33**

Table 5: Results for **English-only** data setting. In each entry, we report micro and macro F1 scores. †The reported average scores are over non-English languages. * McNemar's p-value $< 10^{-5}$ (valid for micro-F1 comparison).

single-label queries in this regime – likely due to the LLM grounding on incorrectly retrieved exemplars—this scenario accounts for only $\sim$14% of all queries. Crucially, for multi-label queries where the gold label is missing, HYDRE still surpasses zero-shot prompting (+3 F1). This demonstrates that exemplar grounding remains beneficial for complex queries even when the initial DSRE candidate recall is imperfect.

A.7 Recall@5 for Supervised DSRE models

We present Recall@5 for both *PARE-X* and *CIL-X* models in table 13 for English and other languages.

A.8 Scalability w.r.t. No. of candidates k

Please refer to Fig. 15

A.9 Detailed Language-wise Ablations

Results are presented in tables 14, 15, 16 and 17 for each of the 4 LLMs.

A.10 Dataset statistics

A.10.1 Test data split

We construct the test split from NYT-10m (Gao et al. 2021) using stratified sampling, ensuring a minimum of 30 instances per relation, except when a relation contains fewer than 30 instances in total. This results in 538 sentences, with the distribution of relation counts shown in Table 18. Since NYT-10m is a multi-label dataset, the total number of relation instances in the test split exceeds the number of sentences, amounting to 722.

A.10.2 Training data

We take the original training data from NYT-10m (Gao et al. 2021) and ensure that "NA" bags do not

Model	ory_Orya	sat_Olck	mni_Mtei	tcy_Tulu	Avg. (Micro)	Avg. (Macro)
Supervised						
PARE-X	31/20	29/20	28/19	30/19	30	20
CIL-X	29/18	22/15	25/19	29/20	26	18
zero-shot						
Qwen3-235B-A22B	46/40	6/5	2/1	45/39	25	21
Llama3.1-8b	26/21	21/14	10/7	29/21	22	16
GPT-4o	56/**56**	8/6	8/6	56/53	32	30
Llama3.1-8B-FT$_{En}$	33/23	10/7	2/1	36/21	20	13
Llama3.1-8B-FT$_{X}$	51/37	33/18	24/11	44/26	38	23
5-shot						
HYDRE(Qwen3-235B-A22B)	54/52	23/14	15/5	57/54	37	31
HYDRE(Llama3.1-8B)	36/24	25/18	24/13	39/26	31	20
HYDRE(GPT-4o)	**57/56**	20/11	21/14	**58/57**	39	**35**
HYDRE(Llama3.1-8B-FT$_{En}$)	48/33	27/15	21/8	47/31	36	22
HYDRE(Llama3.1-8B-FT$_{X}$)	56/40	**39*/23**	**34*/15**	50/34	**45***	28

Table 6: Results for **Translate-train** setting. * McNemar's p-value $< 10^{-5}$.

Model	ory_Orya	sat_Olck	mni_Mtei	tcy_Tulu	Avg. (Micro)	Avg. (Macro)
Supervised						
PARE	35/25	31/21	33/25	32/22	33	23
CIL	34/23	33/21	34/25	36/25	34	24
zero-shot						
Qwen3-235B-A22B	44/37	43/34	43/35	44/34	44	35
Llama3.1	29/25	26/22	29/25	30/26	29	25
Llama3.1-8B-FT$_{En}$	49/35	45/29	49/35	52/39	49	34
GPT-4o	53/51	50/45	55/**57**	56/51	54	51
5-shot						
HYDRE(Qwen3-235B-A22B)	52/46	51/42	55/47	56/48	54	46
HYDRE(Llama3.1-8B)	46/38	41/31	46/38	44/38	44	36
HYDRE(Llama3.1-8B-FT$_{En}$)	51/37	50/33	50/37	54/40	51	37
HYDRE(GPT-4o)	**55*/54**	**53*/49**	**56/57**	**59*/55**	**56***	**54**

Table 7: Results for **Translate-test** setting. * McNemar's p-value $< 10^{-5}$.

Ablation Variant	GPT-4o	Qwen3	Llama3.1
HYDRE (Aggregate Scoring)	**63**/60	**63/62**	**52/47**
Candidate-only Scoring	**63/62**	62/60	47/35
+ Single-Label Prompting	59/58	58/56	41/41

Table 8: **HYDRE's stage 3 design ablation** (micro/macro-F1, English NYT-10m).

	PARE Correct	PARE Wrong
HYDRE Correct	172	223
HYDRE Wrong	25	272
	0-Shot Correct	**0-Shot Wrong**
HYDRE Correct	312	83
HYDRE Wrong	30	267

Table 9: **Confusion analysis for HYDRE vs. the baselines**, aggregated over all relations (English, NYT-10m).

exceed 10% of total bags to avoid model overfit on "NA" label. This leads to a total number of 41624 training bags.

A.11 Data annotation for Indic languages

A.11.1 Annotator details

We conduct human verification for four Indic languages using native speakers—either students or IT professionals—who were proficient in reading and typing in their respective scripts. Each annotator was compensated approximately $60 for verifying translations of 538 sentences. Prior to annotation, the speakers were informed that the task was intended solely for research purposes and posed no risk to them.

Each annotator was presented with the following questionnaire, with binary (YES/NO) responses and rectifications requested in case of a NO:

1. **Q1.** Is the translation of the given English sentence correct?
2. **Q2.** Is the *head entity* correctly translated into your native language?

Method	Micro Recall	Macro Recall
PARE	24	19
0-shot	38	30
HYDRE	**47**	**41**

Table 10: **Multi-label coverage.** Comparison of recall metrics on multi-label queries, demonstrating HYDRE's ability to actively recover multiple relations.

Method	Single-label	Multi-label
0-shot prompting	59	52
single-label prompting	**65**	49
HYDRE	64	**61**

Table 11: **Impact of single-label prompting.** Performance breakdown (Micro-F1) of GPT-4o on single-label vs. multi-label queries when restricting exemplars to a single relation label.

3. **Q3.** Is the *head entity* correctly projected in your native language?

4. **Q4.** Is the *tail entity* correctly translated into your native language?

5. **Q5.** Is the *tail entity* correctly projected in your native language?

A.11.2 Quality Assessment and Interannotator Agreement

We first assess the quality of the system-generated translations presented to annotators. Native speakers across all languages found the translations to be generally decent—likely due to high-quality output from IndicTrans2 (and Google Translate in the case of Tulu).

To quantify this, Table 19 reports:

1. The percentage of translations that required no human correction

2. The character-level F1 score (Char-F1) between the original system-generated translation and the human-corrected version (for cases requiring rectification).

To further evaluate annotation reliability, we conducted an inter-annotator agreement study. A second native speaker independently judged the quality of translations for 100 randomly sampled sentences in each language. Agreement is reported as the percentage of samples where the second speaker's judgment matched that of the first. Results are shown in Table 20.

	Single-Label		Multi-Label	
Method	**Micro-F1**	**Macro-F1**	**Micro-F1**	**Macro-F1**
Gold in top-5 (427 queries: Single - 323, Multi - 104)				
PARE	49.6	27.5	43.0	16.8
0-shot	63.7	48.6	54.5	23.5
HYDRE	**69.6**	**53.6**	**63.1**	**31.5**
Gold out of top-5 (111 queries: Single - 74, Multi - 37)				
PARE	0.0	0.0	0.0	0.0
0-shot	**42.3**	**15.9**	51.6	30.6
HYDRE	27.5	10.6	**54.5**	**34.6**

Table 12: **Impact of DSRE candidate recall**. Performance breakdown based on whether gold relations are in *PARE*'s top-5.

Language	*PARE*	*CIL*
English	**84**	82
Oriya	**75**	72
Santali	**72**	64
Manipuri	69	69
Tulu	69	**71**
Avg.*	**71**	69

Table 13: **Recall@5 scores for *PARE* and *CIL* models** on all languages. *Averaged over Indic languages

In summary, on average, 74% of system-generated translations were accepted as correct by native speakers, and for the remaining, the human-corrected outputs had a 93% Char-F1 match with the original translations. Inter-annotator agreement for human-corrected outputs averaged 91%, indicating strong consistency and translation quality across languages.

Ablation variant	Ory	Sat	Mni	Tcy	mi.	mu.
HYDRE	36/24	25/18	24/13	39/26	31	**20**
w/o sem.sim.	38/22	24/16	**25/13**	**42/28**	32	**20**
w/o f_{PARE}	34/24	21/16	14/10	35/22	26	18
w/o both (Random)	33/22	24/14	17/10	30/18	26	16
w/o stage 1	**39/26**	**33/20**	23/10	38/25	**33**	**20**
w/o ICL	32/21	28/19	25/16	30/19	29	19

Table 14: Average F1 scores for **Llama3.1-8b** over our target languages for different ablation variants of our few-shot approach in *translate-train* setting

Ablation variant	Ory	Sat	Mni	Tcy	mi.	mu.
HYDRE	**56/40**	**39/23**	34/15	50/34	**45**	**28**
w/o sem. sim.	**56**/39	**39**/22	**35/16**	**51/36**	**45**	**28**
w/o f_{PARE}	54/**40**	38/**22**	33/14	49/32	44	27
w/o *PARE* (Random)	53/34	31/15	21/8	47/28	38	21
w/o stage 1	54/35	35/21	33/13	48/30	43	25
w/o ICL	50/30	27/15	10/6	44/22	33	18

Table 15: Average F1 scores for **Llama3.1-8b-FT**$_X$ over our target languages for different ablation variants of our few-shot approach in *translate-train* setting

Ablation variant	Ory	Sat	Mni	Tcy	mi.	mu.
HYDRE	54/52	23/14	15/5	57/**54**	37	**31**
w/o sem. sim.	**56**/53	23/12	14/6	57/52	**38**	**31**
w/o f_{PARE}	55/53	**24/15**	15/5	55/52	37	**31**
w/o both (Random)	55/**55**	19/11	15/5	**58**/53	37	**31**
w/o stage 1	51/52	17/9	**16/7**	52/50	34	30
w/o ICL	54/36	15/8	8/4	53/33	33	20

Table 16: Average F1 scores for **Qwen3-235B-A22B** over our target languages for different ablation variants of our few-shot approach in *translate-train* setting

Ablation variant	Ory	Sat	Mni	Tcy	mi.	mu.
HYDRE	57/56	20/11	21/14	**58/57**	**39**	**35**
w/o sem. sim.	**58**/56	17/11	20/**15**	57/55	38	34
w/o f_{PARE}	57/**57**	18/**12**	**22/15**	56/52	38	34
w/o both (Random)	55/55	15/8	16/10	57/55	36	32
w/o stage 1	57/56	**20**/11	20/**15**	57/**57**	**39**	**35**
w/o ICL	54/38	12/6	12/7	48/32	32	21

Table 17: Average F1 scores for **GPT-4o** over our target languages for different ablation variants of our few-shot approach in *translate-train* setting

ICL Exemplars (HYDRE)

Input: Mr. Marek started the <Head> Fairfield </Head> Theater Company in 2001 ... at <Tail> Fairfield University </Tail> .
Output: /location/location/contains

Input: But Mr. Wallace, 45, a businessman and an investor in partnership for a time with <Tail> Mark Thatcher </Tail> , son of the former British Prime Minister <Head> Margaret Thatcher </Head>, has detractors, including ... . ”
Output: /people/person/children

Input: Behind the News – Founded eight years ago in a Silicon Valley garage by two <Tail> Stanford University </Tail> graduate students, <Head> Google </Head> went public two years ago at $ 85 a share.
Output: /business/company/place_founded

Input: A notch or two down-market, the 777 ’s, 767 ’s, and 757 ’s are often coveted by corporate titans, among them Larry Page and <Tail> Sergey Brin </Tail>, the co-founders of <Head> Google </Head>, who bought ... .
Output: /business/company/founders /business/company/majorshareholders

Input: When he was 32 , <Tail> Bill Hambrecht </Tail> was a co-founder of <Head> Hambrecht & Quist </Head> , a West Coast investment bank that ... .
Output: /business/company/founders

Query:

Input: Mr. Bechtolsheim , one of the first investors in <Head> Google </Head> , co-founded Kealia in 2001 with <Tail> David Cheriton </Tail> , a Stanford professor who was another early Google investor .

Outputs:

HYDRE: **/business/company/majorshareholders**
0-shot: **/people/person/company**
PARE: **NA**

Figure 3: **Working example 1:** HYDRE correctly predicts the “*Majorshareholders*" relation while both 0-shot and *PARE* fail to recognize it.

ICL Exemplars (HYDRE)

Input: .... [illegible] **Output**: /business/company/place_founded

Input: [illegible] ... 1,51,000 [illegible] ‖ (*Cardinal Health reported that it spent $151,000 ... for its Chairman and Chief Executive Robert D. Walter.*)
Output: /business/company/founders

Input: [illegible]
Output: /people/person/nationality /business/person/company

Query:

Input: [illegible] ... *(David Dagon , a Georgia Institute of Technology researcher who is a co-founder of Damballa, a start-up company focusing on controlling botnets , said ... .))*

Outputs:

HYDRE: **/business/company/founders**
0-shot: **NA**
PARE: **/business/person/company**

Figure 4: **Working example 2:** HYDRE correctly predicts “*founders*" while 0-shot and *PARE* fail to predict it.

ICL Exemplars (HYDRE)

Input: Mr. Marek started the <Head> Fairfield </Head> Theater Company in 2001 ... at <Tail> Fairfield University </Tail> .
Output: /location/location/contains

Input: But Mr. Wallace , 45 , a businessman and an investor in partnership for a time with <Tail> Mark Thatcher </Tail> , son of the former British Prime Minister <Head> Margaret Thatcher </Head> , has detractors , including ... . ”
Output: /people/person/children

Input: Behind the News – Founded eight years ago in a Silicon Valley garage by two <Tail> Stanford University </Tail> graduate students , <Head> Google </Head> went public two years ago at $ 85 a share .
Output: /business/company/place_founded

Input: But <Tail> Chicago </Tail> , the birthplace of <Head> Curtis Mayfield </Head> and Quincy Jones and the locus of King Oliver ’s Creole Jazz Band , has its own clanging , speakeasy sway .
Output: /people/person/place_of_birth

Input: This ocean of gray stone is <Tail> Paris </Tail> ’s most literary cemetery , holding the graves of Maupassant , Beckett , Duras and Man Ray , as well as arrivistes like Susan Sontag and <Head> Serge Gainsbourg </Head> .
Output: /people/person/place_lived

Query:

Input: The restaurant is across the street from <Tail> Atlanta </Tail> ’s oldest and loveliest cemetery , Oakland Cemetery (404-688-2107 ; www.oaklandcemetery.com) , where <Head> Margaret Mitchell </Head> , 25 mayors and thousands of unidentified Confederate soldiers are buried .

Outputs:

Gold: place_of_burial
HYDRE: **place_of_burial**
0-shot: **place_lived**
PARE: **place_lived**

Figure 5: **Example 1:** HYDRE correctly predicts “place_of_burial" – absent from PARE’s top-5 – while 0-shot and PARE predict “place_lived".

ICL Exemplars (HYDRE)

Input: ... Mr. Koppel to take another look at a once-unknown man , <Head> Morrie Schwartz </Head> , a <Tail> Brandeis University </Tail> professor who ... .
Output: person/company

Input: ... as you race away from the pleasant corporate maw of <Tail> Seattle </Tail> , from Starbucks and <Head> Boeing </Head> , Amazon and Microsoft .
Output: company/place_founded

Input: ... <Head> Ernest Hemingway </Head> was born in <Tail> Oak Park </Tail> in 1899 and lived here through high school .
Output: place_of_birth, place_lived

Input: ... Mr. Narayanan ’s body will be cremated ... in <Tail> New Delhi </Tail> , near the funeral ground of <Head> Jawaharlal Nehru </Head> , India ’s first prime minister ... .
Output: place_of_death, place_lived

Input: The easiest ... way to see <Tail> Philadelphia </Tail> is to stick with the older , central parts of town , emulate <Head> Benjamin Franklin </Head> ...
Output: place_of_death, place_lived

Query:

Input: Poe , Evermore A mystery man arrived at <Head> Edgar Allan Poe </Head> ’s grave at the Westminster Burial Grounds in <Tail> Baltimore </Tail> on Friday morning , as he has on Poe ’s birthday (Jan. 19) every year since 1949 , ...

Outputs:

Gold: place_lived, place_of_burial
HYDRE: **place_of_burial, place_lived**
0-shot: **place_of_burial**
PARE: **place_lived**

Figure 6: **Example 2:** HYDRE recovers both “place_of_burial" and “place_lived”; 0-shot and PARE each capture only one.

ICL Exemplars (HYDRE)

Input: ... Mendelssohn , who was born Jewish and converted to Christianity , and <Head> Otto Klemperer </Head> , who converted to Christianity and then back to <Tail> Judaism </Tail> .
Output: /religion

Input: As the <Head> <Tail> Baltimore </Tail> Orioles </Head> return home ... , Baltimore embraces its rich sports and maritime history .
Output: /business/location

Input: <Head> Lorenzo Da Ponte </Head> , a Bridge From <Tail> Italy </Tail> to New York " includes three vocal recitals , beginning tonight with ... .
Output: /nationality

Input: The Museum of Modern Art 's exhibition of four films starring the <Tail> Italian </Tail> actress <Head> Laura Morante </Head> concludes this weekend with four films , including ... .
Output: /ethnicity

Input: ... <Head> Madhesi </Head> ethnic group , which by some estimates represents as much as a third of <Tail> Nepal </Tail> 's population of 29 million , has been granted citizenship rights ... .
Output: /geographic_distribution

Query:

Input: Among the performances of note : ... the <Head> **Italian** </Head> dancer ALESSANDRA FERRI gives her final performance with the company on Saturday night , with ROBERTO BOLLE , a guest artist also from <Tail> **Italy** </Tail> .

Outputs:

Gold: /geographic_distribution, /nationality
HYDRE: **/geographic_distribution**
0-shot: **/nationality**
PARE: **NA**

Figure 7: **Example 3:** HYDRE and 0-shot each partially predict one of two gold labels (`geographic_distribution` vs. `nationality`).

ICL Exemplars (HYDRE)

Input: ... owned by a <Tail> Cincinnati </Tail> company , American Financial Group , whose chairman and chief executive officer is Carl H. Lindner III , who is also an owner of the <Head> Cincinnati Reds </Head> .
Output: /business/location

Input: A biomedical research institute in <Tail> Chengdu </Tail> , <Head> China </Head> , is planning to show true commitment to scientific principles ... .
Output: /location/location/contains /location/country/administrative_divisions

Input: <Tail> Robert Bigelow </Tail> , the founder of <Head> Budget Suites of America </Head> , is likely to push forward ... .
Output: /founders

Input: ... the 777 's , 767 's and 757 's are often coveted by corporate titans , among them Larry Page and <Tail> Sergey Brin </Tail> , the co-founders of <Head> Google </Head> , who bought ... .
Output: /founders
/majorshareholders

Input: The N.F.L. ... is very popular , ... , " said Tony Ponturo , vice president for global media and sports marketing at <Head> Anheuser-Busch </Head> in <Tail> St. Louis </Tail> , ... "
Output: /place_founded

Query:

Input: <Head> **Nestlé** </Head> , based in <Tail> **Vevey** </Tail> , Switzerland , ...

Outputs:

Gold: **/business/location**
HYDRE: **/place_founded**
0-shot: **/place_founded**
PARE: **NA**

Figure 8: **Example 4:** Both HYDRE and 0-shot misclassify `business/location` as `place_founded`.

ICL Exemplars (HYDRE)

Input: ... the epic poem " Paterson " by <Head> William Carlos Williams </Head> , a native of <Tail> Rutherford </Tail> .
Output: /people/person/place_of_birth

Input: It goes on to list notable <Tail> Mississippi </Tail> writers including William Faulkner , <Head> Richard Wright </Head> , ...
Output: /people/person/place_lived

Input: ... , including Giocangga , the founder of the <Head> Manchu </Head> dynasty in <Tail> China </Tail> , and Niall of the Nine Hostages , ... "
Output: /people/ethnicity/geographic_distribution

Input: <Head> Anthony Trollope </Head> , the brilliant depicter of the 19th-century social strata in <Tail> England </Tail> , ...
Output: /people/person/nationality

Input: ... , Mr. Delli Colli worked with generations of <Tail> Italian </Tail> directors , including <Head> Pier Paolo Pasolini </Head> , ...
Output: /people/person/ethnicity

Query:

Input: Lukacs , ... , makes very large claims for his subject in " George Kennan " : He was " a better writer and a better thinker " than <Head> Henry Adams </Head> ; he was " the best and finest <Tail> American </Tail> writer about Europe " in the interwar years ... .

Outputs:

Gold: **/people/person/ethnicity**
HYDRE: **/people/person/nationality**
0-shot: **/people/person/nationality**
PARE: **NA**

Figure 9: **Example 5:** Both HYDRE and 0-shot confuse `ethnicity` with `nationality`.

ICL Exemplars (HYDRE)

Input: Mr. White , 54 , of <Tail> Centerport </Tail> , has held top environmental posts with <Head> Suffolk County </Head> and the Town of Huntington .
Output: /location/location/contains

Input: Being able to combine Loudeye services with <Head> Nokia </Head> terminals provides a good base for innovative and useful consumer services , " said Ilkka Raiskinen , senior vice president for multimedia experiences at Nokia , based in <Tail> Espoo </Tail> , Finland .
Output: /business/company/place_founded

Input: ... ' It 's a gentle soul with a naughty sense of humor , ' Jerry was just as responsible for that as my dad , " ' said <Tail> Brian Henson </Tail> , <Head> Jim Henson </Head> 's son , who serves with his sister Lisa as chairman and chief executive of the Jim Henson Company .
Output: /people/person/children

Input: ... the 777 's , 767 's and 757 's are often coveted by corporate titans , among them Larry Page and <Tail> Sergey Brin </Tail> , the co-founders of <Head> Google </Head> , who bought a used 767 last year and spent millions converting it into a private jet .
Output: /business/company/founders, /business/company/majorshareholders

Input: The service has more ways to shield users ' identities compared with many social networking sites , said <Tail> Antony Brydon </Tail> , <Head> Visible Path </Head> 's chief executive , above . "
Output: /business/company/founders

Query:

Input: " <Head> MySpace </Head> is dedicated to ensuring that content owners , whether large or small , can both promote and protect their content in our community , " <Tail> Chris DeWolfe </Tail> , the chief executive of MySpace , said in a statement . "

Outputs:

Gold: **/business/company/advisors**
HYDRE: **/business/company/advisors**
0-shot: **/business/person/company**
PARE: **NA**

Figure 10: **Example 6:** HYDRE correctly predicts `company/advisors`; 0-shot and PARE fail.

ICL Exemplars (HYDRE)

Input: [illegible] 2001 [illegible] ....

Output: /people/person/nationality /people/person/place_lived

Input: [illegible] ||

Output: /location/neighborhood/neighborhood_of

Input: .... [illegible] ||

Output: /business/company/place_founded

Query:

Input: [illegible] .... [illegible] || (*Mr. Flood said ... had bought the commercial operations of Allied Waste Industries in New York City, ...*)

Outputs:

Gold: /business/location
HYDRE: /business/location
0-shot: NA
PARE: NA

Figure 11: **Example 7 (Manipuri)**: HYDRE correctly predicts `business/location` despite its absence from PARE's top-5 candidates.

ICL Exemplars (HYDRE)

Input: [illegible] ||

Output: /location/location/contains, /location/country/capital

Input: [illegible] ||

Output: /location/location/contains, /location/country/administrative_divisions

Input: [illegible] || (*One of the newest auditoriums in New York, the bedrock building under Carnegie Hall added to the experience of the musical event.*)

Output: /location/location/contains

Query:

Input: [illegible] 1,700 [illegible] || (*This 1,700-acre project on the Sea of Cortez will cover a peninsula across a bay from the center of La Paz, the capital of Baja California Sur state .*)

Outputs:

Gold: /location/region/capital, /location/location/contains
HYDRE: /location/location/contains
0-shot: NA
PARE: /location/location/contains

Figure 12: **Example 8 (Manipuri):** HYDRE and PARE partially predict `contains`; both miss `region/capital`. 0-shot predicts NA.

Relation	Count
/people/person/place_lived	73
/people/person/nationality	34
/business/person/company	34
/people/person/place_of_birth	31
/location/location/contains	102
/location/country/administrative_divisions	58
/business/location	33
/location/administrative_division/country	31
/business/company/advisors	37
/business/company/founders	31
/business/company/majorshareholders	4
/location/neighborhood/neighborhood_of	30
/location/country/capital	30
/film/film/featured_film_locations	1
/location/us_county/county_seat	6
/people/person/children	30
/people/deceasedperson/place_of_death	30
/people/deceasedperson/place_of_burial	4
/people/ethnicity/geographic_distribution	30
/location/region/capital	12
/business/company/place_founded	4
/people/person/religion	21
/time/event/locations	3
/people/person/ethnicity	23
NA	30
Total	538

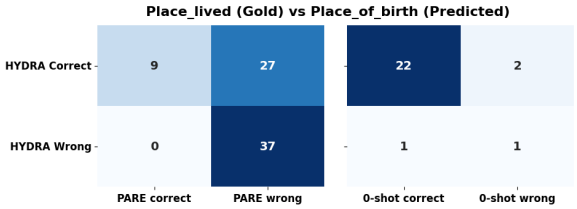
Table 18: **Label-wise statistics of NYT-10m test data.** Total number of sentences (538) in our test split is different from total number of labels (722) due to multi-label characteristics of NYT-10m dataset.

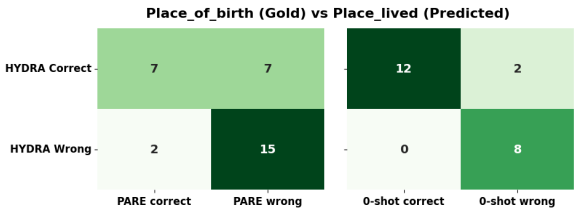
Language	No Rectification Needed (%)	Char-F1 Match (if rectified)
Oriya	69	92
Santali	72	88
Manipuri	83	96
Tulu	73	95
Average	74	93

Table 19: **Translation quality assessment.** Percentage of system translations requiring no rectification, and character-level F1 match for rectified translations.

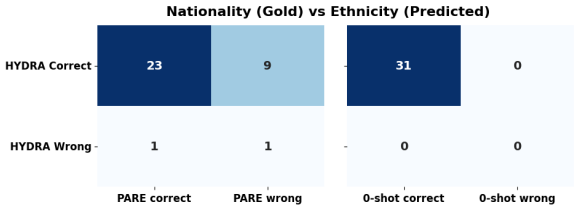
Lang.	Translation	Head projection	Tail projection
Oriya	92	92	92
Santali	89	84	87
Manipuri	85	98	100
Tulu	96	92	88
Average	91	92	92

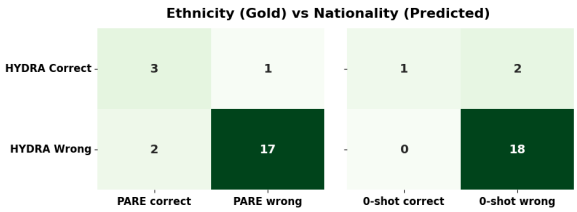
Table 20: **Inter-annotator agreement** for 100 samples in each language



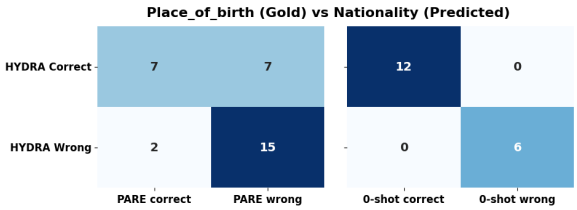


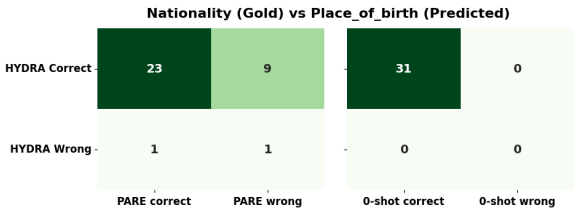
a Place of birth vs Place lived



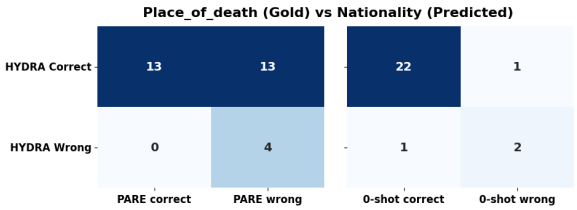


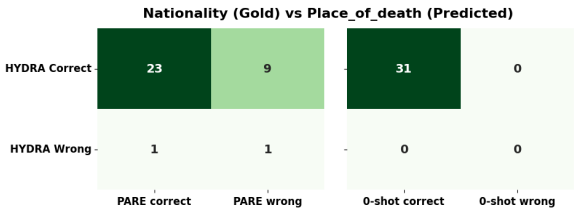
b Nationality vs Ethnicity



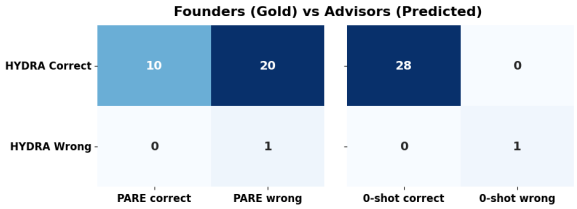


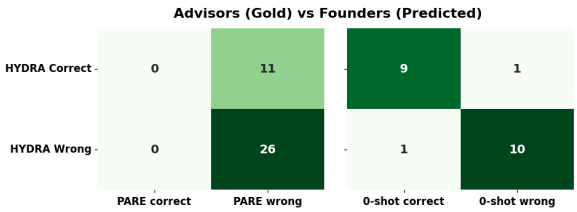
c Nationality vs Place of Birth



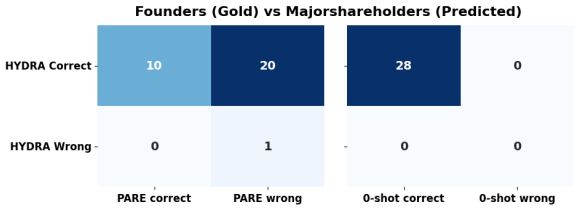


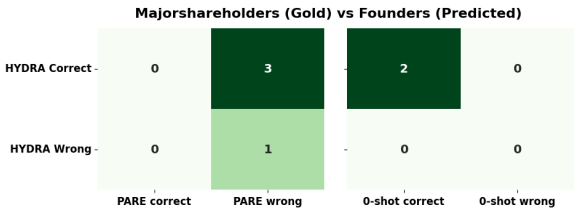
d Nationality vs Place of Death





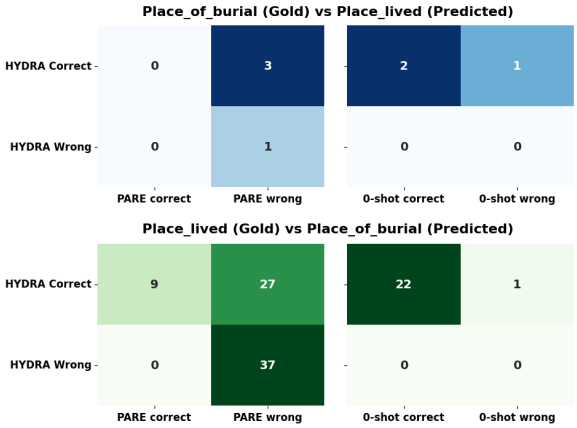
e Company founders vs Company Advisors



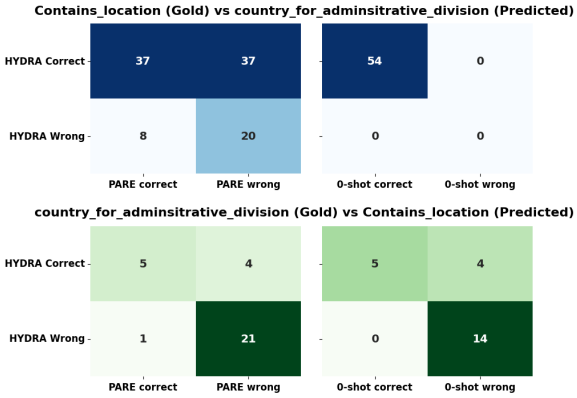


f Company founders vs Company Majorshareholders

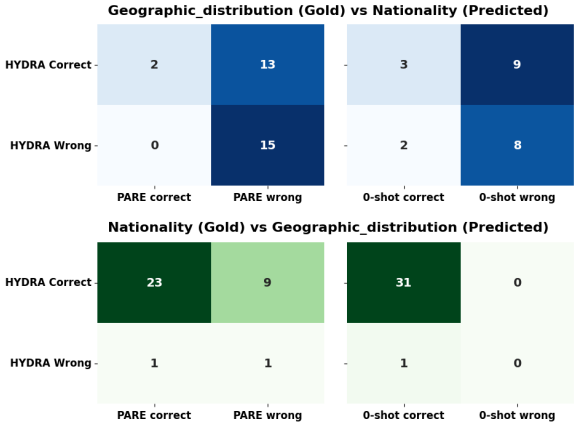
Figure 13: **Confusion matrices for HYDRE vs. the baselines (part 1 of 2)**.



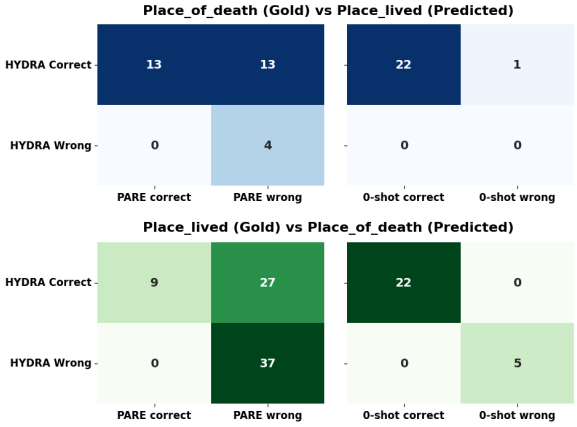
a Place of Burial vs Place lived



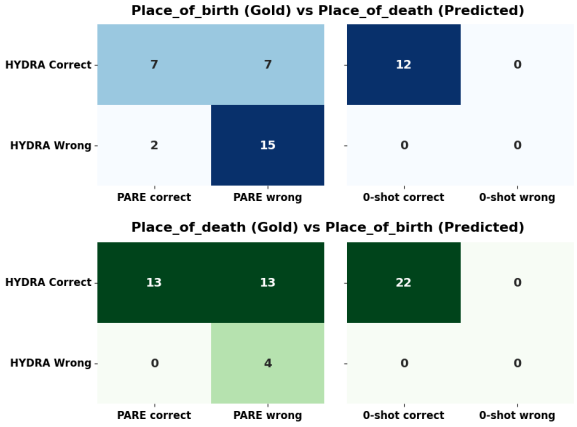
b Country for Administrative divisions vs Contains



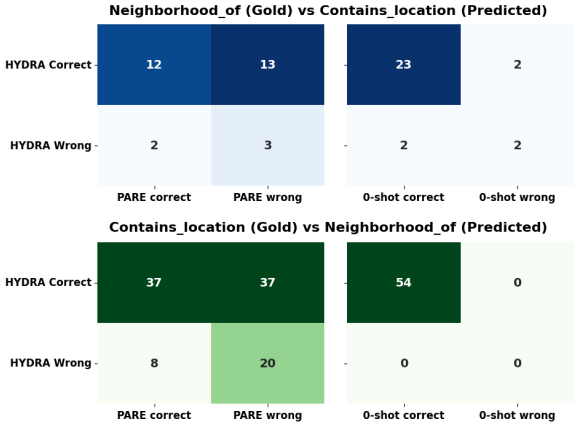
c Nationality vs Geographic distribution



d Place lived vs Place of Death

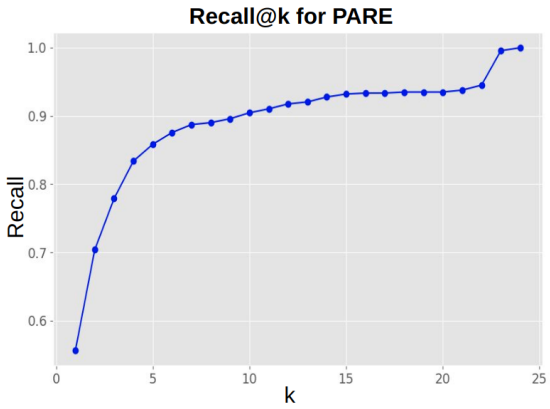


e Place of Birth vs Place of Death

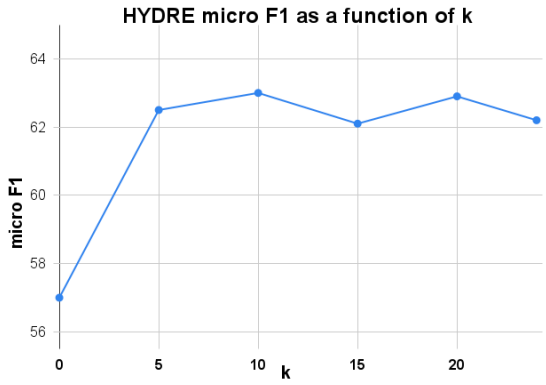


f Neighborhood_of vs Contains_location

Figure 14: **Confusion matrices for HYDRE vs. the baselines (part 2 of 2)**.



a *PARE* Recall@k v/s k



b HYDRE (GPT-4o) micro-F1 v/s k

Figure 15: Analysis of PARE's Recall@k and HYDRE (GPT-4o) downstream performance on NYT-10m English dev set