Relation Extraction or Pattern Matching? Unravelling the Generalisation Limits of Language Models for Biographical RE

Varvara Arzt[1,2], Allan Hanbury[1*], Michael Wiegand[3], Gábor Recski[1], Terra Blevins[4*]
[1]Faculty of Informatics, TU Wien, [2]D!ARC, University of Klagenfurt,
[3]Digital Philology, University of Vienna,
[4]Khoury College of Computer Sciences, Northeastern University
Correspondence: varvara.arzt@tuwien.ac.at

Abstract

Analysing the generalisation capabilities of relation extraction (RE) models is crucial for assessing whether they learn robust relational patterns or rely on spurious correlations. Our cross-dataset experiments find that RE models struggle with unseen data, even within similar domains. Notably, higher intra-dataset performance does not indicate better transferability, instead often signaling overfitting to dataset-specific artefacts. Our results also show that data quality, rather than lexical similarity, is key to robust transfer, and the choice of optimal adaptation strategy depends on the quality of data available: while fine-tuning yields the best cross-dataset performance with high-quality data, few-shot in-context learning (ICL) is more effective with noisier data. However, even in these cases, zero-shot baselines occasionally outperform all cross-dataset results. Structural issues in RE benchmarks, such as single-relation per sample constraints and non-standardised negative class definitions, further hinder model transferability. We release our dataset splits with sample IDs and code for reproducibility.[1]

1 Introduction

Relation extraction (RE) is the core information extraction task of identifying the semantic relationship between entities in text. Traditional RE evaluations rely predominantly on in-distribution testing, but this approach often overestimates true model performance by implicitly assuming that individual datasets wholly represent the underlying task (Linzen, 2020; Kovatchev and Lease, 2024). While model generalisation has gained increasing attention in NLP, RE remains relatively unexplored in this context (§ 2).

However, understanding RE generalisation to out-of-distribution (OOD) data is crucial both for the task itself as well as for the robust application of RE systems in downstream tasks like question answering and knowledge-base population (Bassignana and Plank, 2022a). Given the popularity of representing internal language model (LM) knowledge as relational triples (Geva et al., 2023; Hernandez et al., 2024), building robust RE systems beyond the mere memorisation of dataset-specific patterns may also be key to more interpretable and trustworthy models.

This paper systematically analyses how well RE systems generalise across datasets focusing on sentence-level RE. Due to the limited relation overlap in popular RE datasets, we focus our experiments on biographical relations, which are pervasive in RE settings; this also allows us to include a domain-specific dataset for grounded analysis (§ 3). Through our cross-dataset experiments, this paper makes the following contributions:

- We document key challenges in analysing RE generalisation, including inconsistent relation schemas and highly imbalanced class distributions (§ 3), as well as propose methods for overcoming these issues.
- We find that strong in-distribution RE performance often masks fundamental generalisation failures, with models that excel on intra-dataset evaluations frequently failing to transfer effectively (§ 5).
- Our cross-dataset analysis suggests that how data is annotated influences the best adaptation method: in our experiments, fine-tuning achieves better cross-dataset performance for manually annotated data, while few-shot in-context learning (ICL) performs better on distantly supervised data. However, zero-shot

*Supervised by
[1]https://github.com/kleines-gespenst/re_cross_dataset

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1463–1484

prompting outperforms all cross-dataset methods in some settings (§ 5.2).

- We identify structural issues in current RE benchmarks that lead to generalisation errors, including single-relation constraints, external knowledge reliance, and coverage biases (§ 6).

These findings reveal that while current RE systems achieve high in-distribution results, their cross-dataset performance shows critical gaps in genuine relation understanding, limiting their real-world applicability.

2 Related Work

2.1 Approaches for RE

RE is traditionally framed as a classification task, tackled via either a pipeline approach—where subtasks like named entity recognition (NER), coreference resolution, and relation classification (RC) are performed sequentially—or a joint model that processes them simultaneously (Taillé et al., 2020; Bassignana and Plank, 2022b; Saini et al., 2023). It is further categorised into sentence- (Alt et al., 2020; Plum et al., 2022) and document-level RE (Yao et al., 2019; Meng et al., 2024).

Since the introduction of BERT (Devlin et al., 2019), encoder-based models have dominated RE due to their bidirectional attention mechanism, which effectively captures context for classification tasks (Alt et al., 2020; Plum et al., 2022). However, the rise of autoregressive models has led to increasing adoption of decoder-based architectures to RE (Wang et al., 2022; Sun et al., 2023; Xu et al., 2023; Liu et al., 2024; Efeoglu and Paschke, 2024). While encoder-decoder models have been explored (Huguet Cabot and Navigli, 2021; Li et al., 2023b), our experiments focus on the dominant encoder-only and decoder-only architectures for RE.

2.2 Generalisation Capabilities of RE Models

Recent work advocates for transparent evaluation (Neubig et al., 2019; Liu et al., 2021) and OOD testing (Linzen, 2020; Allen-Zhu and Li, 2024; Qi et al., 2023) to assess model robustness. Common strategies include cross-dataset (Antypas and Camacho-Collados, 2023; Jang and Frassinelli, 2024) and cross-domain (Fu et al., 2017; Liu et al., 2020; Bassignana and Plank, 2022a; Calderon et al., 2024) experiments, as well as testing on perturbed and adversarial sets (Wu et al., 2019; Gardner et al., 2020; Goel et al., 2021; Rusert et al., 2022).

Recent studies have explored various ways to improve RE model robustness. Bassignana and Plank (2022a) introduce a cross-domain RE dataset with broad relation types, while Meng et al. (2024) and Chen et al. (2023) evaluate state-of-the-art (SOTA) document-level RE models on perturbed test sets. Chen et al. (2023) reveal that even when models predict correctly, they often rely on spurious correlations, showing vulnerability to minor evaluation shifts. To reduce dependence on mere pattern matching, Allen-Zhu and Li (2024) propose augmenting training data with synthetic samples reformulated by an auxiliary model. Most closely related to our work, Bassignana and Plank (2022b) analysed cross-dataset model transfer for scientific RE with large data overlap, while our work examines cross-dataset generalisation across models for general-purpose RE and analyses factors influencing transfer considering datasets with no overlap.

3 Methodology

We assess the RE systems' robustness by evaluating their OOD performance. Standard in-distribution evaluations may overestimate RE performance (Linzen, 2020), as models can exploit spurious cues rather than learning genuine RE task (Chen et al., 2023; Meng et al., 2024; Arzt and Hanbury, 2024).

To systematically evaluate generalisation capabilities of RE models, we conduct both intra- and cross-dataset experiments. The intra-dataset experiments act as a control, evaluating RE models on data drawn from the distribution used for model adaptation, while the cross-dataset experiments measure model robustness with OOD test sets derived from a different RE dataset.

For our experiments, we use three sentence-level RE datasets: TACRED-RE (Alt et al., 2020), NYT (Riedel et al., 2010), and Biographical (Plum et al., 2022). While TACRED-RE and NYT are general-purpose RE datasets, we focus only on biographical relations, or relations describing aspects of an individual's life like *place_of_birth* or *children* for two key factors: (1) TACRED-RE and NYT share six biographical relations but only two non-biographical ones, and (2) focusing on biographical relations allows for additional cross-dataset evaluations with the Biographical dataset, which only contains biographical relations. This setup thus allows us to compare the generalisation of two popular RE datasets in a third, held-out evaluation setting.

3.1 Data

We now briefly describe three RE datasets used.

TACRED-RE (Alt et al., 2020) is a general-purpose RE dataset with 41 relations and a 'no_relation' class.[2] It contains over 106k instances but is highly imbalanced, with ∼80% labeled as 'no_relation'. Built from English newswire and web text, it is a revised version of TACRED (Zhang et al., 2017), with challenging samples re-annotated by professional annotators to reduce noise from crowdsourcing. Experimental results show improved performance on TACRED-RE compared to TACRED (Appendix Tables 13, 14), leading to its use in our experiments. Figure 1 shows a TACRED-RE example. We focus on its 26 biographical relations, including 'no_relation' (Appendix Table 5).

per:city_of_birth

Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkady.

Figure 1: TACRED-RE example (Zhang et al., 2017).

NYT (Riedel et al., 2010) is a general-purpose RE dataset with 24 relations and a 'None' class. It contains over 266k sentences, with 64% labeled as 'None' and half of positive instances containing a single dominant relation, '/location/location/contains'.[3] NYT was constructed via distant supervision, by applying `Freebase` (Bollacker et al., 2008) as external supervision to New York Times articles (Sandhaus, 2008). Figure 2 shows an NYT example. We focus on its subset with 7 biographical relations, including a 'None' class (Table 6, Appendix).

"I think Amlo truly feels he's the Redeemer of Mexico, but his reign is of this world," said Enrique Krauze, a historian.

people/person/nationality

Figure 2: NYT example.

Biographical is an RE dataset for the biographical text domain, with 10 relation types (Plum et al., 2022). Built from Wikipedia articles on prominent individuals and containing 346,257 instances, Biographical was created using a semi-supervised approach.[4] Named entities were automatically extracted using `spaCy` (Honnibal et al., 2020) and `Stanford CoreNLP` (Manning et al., 2014). `Wikipedia` sentences with these entities were matched with `Pantheon` and `Wikidata` to automatically infer relations. Figure 3 shows an example[5] from Biographical. Statistics for Biographical, downsampled to match the size of the TACRED-RE and NYT subsets, appear in Appendix Table 6.

bplace_name

Kharlan was born in Mykolaiv a shipbuilding town in the south of Ukraine.

Figure 3: Biographical example.

3.2 Cross-Dataset Comparison: Challenges

Single Relation per Sample: Both TACRED-RE and Biographical restrict each sample to at most two entities and one relation, even when multiple relations exist in a sentence. For instance, the TACRED-RE example in Figure 1 is labeled with 'per:city_of_birth' but also contains 'per:stateorprovinces_of_residence', 'per:employee_of', and 'per:spouse', all within TACRED-RE's relation set. This constraint may confuse a model trained on such data, as it enforces a single-label assignment. While NYT better reflects real-world scenarios by allowing multiple relations per sentence, we filtered it to two-entity, single-relation samples for fair cross-dataset comparison, retaining only examples like in Figure 2.

Unclear 'Negative' Class: Clear negative samples—instances with entities but no meaningful relation—are crucial for RE. While TACRED-RE has an explicit 'no_relation' class, NYT's 'None' class lacks clear definition (Riedel et al., 2010), potentially confusing models about whether it indicates absence of predefined relations or any relation. Similarly, Biographical lacks an explicit negative class, using instead an 'Other' class for unspecified relations (Plum et al., 2022), exemplified in Appendix Figure 5. This inconsistency between choosing a 'no_relation' versus a 'none_of_the_above' class in RE benchmarks highlights the general challenge of consistently defining the boundaries between presence and absence of semantic relations in text (Bassignana and Plank, 2022b).

Expected Factual Knowledge: The design of RE datasets influences whether models genuinely

[2]Licensed by the Linguistic Data Consortium (LDC).
[3]Available at `https://github.com/INK-USC/ReQuest`.
[4]We use the *m2_normal_final1* version.

[5]Sample ID 'mS1/18860978' shows punctuation revomal by Plum et al. (2022) that may affect model comprehension.

learn RE or rely on dataset-specific cues. NYT's distant supervision approach incorporates Freebase-derived relations not stated in text requiring external world knowledge rather than textual evidence, as shown in Figure 2 where the text lacks explicit information about Enrique Krause's nationality—such annotations extend beyond RE's scope and corrupt models trained on such data. Similarly, although manually curated, TACRED-RE encompasses relations like 'per:city_of_birth', which require factual knowledge from a model, limiting generalisation to instances seen during adaptation.

3.3 Cross-Dataset Label Overlap

Following Bassignana and Plank (2022b), we manually analysed instances for each relation in TACRED-RE, NYT, and Biographical to establish a cross-dataset label mapping. Appendix Table 8 shows six overlapping biographical relations between NYT and TACRED-RE, with twelve fine-grained TACRED-RE relations mapping to six broader NYT labels (e.g., NYT's 'place_of_birth' encompasses three TACRED-RE birth location relations). Treating Biographical's 'Other' class as negative—supported by manual analysis of 30 random instances showing negative rather than unspecified relations—we find four overlapping relations across three datasets (Appendix Table 10). NYT and Biographical share these same four relations (Appendix Table 7), while TACRED-RE and Biographical share nine relations (Appendix Table 9).

4 Experiments

4.1 Data Format and Standardisation

To enable cross-dataset evaluations and focus exclusively on *relation classification*, we standardise our data using unified format with entity spans marked as `<e1>`head entity`</e1>` and `<e2>`tail entity`</e2>`.

To address class imbalance, we randomly downsampled negative instances across three datasets to balance the number of positive and negative instances, and downsampled Biographical (∼350k) to match other biographical subset sizes for fair comparison (Appendix A). For cross-dataset evaluation, we mapped TACRED-RE's fine-grained relations to broader NYT labels (Appendix C.1).

4.2 Model Selection, Training, and Evaluation

We consider two types of models: an encoder-only (DeBERTa-v3-large 304M; He et al. (2021)) and a decoder-only model (an instruction-tuned LLaMA 3.1 8B; Grattafiori et al. (2024)). For DeBERTa, we follow common practice for RE encoder models: we mark entity spans with custom markers and use the concatenated hidden states of the entity start tokens, `<e1>` and `<e2>`, as inputs for the classification head (Baldini Soares et al., 2019). We also evaluate three SOTA systems (or replications of these systems) on our biographical test sets, one for each considered dataset, in order to compare our models to prior work. We provide more details on our replication study in Appendix C.3.

We employ two commonly used model adaptation strategies for RE: fine-tuning and in-context learning (ICL). Specifically, we consider direct fine-tuning with DeBERTa, fine-tuning LLaMA using low-rank adaptation (LoRA; Hu et al., 2022), and zero-shot and five-shot ICL (Brown et al., 2020) with LLaMA. For few-shot ICL, we perform five runs with different demonstration sets to account for demonstration sensitivity (Zhang et al., 2022; Webson and Pavlick, 2022; Lu et al., 2022). However, due to computational constraints, fine-tuning experiments are limited to a single run. For NYT and TACRED-RE, we conduct experiments in two adaptation settings: adaptation on all biographical relations in each dataset (Appendix Tables 5 and 6) and adaptation on only overlapping relations (Appendix Table 8). This applies to both fine-tuning and ICL, where zero-shot prompts and few-shot demonstrations are selected accordingly.

We then perform two types of evaluations: **intra-dataset**, where models are evaluated on the same dataset they were adapted to; and **cross-dataset**, where the adapted models were tested on OOD data to assess their generalisation. Further implementation details, including hyperparameter settings, and prompting details, are provided in Appendix C.2.

5 Results

Overview of Reported Results: Table 1 shows intra- and cross-dataset results for NYT and TACRED-RE on six overlapping relations, using models adapted on all biographical relations. For TACRED-RE, which maps its 12 fine-grained labels to NYT's shared label space, we report both dataset-specific and shared label results, denoted as 'Dataset Labels' and 'Shared Labels' in Table 1. Table 2 shows model generalisation to Biographical across three overlapping relation sets: (1) four relations shared by all datasets, (2) same four relations shared between NYT and Biographical, and

Model	Setting	Dataset	Intra-Dataset		Cross-Dataset	
			Shared Labels	Dataset Labels	NYT	TACRED-RE
DeBERTa-v3 large 304M	Fine-tuned on	NYT	0.81	0.81	–	0.26
		TACRED-RE	0.72	0.62	0.53	–
LLaMA 3.1 8B	Fine-tuned on	NYT	**0.87**	0.87	–	0.45
		TACRED-RE	**0.82**	0.76	**0.62**	–
LLaMA 3.1 8B	Zero-Shot	NYT	0.31	0.31	–	–
		TACRED-RE	0.58	0.37	–	–
LLaMA 3.1 8B	5-Shot	NYT	0.45 ± 0.07	0.45 ± 0.07	–	**0.52 ± 0.06**
		TACRED-RE	0.63 ± 0.06	0.43 ± 0.07	0.39 ± 0.02	–

Table 1: Macro F1-scores for intra- and cross-dataset predictions on six overlapping relations. Results include shared and dataset-specific labels, with models adapted on all biographical relations via fine-tuning or ICL. Best intra- and cross-dataset results are in bold.

Model	Setting	Dataset	Full Overlap	Overlap w. NYT	Overlap w. TACRED-RE
DeBERTa-v3-large 304M	Fine-tuned on	NYT	0.48	0.48	–
		TACRED-RE	0.62	–	0.70
		Biographical	**0.80**	0.80	0.81
LLaMA 3.1 8B	Fine-tuned on	NYT	0.30	0.30	–
		TACRED-RE	**0.69**	–	0.70
		Biographical	0.79	0.79	0.74
LLaMA 3.1 8B	Zero-Shot	Biographical	0.24	0.24	0.35
LLaMA 3.1 8B	5-Shot	NYT	0.48 ± 0.04	0.48 ± 0.04	–
		TACRED-RE	0.51 ± 0.04	–	0.58 ± 0.02
		Biographical	0.53 ± 0.05	0.53 ± 0.05	0.54 ± 0.03

Table 2: Evaluation on Biographical Dataset (macro F1-scores). Models adapted on all biographical relations through fine-tuning or ICL. Best intra- and cross-dataset results on full overlap are in bold.

(3) nine TACRED-RE/Biographical shared relations. Models were adapted on each dataset's full biographical relations, with Biographical's intra-dataset results for comparison. We focus on results with models adapted on the full overlap, as they perform similarly to those adapted only on overlap (Appendix Table 16) while better reflecting real-world settings. Cross-dataset experiments with Biographical as training source appear in Appendix E. Given Biographical's ambiguous 'Other' class (§ 3.2), we use it only for evaluation in the main paper.

While we primarily focus on macro F1 to address class imbalance, we report micro F1 (commonly reported for RE tasks) for our fine-tuned models in Table 3; additional experimental results, including per-class breakdowns, appear in Appendix F.

5.1 Intra-Dataset Results

We evaluate our RE models on their training data distribution for comparison of cross-dataset generalisation. Unsurprisingly, we find that fine-tuning performs best for intra-dataset evaluations: fine-tuned LLaMA outperforms DeBERTa on TACRED-RE and NYT (Table 1), while DeBERTa outperforms LLaMA on Biographical (Table 2).

Our ICL experiments similarly show expected results, with the five-shot prompting moderately outperforming zero-shot prompting but underperforming full model fine-tuning. For Biographical, this few-shot ICL gain over zero-shot is significant, increasing from 0.24 to 0.53 ± 0.05 with five demonstrations (Table 2).

We observe different performance trends within TACRED-RE and NYT intra-dataset evaluations. While fine-tuning yields higher intra-dataset performance on NYT than on TACRED-RE, this trend flips for the zero- and few-shot ICL settings, with prompting on NYT performing significantly *worse* despite TACRED-RE's finer-grained relation schema. This likely stems from differing data quality between datasets: the noisy labeling during NYT creation (Yaghoobzadeh et al., 2017) likely leads to over-fitting during fine-tuning (Tänzer et al., 2022) (rather than learning robust relational patterns), but harms model generalisation to NYT when not fine-tuned for that data distribution.

Comparison with SOTA: As our cross-dataset analysis focuses on biographical relations, direct comparisons with prior work are challenging as they rarely report per-class results. Thus, we perform a replication study to re-evaluate prior work

Dataset	SOTA*		DeBERTa-v3 large 304M		LLaMA 3.1 8B	
	Macro	Micro	Macro	Micro	Macro	Micro
NYT	0.87	<u>0.90</u>	0.84	0.93	0.89	0.94
TACRED-RE	0.78	<u>0.87</u>	0.71	0.85	0.78	0.89
Biographical	<u>0.87</u>	0.93	0.83	0.92	0.75	0.91

Table 3: F1 score comparison of our models (trained on biographical subsets) vs. SOTA systems (trained on full datasets) when evaluated on biographical test sets. Underlined values indicate which metric (macro or micro F1) was originally reported in SOTA* papers from Orlando et al. (2024) (NYT), Zhou and Chen (2022) (TACRED-RE), and Plum et al. (2022) (Biographical).

on our biographical evaluation sets, comparing SOTA models—trained using full training and validation datasets—against our models fine-tuned exclusively on biographical subsets.

Table 3 shows that our models perform well on all three biographical test sets, achieving competitive or superior results to existing RE systems despite their more limited training data.

In most settings, our models outperform prior work, with one exception: on the Biographical dataset, our DeBERTa model trails the model with entity tags introduced in Plum et al. (2022) by 4 and 1 points in macro/micro F1. This gap is likely attributable to training data disparity (our 20K vs. their 346K), particularly since both approaches use identical text formatting with entity markers and entity token representations for the classification head. Replication details appear in Appendix C.3.

5.2 Cross-Dataset Results

We now turn to examining the cross-dataset generalisation of our RE systems. Unsurprisingly, we find that performance almost always declines with cross-dataset evaluations. However, models adapted on TACRED-RE exhibit relatively strong generalisation capabilities—the few exceptions of better cross-dataset performance stemming from TACRED-RE models applied to the Biographical dataset–while those adapted on NYT struggle to transfer effectively, likely due to dataset noise.

RE Models Struggle to Generalise across Datasets Cross-dataset evaluations (almost) always perform worse than the comparable intra-dataset experiment: NYT and TACRED-RE show substantial drops of 20-30 points, while Biographical exhibits a smaller decrease of $\sim$10 points for both full and TACRED-RE/Biographical relation overlap (Tables 1, 2). We also observe somewhat different performance trends across model and adaptation approaches from the intra-dataset experiments; while fine-tuning LLaMA on TACRED-RE achieves the best cross-dataset performance on NYT, the best TACRED-RE cross-dataset results are obtained using few-shot ICL with NYT demonstrations (rather than fine-tuning). However, these remain below the zero-shot TACRED-RE baseline.

The cross-dataset experiments on Biographical similarly perform worse than the corresponding intra-dataset experiments in most settings (Table 2); one notable exception is LLaMA prompted with five TACRED-RE examples, which outperforms the intra-dataset few-shot experiments on the TACRED-RE/Biographical label overlap. The best Biographical cross-dataset results are achieved with fine-tuning LLaMA on TACRED-RE, though this still underperforms intra-dataset fine-tuning.

NYT Models Generalise Worse than TACRED-RE Models NYT-adapted models exhibit significantly poorer generalisation than those adapted on TACRED-RE (Tables 1 and 2). For example, fine-tuning LLaMA with TACRED-RE surpasses zero- and cross-dataset ICL on NYT (by $\sim$30 and $\sim$20 points, respectively), while all cross-dataset experiments transferring from NYT to TACRED-RE underperform zero-shot evaluations with no cross-dataset signal. This is also clear from the evaluations on the held-out Biographical dataset, where transferring from TACRED-RE always performs better than NYT (and occasionally outperforms the intra-dataset performance).

This performance gap is unlikely due to domain differences, as both datasets contain newspaper articles (TACRED-RE includes some NYT newspaper content without instance overlap), while Biographical uses Wikipedia. We attribute it to NYT's distant supervision annotations, which introduce noise and limit model robustness. This is likely why LLaMA fine-tuned on NYT and evaluated on Biographical (0.30) underperforms DeBERTa (0.48; Table 2)—the over-parametrised LLaMA exhibits stronger overfitting to NYT noise and generalises poorly to unseen data, a phenomenon also noted by Liu et al. (2022) with corrupted training data.

Effect of Adaptation Strategy on Generalisation While prior work suggests that ICL often generalises more effectively to OOD data than fine-tuning (Awadalla et al., 2022; Song et al., 2023; Si et al., 2023), our results indicate this advantage depends heavily on data quality. With high-quality data like TACRED-RE, fine-tuning consistently achieves the best cross-dataset per-

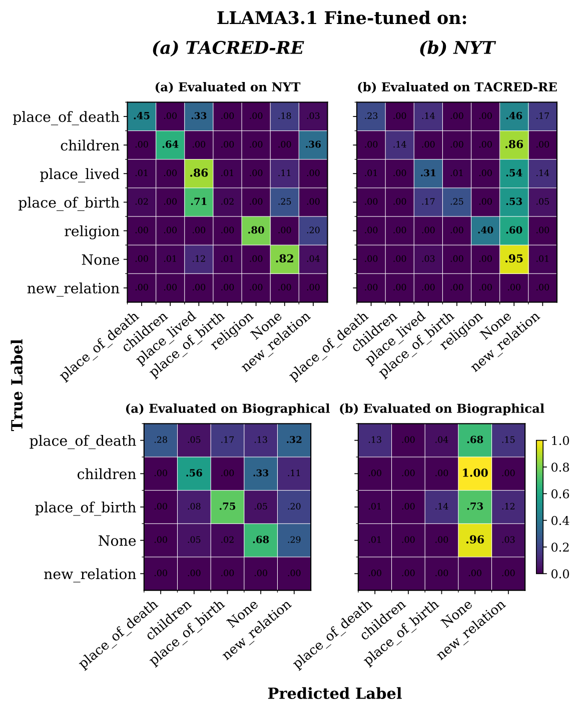

Figure 4: Confusion matrices comparing cross-dataset results for LLaMA fine-tuned on TACRED-RE/NYT.

formance, surpassing few-shot ICL on both NYT and Biographical evaluations. In fact, TACRED-RE adaptations can even perform comparably to intra-dataset ones: LLaMA (0.70) fine-tuned on TACRED-RE achieves similar results to its Biographical intra-dataset performance (0.74) on the TACRED-RE/Biographical label overlap (Table 2).

However, when adaptation data are noisy, as with NYT, few-shot ICL becomes a more effective strategy: few-shot ICL via NYT consistently performs better than fine-tuning on NYT for both TACRED-RE and Biographical evaluations (Tables 1, 2). This is likely because ICL limits the signal from noisy training data, which in turn reduces the overfitting to dataset-specific artefacts and catastrophic forgetting compared to fine-tuning (Tran et al., 2024; Kotha et al., 2024; Li et al., 2025).

6 Analysing RE Generalisation Failure Cases

Given RE benchmarks' numerous relations and class imbalance, we analyse the strongest cross-dataset performing model, fine-tuned LLaMA, beyond aggregated metrics. We examine per-relation performance (Figure 4) and qualitatively analyse 30 random misclassifications from four evaluation settings to identify their likely underlying causes.[6] Through these analyses, we find the following causes of RE generalisation mistakes:

Effect of Noisy Supervision on Generalisation Figure 4 reveals that NYT-adapted models systematically overpredict the 'None' class on both TACRED-RE and Biographical, with manual analysis showing misclassifications stem primarily from NYT's *distant supervision* (Appendix Table 23) rather than vocabulary differences (Appendix Figure 6) or domain shift. This issue is particularly evident in NYT's location-based relations, where reliance on external knowledge leads to conflicting annotations that hinder pattern learning—for instance, "`Henryk Tomaszewski` [...] died on Sunday at his home in `Warsaw`"[7] is labeled as birthplace despite clear evidence of death location. Similar issues arise with Biographical's *semi-supervised* data, where models adapted on cleaner datasets like TACRED-RE fail to replicate ground truth labels that lack textual evidence. Notably, despite higher lexical overlap between Biographical and NYT (Appendix Figure 7), TACRED-RE-adapted models perform better on Biographical, indicating that adaptation data quality matters more than lexical similarity.

Single Relation Constraint & Negative Class The issue extends beyond noisy supervision to fundamental RE task design constraints. Models often detect valid but unlabeled relations (marked as 'new_relation' in Figure 4), revealing limitations of single-relation per sample. This manifests in cases like "`Gross`, who is himself Jewish [...] was sent to `Cuba`"[8], where only 'place_lived' is labeled while 'religion' is omitted due to TACRED-RE's constraint. Also, unclear negative class affects cross-dataset evaluation: while Biographical's 'Other' class is intended to cover undefined relations, our analysis reveals it contains both instances without meaningful relations and those with valid but undefined relations. This explains the high frequency of 'new_relation' predictions for Biographical when using LLaMA fine-tuned on TACRED-RE with their finer-grained relation schema (Figure 4) and highlights the fundamental difficulty in defining boundaries between relation presence or absence. Despite this ambiguity, the class achieves strong

[6] More misclassification patterns in Appendix Table 23.

[7] NYT instance ID: '/m/vinci8/data1/riedel/projects/relation/kb/nyt1/docstore/nyt-2005-2006.backup/1701917.xml.pb'

[8] TACRED-RE instance ID: '098f6f318be29eddb182'

cross-dataset performance when mapped to 'None' (0.78 and 0.72 for LLaMA fine-tuned on TACRED-RE and NYT (Appendix Tables 19 and 20)), further underscoring the importance of distinguishing between 'no_relation' and 'none_of_the_above' cases (Bassignana and Plank, 2022b).

Reliance on External Knowledge even in Manually Curated Datasets Even with high-quality manual annotations, RE often requires external knowledge and complex reasoning capabilities. Our analysis reveals this challenge manifests in two key ways: through implicit relations requiring inference, and through necessary world knowledge for entity interpretation. For example, in "`Gross` [...] was sent to `Cuba` as a spy"[8], the NYT-adapted model predicts 'None' instead of 'place_lived', failing to infer that being sent somewhere as a spy implies residence. While detecting implicit relations is crucial (Geva et al., 2021), ensuring consistent and objective interpretation remains challenging.

Beyond implicit relations, models must also rely on world knowledge for basic entity understanding-as in cases like 'Idaho businesswoman'[9], where identifying entity types requires knowing Idaho as a location. TACRED-RE fine-grained relation schema further demonstrates this issue, where even with world knowledge, distinguishing between relations like city of birth and state/province of birth can be ambiguous (e.g., whether New York refers to the city or state). As Chen et al. (2023) note, even human annotators tend to rely on such prior knowledge despite the lack of rationales, motivating the need for finer-grained word evidence annotation.

Dataset Composition and Coverage Biases Analysis of the most frequent words across NYT and TACRED-RE shows a strong US-centric coverage bias, likely limiting generalisation to non-US contexts (Appendix Table 24). NYT also exhibits topical skews in specific relations, like religion being primarily associated with Islam, potentially biasing model relation representations.

Analysing part-of-speech distributions also reveals distinct patterns across all three datasets (Appendix Table 22). While proper nouns dominate head and tail entities in all datasets (nearly 100% in NYT), TACRED-RE shows more linguistic diversity with 17% of head entities as pronouns and 17% of tail entities as common nouns. Biographical, sourced from Wikipedia, contains a high proportion (26%) of numerical tail entities, primarily dates. These compositional differences, along with TACRED-RE's longer, compound sentences and higher average entity distance (∼12 tokens vs NYT's ∼8 tokens), most likely impact cross-dataset performance; NYT-adapted models struggle with these more complex patterns, which are absent from their training data (Appendix Table 23).

7 Conclusion

This work examines cross-dataset generalisation in language model-based RE systems in biographical settings. We find RE models struggle to generalise even within similar domains, with high intra-dataset performance potentially masking spurious overfitting rather than indicating genuine learning of relational patterns. Our empirical results suggest that data annotation quality can significantly influence transfer; specifically, we observe that datasets with human annotations provide notably better cross-dataset performance than those created through distant supervision. The best adaptation strategy also appears dataset-dependent, with fine-tuning yielding the best cross-dataset performance with TACRED-RE as the adaptation source while few-shot ICL appears to offer advantages with noisier data from our experiments. However, in some cases, a zero-shot baseline surpasses all cross-dataset results, further underscoring the limitations of current RE systems.

Our analysis also reveals several structural issues in all current RE benchmarks: (1) single-relation constraints that ignore other valid relations between entities in text, (2) the lack of a well-defined negative class with challenging samples (e.g., sentences containing commonly used tokens for relations like 'born' or 'died') to enforce deeper semantic understanding beyond pattern matching, and (3) limited diversity in data sources. These issues, compounded by inconsistent relation definitions and limited overlap across datasets, hinder meaningful evaluations of RE generalisation.

These findings thus highlight the need for more transparent evaluation beyond in-distribution testing and aggregated metrics, as limiting evaluation to these may not reflect genuine improvements in capturing relational patterns or account for class imbalance and the large number of relations in RE benchmarks. We see many promising directions for future work, including testing RE robustness on perturbed evaluation sets and applying interpretabil-

[9]TACRED-RE instance ID: '098f6bd9fa786293e49d'

ity methods to better understand how models infer relational knowledge.

Limitations

Our cross-dataset analysis is limited to a particular set of biographical relations but reflects a broader challenge in RE evaluation where datasets, even covering the same domain, typically share a small relation overlap. We also constrain our analysis to single-relation examples: while, real-world scenarios often involve multiple relations per instance (and NYT allows multiple relations), we focused on single-relation setting for fair cross-dataset comparison, as TACRED-RE and Biographical are annotated with single relations. Similarly, we exclusively evaluate relation classification (RC) due to dataset constraints: TACRED-RE and Biographical assume a single relation triple per sentence, unlike real-world text where multiple relations can coexist. By focusing on RC with entity tags as guidance, we aim to minimise the prediction of other potential relations present in a sentence, but not between the specified entities.

The adaptation sets we used contain a large class imbalance due to the underlying distributions of the datasets, even after we perform data rebalancing. While this could be viewed as a limitation, it reflects real-world scenarios where models must adapt with limited training data (Bassignana and Plank, 2022a). Additionally, we evaluate replicated SOTA systems on identical biographical subsets as our models rather than full test sets to ensure controlled comparison with our models. Finally, our work is limited to examining cross-dataset generalisation across three general-purpose datasets due to limited relation type overlap across RE datasets.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. *Preprint*, arXiv:2309.14316.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Varvara Arzt and Allan Hanbury. 2024. Beyond the numbers: Transparency in relation extraction benchmark creation and leaderboards. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 120–130, Miami, Florida, USA. Association for Computational Linguistics.

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Elisa Bassignana and Barbara Plank. 2022a. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elisa Bassignana and Barbara Plank. 2022b. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. 2024. Measuring the robustness of NLP models to domain shifts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 126–154, Miami, Florida, USA. Association for Computational Linguistics.

Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. 2023. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435, Toronto, Canada. Association for Computational Linguistics.

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.

Sefika Efeoglu and Adrian Paschke. 2024. Relation extraction with fine-tuned large language models in retrieval augmented generation frameworks. *Preprint*, arXiv:2406.14745.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–429, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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 *International Conference on Learning Representations*.

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.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*.

Venelin Kovatchev and Matthew Lease. 2024. Benchmark transparency: Measuring the impact of data on evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1536–1551, Mexico City, Mexico. Association for Computational Linguistics.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023a. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore. Association for Computational Linguistics.

Shaoshuai Li, Wenbin Yu, Zijian Chen, and Yangyang Luo. 2023b. A joint entity and relation extraction model based on encoder-decoder. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 3, pages 996–1000.

Xinlong Li, Weijieying Ren, Wei Qin, Lei Wang, Tianxiang Zhao, and Richang Hong. 2025. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. 2022. Robust training under label noise by overparameterization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14153–14172. PMLR.

Siyi Liu, Yang Li, Jiang Li, Shan Yang, and Yunshi Lan. 2024. Unleashing the power of large language models in zero-shot relation extraction via self-prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13147–13161, Miami, Florida, USA. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition. In *AAAI Conference on Artificial Intelligence*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Shiao Meng, Xuming Hu, Aiwei Liu, Fukun Ma, Yawen Yang, Shuang Li, and Lijie Wen. 2024. On the robustness of document-level relation extraction models to entity name variations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16362–16374, Bangkok, Thailand. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.

Alistair Plum, Tharindu Ranasinghe, Spencer Jones, Constantin Orasan, and Ruslan Mitkov. 2022. Biographical semi-supervised relation extraction dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3121–3130, New York, NY, USA. Association for Computing Machinery.

Ji Qi, Chuchun Zhang, Xiaozhi Wang, Kaisheng Zeng, Jifan Yu, Jinxin Liu, Lei Hou, Juanzi Li, and Xu Bin.

2023. Preserving knowledge invariance: Rethinking robustness evaluation of open information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Singapore. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. 2022. On the robustness of offensive language classifiers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7424–7438, Dublin, Ireland. Association for Computational Linguistics.

Pratik Saini, Samiran Pal, Tapas Nayak, and Indrajit Bhattacharya. 2023. 90% f1 score in relation triple extraction: Is it real? In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 1–11, Singapore. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.*, 55(13s).

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online. Association for Computational Linguistics.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Quyen Tran, Nguyen Xuan Thanh, Nguyen Hoang Anh, Nam Le Hai, Trung Le, Linh Van Ngo, and Thien Huu Nguyen. 2024. Preserving generalization of language models in few-shot continual relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13771–13784, Miami, Florida, USA. Association for Computational Linguistics.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks. *Preprint*, arXiv:2407.12994.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194, Valencia, Spain. 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.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–

9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

A Dataset Statistics: Class Distribution

Relation	# of Samples
Other	10,000
birthdate	2914
bplace_name	2845
dplace_name	1138
occupation	1105
deathdate	1011
parent	394
educatedAt	339
child	136
sibling	118
Positive Samples	10,000
Negative Samples	10,000
All	20,000

Table 4: Balanced Biographical Dataset.

Relation	# of Samples
no_relation	14,192
per:title	3805
per:employee_of	2104
per:age	818
per:countries_of_residence	695
per:cities_of_residence	596
per:origin	652
per:stateorprovinces_of_residence	444
per:spouse	463
per:date_of_death	343
per:children	347
per:cause_of_death	318
per:parents	282
per:charges	270
per:other_family	241
per:siblings	238
per:schools_attended	219
per:city_of_death	204
per:religion	145
per:alternate_names	132
per:city_of_birth	107
per:stateorprovince_of_death	100
per:date_of_birth	99
per:stateorprovince_of_birth	77
per:country_of_death	57
per:country_of_birth	45
Positive Samples	12,801
Negative Samples	14,192
All	26,993

Table 5: Balanced TACRED-RE Subset with Biographical Relations (26 relations).

Relation	# of Samples
None	5068
/people/person/nationality	2160
/people/person/place_lived	2016
/people/person/place_of_birth	437
/people/deceased_person/place_of_death	284
/people/person/children	147
/people/person/religion	24
Positive Samples	5068
Negative Samples	5068
All	10,136

Table 6: Balanced NYT Subset with Biographical Relations after removal of instances with multiple labels (7 relations).

B Relation Type Overlap

NYT	Biographical
/people/person/children	child
/people/person/place_of_birth	bplace_name
/people/deceased_person/place_of_death	dplace_name
None	Other

Table 7: NYT/Biographical Relation Overlap.

NYT	TACRED-RE
None	no_relation
/people/person/children	per:children
/people/person/religion	per:religion
/people/person/place_lived	per:stateorprovinces_of_residence per:countries_of_residence per:cities_of_residence
/people/person/place_of_birth	per:stateorprovince_of_birth per:country_of_birth per:city_of_birth
/people/deceased_person/place_of_death	per:stateorprovince_of_death per:country_of_death per:city_of_death

Table 8: NYT/TACRED-RE Relation Overlap.

Biographical	TACRED-RE
bplace_name	per:stateorprovince_of_birth per:country_of_birth per:city_of_birth
birthdate	per:date_of_birth
deathdate	per:date_of_death
parent	per:parents
educatedAt	per:schools_attended
dplace_name	per:stateorprovince_of_death per:country_of_death per:city_of_death
sibling	per:siblings
child	per:children
Other	no_relation

Table 9: Biographical/TACRED-RE Relation Overlap.

Biographical	TACRED-RE	NYT
child	per:children	/people/person/children
bplace_name	per:stateorprovince_of_birth per:country_of_birth per:city_of_birth	/people/person/place_of_birth
dplace_name	per:stateorprovince_of_death per:country_of_death per:city_of_death	/people/deceased_person/ place_of_death
Other	no_relation	None

Table 10: Biographical/TACRED-RE/NYT Overlap.

Ruggiero was awarded the Grand Cross of the Order of the Sacred Treasure by the government of Japan.

Other

Figure 5: Example for 'Other' relation from the Biographical dataset (sample ID 'mS7/1269356').

C Implementation Details

C.1 Data Formatting Details

TACRED-RE's fine-grained relations (e.g., "per:city_of_birth" and "per:country_of_birth") were mapped to broader categories (e.g., "place_of_birth") used in NYT and Biographical datasets, as shown in Tables 8 and 9. Cross-dataset results are reported using NYT label names (Table 17).

C.2 Model Implementation

We fine-tuned DeBERTa-v3-large[10] for 10 epochs employing early stopping. Following prior work using encoder-based models for RE (Baldini Soares et al., 2019; Plum et al., 2022), we extended the `DeBERTa-v3-large` tokeniser with entity marker tokens, namely, <e1> and </e1> for the head entity and <e2> and </e2> for the tail entity, and used the concatenated final hidden states of the entity start tokens, specifically <e1> and <e2>, as input to the classification head. For LLaMA 3.1[11], we used LoRA fine-tuning ($r = 8$) over three epochs, applying it to attention and feedforward modules. Both models were fine-tuned using HuggingFace's Trainer class. For evaluation of LLaMA 3.1, predictions were considered correct only if they matched ground-truth labels exactly (Hendrycks et al., 2021).

For prompting, we used `vanilla` prompting (Li et al., 2023a; Vatsal and Dubey, 2024) and tested several RE-specific prompt designs (Leidinger et al., 2023; Li et al., 2023a; Efeoglu and Paschke, 2024), given LLaMA's sensitivity to prompt formulation (Leidinger et al., 2023). The prompt 1 performed best and was used across all datasets with adapted label sets. Further prompt optimisation techniques were not considered, as they were beyond the scope of this paper. Hyperparameter settings for all experiments are detailed in Table 21.

[10] https://huggingface.co/microsoft/deberta-v3-large

[11] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

```
system_message = {
    "role": "system",
    "content": (
            "You are an intelligent assistant
                specializing in identifying
                relations between entities in a
                sentence. "
            "Question: What is the relation between
                two tagged entities <e1>entity1</e1
                > and <e2>entity2</e2> in the
                following sentence? "
            "Choose one relation from the list: "
            "['/people/person/children', '/people/
                person/nationality', '/people/
                person/place_lived',"
             "'/people/person/place_of_birth', '/
                 people/deceased_person/
                 place_of_death', '/people/person/
                 religion',"
             "'None']. "
            "Rules: Select exactly one relation from
                 the list. If none of the listed
                 relations apply, select 'None'. "
             "Output must strictly follow this
                 format: <relation_type>. Provide
                 no additional text or explanation
                 ."
    )
}
```

Listing 1: System prompt for LLaMA 3.1 8B zero-shot, few-shot, and fine-tuning experiments, shown here with NYT relation inventory.

Due to Biographical's ambiguous 'Other' class (Section 3), we use it only for evaluation in the main paper (Table 2, Section 5) and add supplementary cross-dataset experiments with it in the Appendix Section E.

All experiments with DeBERTa-v3-large were run on a single NVIDIA® TITAN RTX 24GB GPU; all experiments with LLaMA-3.1-8B-Instruct were run on a single NVIDIA® A100 80GB GPU. All experiments were performed with a fixed random seed for reproducibility.

C.3 Replication Study

For our replication of SOTA models for NYT (Orlando et al., 2024), TACRED-RE (Zhou and Chen, 2022), and Biographical (Plum et al., 2022), we either used the provided trained model, implementation, or closely followed the implementation details provided in the corresponding paper. Thus, all models were trained on full training data (i.e., train and validation) (NYT: ∼266k, TACRED-RE: ∼107k, Biographical: ∼346k instances), while our models used only downsampled biographical subsets (NYT: ∼10k, TACRED-RE: ∼26k, Biographical: ∼20k instances), as reflected in Tables 4, 5, and 6. The evaluation was performed on the biographical test sets used in our experiments.

For NYT (Orlando et al., 2024), the retriever-reader `ReLiK Large` model is accessible via HuggingFace[12] ecosystem and it was used for evaluation on the biographical test set. Detailed per-class results of the model by Orlando et al. (2024) on this test set along with results using our models are reflected in Table 12. While Orlando et al. (2024) reports micro F1 of 0.95 on the full test set, ReLiK-Large achieves 0.90 micro F1 on the biographical subset, underperforming our fine-tuned LLaMA model by 4 micro F1 points, as reflected in Table 12.

For TACRED-RE, we replicated the SOTA model by Zhou and Chen (2022), RoBERTa-large with typed entity markers that uses entity-specific hidden representations as input for the classification head, using provided code[13]. Detailed per-class results of the model by Zhou and Chen (2022) on this test set along with results using our models are reflected in Table 14. On the full test set of TACRED-RE (Alt et al., 2020), Zhou and Chen (2022) report micro F1 of 0.83, whereas it achieves a micro F1 of 0.87 on the biographical subset trailing our fine-tuned LLaMA model by 2 micro F1 points.

For Biographical, we replicated the SOTA model Plum et al. (2022), who used BERT Base with entity markers and entity-specific hidden representations for classification—the same representation strategy we employed in our DeBERTa experiments. Since neither the trained BERT Base model nor the train-validation-test split used by Plum et al. (2022) were made publicly available, the results cannot be considered fully replicable. We followed all implementation details provided by Plum et al. (2022) to replicate their model. On the full test set of the `normal` version of Biographical, as designated by the authors, Plum et al. (2022) report macro F1 of 0.76. With our replicated model, we achieve macro F1 of 0.87 and micro F1 of 0.93 on the downsampled test set, outperforming our DeBERTa by 4 macro F1 points and 1 micro F1 point (Table 15). This gap is likely attributable to training data disparity—we heavily downsampled Biographical to 20K instances to match NYT and TACRED-RE subset sizes, whereas the SOTA approach used the full 346K instances.

[12] `https://huggingface.co/sapienzanlp/relik-relation-extraction-nyt-large`

[13] `https://github.com/wzhouad/RE_improved_baseline`

D Vocabulary Overlap between Datasets

Figure 6 depicts vocabulary overlap between NYT and TACRED-RE per overlapping relation. Figure 7 depicts vocabulary overlap between Biographical and TACRED-RE as well as Biographical and NYT per overlapping relation.

E Cross-Dataset Results with Biographical-adapted Models

We extend our cross-dataset experiments with a cross-dataset evaluation of models adapted on Biographical (Table 11). We compare these cross-dataset experiments with NYT- and TACRED-RE-adapted models evaluated on the same four overlapping relations between the three datasets.

As observed with other datasets, Biographical-adapted models also exhibit performance degradation in cross-dataset scenarios, with NYT and TACRED-RE scores showing drops of 14-17 points. Following patterns observed in Section 5, LLaMA fine-tuned on Biographical achieves higher scores than DeBERTa on both NYT and TACRED-RE.

Notably, Biographical-adapted LLaMA achieves the best cross-dataset results on NYT, even outperforming its TACRED-RE counterpart. However, closer manual analysis of errors suggests that this stems from the Biographical dataset's alignment with NYT's annotation methodology: for example, in "`Arne Duncan`, the chief executive of `Chicago` public schools [...]"[14], the Biographical-adapted model predicts 'place_of_birth' (matching NYT's gold label) despite absent textual evidence, while TACRED-RE more appropriately predicts 'no_relation'. Thus, this finding does not necessarily indicate Biographical-adapted models possess better cross-dataset transfer capabilities than models finetuned on the other datasets.

In few-shot ICL, LLaMA with Biographical demonstrations underperforms its TACRED-RE counterpart on NYT by 5 points, while achieving nearly identical results on TACRED-RE as compared to intra-dataset few-shot ICL results. Biographical-adapted models also achieve their best cross-dataset results on TACRED-RE via few-shot ICL adaptation (0.68 ± 0.02), which aligns with our hypothesis in Section 5.2 that ICL mitigates noisy training signals (see Section 6 for more details on Biographical annotation noise).

[14] NYT instance ID: /m/vinci8/data1/riedel/projects/relation /kb/nyt1/docstore/nyt-2005-2006.backup/1653431.xml.pb

Model	Setting	Dataset	Intra-Dataset		Cross-Dataset		
			Shared Labels	Dataset Labels	NYT	TACRED-RE	Biographical
DeBERTa-v3 large 304M	Fine-tuned on	NYT	**0.83**	0.83	–	0.28	0.48
		TACRED-RE	0.73	0.57	0.55	–	0.62
		Biographical	**0.80**	0.80	0.46	0.62	–
LLaMA 3.1 8B	Fine-tuned on	NYT	0.82	0.82	–	0.43	0.30
		TACRED-RE	**0.82**	0.74	0.55	–	**0.69**
		Biographical	0.79	0.79	**0.65**	0.65	–
LLaMA 3.1 8B	Zero-Shot	NYT	0.30	0.30	–	–	–
		TACRED-RE	0.58	0.39	–	–	–
		Biographical	0.24	0.24	–	–	–
LLaMA 3.1 8B	5-Shot	NYT	0.45 ± 0.04	0.45 ± 0.04	–	0.62 ± 0.06	0.48 ± 0.04
		TACRED-RE	0.70 ± 0.13	0.47 ± 0.08	0.37 ± 0.03	–	0.51 ± 0.04
		Biographical	0.53 ± 0.05	0.53 ± 0.05	0.32 ± 0.05	**0.68 ± 0.02**	–

Table 11: Macro F1-scores for intra- and cross-dataset predictions on four overlapping relations. Results show both shared and dataset-specific labels, with models adapted on all biographical relations through fine-tuning or ICL. The best intra- and cross-dataset results on full overlap are highlighted in bold.

F Results

F.1 Intra-Dataset Results

Model	Orlando et al. (2024)			DeBERTa-v3-large 304M			LLaMA-3.1 8B zero-shot			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
place_of_death	0.84	0.97	0.90	0.82	0.70	0.75	0.93	0.82	0.87	0.75	0.09	0.16	0.93	0.82	0.87
children	0.48	0.93	0.63	0.87	0.93	0.90	0.87	0.93	0.90	0.77	0.71	0.74	0.87	0.93	0.90
nationality	0.94	1.00	0.97	0.98	1.00	0.99	0.99	0.99	0.99	0.96	0.10	0.18	0.99	0.99	0.99
place_lived	0.84	0.98	0.90	0.85	0.93	0.89	0.84	0.95	0.89	0.36	0.58	0.44	0.84	0.95	0.89
place_of_birth	0.70	0.94	0.80	0.76	0.54	0.63	0.88	0.46	0.60	0.07	0.06	0.07	0.88	0.46	0.60
religion	1.00	1.00	1.00	1.00	0.60	0.75	1.00	1.00	1.00	0.83	1.00	0.91	1.00	1.00	1.00
None	0.98	0.81	0.89	0.97	0.96	0.97	0.97	0.97	0.97	0.62	0.76	0.68	0.97	0.97	0.97
macro avg	0.87	0.87	0.87	0.89	0.81	0.84	0.92	0.87	0.89	0.62	0.47	0.45	0.92	0.87	**0.89**
micro avg	0.90	0.90	0.90	0.93	0.93	0.93	0.30	0.22	0.25	0.52	0.52	0.52	0.94	0.94	**0.94**
weighted avg	0.90	0.90	0.90	0.93	0.93	0.93	0.94	0.94	0.94	0.63	0.52	0.47	0.94	0.94	**0.94**

Table 12: NYT Results. SOTA model by Orlando et al. (2024) used full NYT train/dev sets, while ours were adapted exclusively on the subset with biographical relations.

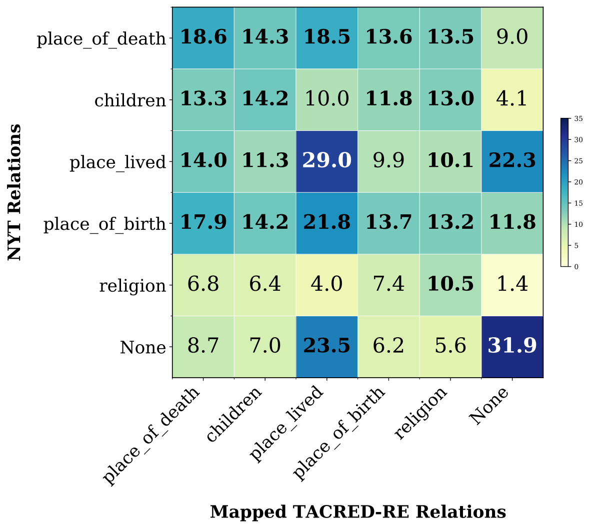

Figure 6: Vocabulary Overlap (%) per overlapping relation between NYT and TACRED-RE.

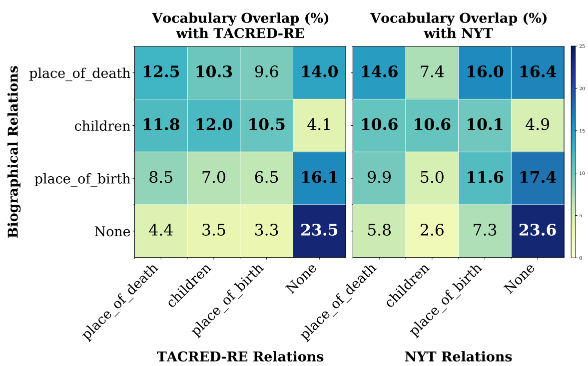

Figure 7: Vocabulary Overlap (%) Between Biographical and TACRED-RE/NYT Relations following lemmatisation and stop word removal using spaCy's `en_core_web_trf`.

Model	DeBERTa-v3-large 304M			LLaMA-3.1 8B zero-shot			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
macro avg	0.67	0.64	0.64	0.49	0.38	0.34	0.37	0.30	0.29	0.76	0.71	**0.73**
micro avg	0.83	0.83	0.83	0.31	0.29	0.30	0.51	0.51	0.51	0.87	0.87	**0.87**
weighted avg	0.83	0.83	0.83	0.68	0.29	0.32	0.59	0.51	0.49	0.87	0.87	**0.87**

Table 13: TACRED Results.

Model	Zhou and Chen (2022)			DeBERTa-v3-large 304M			LLaMA-3.1 8B zero-shot			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
no_relation	0.90	0.95	0.92	0.97	0.80	0.88	0.81	0.01	0.02	0.80	0.61	0.69	0.97	0.87	0.92
per:age	0.96	0.96	0.96	0.92	0.99	0.96	0.97	0.32	0.48	0.93	0.65	0.77	0.95	1.00	0.97
per:cause_of_death	0.92	0.86	0.89	0.77	0.71	0.74	0.46	0.26	0.33	0.53	0.24	0.33	0.65	0.95	0.77
per:charges	0.98	0.88	0.92	0.82	0.97	0.89	0.76	0.30	0.43	0.67	0.43	0.52	0.87	0.99	0.93
per:children	0.93	0.68	0.78	0.65	0.81	0.72	0.23	0.14	0.17	0.19	0.38	0.25	0.96	0.73	0.83
per:cities_of_residence	0.76	0.87	0.81	0.49	0.96	0.65	0.38	0.55	0.45	0.51	0.34	0.41	0.61	0.92	0.73
per:city_of_birth	1.00	0.50	0.67	0.75	0.50	0.60	0.50	0.67	0.57	0.43	0.50	0.46	1.00	0.50	0.67
per:city_of_death	0.60	0.38	0.46	0.64	0.44	0.52	0.33	0.31	0.32	0.28	0.56	0.38	0.43	0.56	0.49
per:countries_of_residence	0.70	0.53	0.61	0.47	0.82	0.60	0.33	0.45	0.38	0.34	0.31	0.32	0.59	0.91	0.71
per:country_of_death	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.56	0.53	0.50	0.44	0.47
per:date_of_birth	1.00	1.00	1.00	0.78	1.00	0.88	0.26	0.86	0.40	0.71	0.71	0.71	0.86	0.86	0.86
per:date_of_death	0.94	0.73	0.82	0.65	0.88	0.74	0.53	0.42	0.47	0.62	0.12	0.21	0.74	0.93	0.82
per:employee_of	0.93	0.81	0.87	0.79	0.88	0.83	0.10	0.97	0.19	0.17	0.73	0.27	0.86	0.89	0.88
per:origin	0.85	0.81	0.83	0.71	0.84	0.77	0.65	0.12	0.21	0.45	0.13	0.20	0.82	0.79	0.80
per:other_family	0.92	0.97	0.95	0.53	0.89	0.67	0.00	0.00	0.00	0.11	0.59	0.19	0.61	0.97	0.75
macro avg	0.84	0.74	0.78	0.67	0.78	0.71	0.43	0.30	0.25	0.53	0.44	0.41	0.77	0.83	**0.78**
micro avg	0.90	0.83	0.87	0.85	0.85	0.85	0.20	0.14	0.17	0.53	0.51	0.52	0.89	0.89	**0.89**
weighted avg	0.90	0.90	0.90	0.88	0.85	0.85	0.70	0.14	0.11	0.72	0.51	0.54	0.91	0.89	**0.89**

Table 14: TACRED-RE Results. SOTA model by Zhou and Chen (2022) used full TACRED-RE train/dev sets, while ours were adapted exclusively on the subset with biographical relations.

Model	Plum et al. (2022)			DeBERTa-v3-large 304M			LLaMA-3.1 8B zero-shot			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Other	0.95	0.94	0.95	0.95	0.93	0.94	0.78	0.29	0.42	0.94	0.54	0.69	0.92	0.93	0.93
birthdate	1.00	1.00	1.00	1.00	1.00	1.00	0.53	0.91	0.67	0.83	0.91	0.87	1.00	1.00	1.00
bplace_name	0.83	0.92	0.87	0.86	0.87	0.87	0.82	0.08	0.14	0.69	0.82	0.75	0.87	0.85	0.86
child	0.67	0.67	0.67	0.38	0.56	0.45	0.02	0.56	0.05	0.00	0.00	0.00	0.57	0.44	0.50
deathdate	1.00	0.98	0.99	1.00	0.95	0.98	0.73	0.82	0.77	0.65	0.85	0.74	0.96	0.99	0.98
dplace_name	0.76	0.67	0.71	0.60	0.72	0.66	0.39	0.25	0.31	0.50	0.55	0.52	0.56	0.81	0.66
educatedAt	0.86	0.97	0.91	0.84	0.84	0.84	0.06	1.00	0.12	0.19	1.00	0.32	0.00	0.00	0.00
occupation	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.41	0.55	0.77	0.94	0.85	1.00	0.98	0.99
parent	0.83	0.77	0.80	0.77	0.82	0.79	0.38	0.59	0.46	0.26	0.84	0.39	1.00	0.68	0.81
sibling	0.85	0.73	0.79	0.83	0.67	0.74	0.10	0.13	0.11	0.42	0.33	0.37	1.00	0.67	0.80
macro avg	0.87	0.86	0.87	0.82	0.84	**0.83**	0.46	0.50	0.36	0.53	0.68	0.55	0.79	0.74	0.75
micro avg	0.93	0.93	0.93	0.92	0.92	**0.92**	0.41	0.40	0.41	0.69	0.69	0.69	0.91	0.91	0.91
weighted avg	0.93	0.93	0.93	0.92	0.92	**0.92**	0.70	0.40	0.43	0.80	0.69	0.71	0.90	0.91	0.90

Table 15: Biographical Results. The model by Plum et al. (2022) used full Biographical train/dev sets, while ours were adapted on its downsampled subset matching NYT and TACRED-RE biographical subset sizes.

F.2 Cross-Dataset Results with Models Fine-tuned on Overlap

Model	Setting	Dataset	Intra-Dataset		Cross-Dataset	
			Shared Labels	Dataset Labels	NYT	TACRED-RE
DeBERTa-v3-large 304M	Fine-tuned on	NYT	0.79	0.79	–	0.27
		TACRED-RE	0.73	0.64	0.49	–
LLaMA 3.1 8B	Fine-tuned on	NYT	0.83	0.83	–	0.45
		TACRED-RE	0.79	0.76	0.58	–
	Shot Setting					
LLaMA 3.1 8B	Zero-Shot	NYT	0.35	0.35	–	–
		TACRED-RE	0.64	0.54	–	–
LLaMA 3.1 8B	5-Shot	NYT	0.46 ± 0.04	0.46 ± 0.04	-	0.50 ± 0.06
		TACRED-RE	0.59 ± 0.03	0.45 ± 0.05	0.44 ± 0.04	-

Table 16: Macro F1-scores for intra- and cross-dataset predictions for six overlapping relations. Results are reported for shared and dataset-specific labels in both fine-tuned and shot settings; only the six overlapping relations are used for adaptation.

F.3 Cross-Dataset Per-Class Results

Model	DeBERTa-v3-large 304M			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1
/people/deceased_person/place_of_death	0.75	0.36	0.49	0.54 ± 0.07	0.39 ± 0.11	0.45 ± 0.08	0.71	0.45	0.56
/people/person/children	0.75	0.64	0.69	0.47 ± 0.05	0.29 ± 0.10	0.34 ± 0.08	0.69	0.64	0.67
/people/person/place_lived	0.62	0.90	0.73	0.27 ± 0.06	0.22 ± 0.15	0.23 ± 0.11	0.63	0.86	0.73
/people/person/place_of_birth	0.20	0.04	0.07	0.12 ± 0.05	0.07 ± 0.04	0.09 ± 0.04	0.17	0.02	0.04
/people/person/religion	1.00	0.20	0.33	0.64 ± 0.03	0.72 ± 0.23	0.66 ± 0.11	1.00	0.80	0.89
None	0.94	0.78	0.85	0.84 ± 0.03	0.46 ± 0.07	0.59 ± 0.06	0.90	0.82	0.86
macro avg	0.71	0.49	0.53	0.48 ± 0.03	0.36 ± 0.04	0.39 ± 0.02	0.68	0.60	**0.62**
micro avg	0.79	0.74	0.77	0.59 ± 0.07	0.37 ± 0.03	0.45 ± 0.04	0.79	0.76	**0.78**
weighted avg	0.80	0.74	0.75	0.62 ± 0.04	0.37 ± 0.03	0.45 ± 0.03	0.78	0.76	**0.76**

Table 17: Adapted on TACRED-RE with all biographical relations present in TACRED-RE. Evaluations on NYT. The labels, although borrowed from NYT dataset, reflect the shared labels between NYT and TACRED-RE. More fine-grained TACRED-RE were mapped to broader shared labels to enable cross-dataset evaluation comparison. The best results on full overlap are highlighted in bold.

Model	DeBERTa-v3-large 304M			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1
/people/deceased_person/place_of_death	0.27	0.17	0.21	0.66 ± 0.11	0.33 ± 0.20	0.40 ± 0.17	0.62	0.23	0.33
/people/person/children	0.40	0.05	0.10	0.49 ± 0.15	0.48 ± 0.20	0.47 ± 0.17	1.00	0.14	0.24
/people/person/place_lived	0.56	0.25	0.34	0.65 ± 0.09	0.38 ± 0.06	0.48 ± 0.05	0.71	0.31	0.43
/people/person/place_of_birth	0.00	0.00	0.00	0.18 ± 0.10	0.88 ± 0.26	0.28 ± 0.11	0.38	0.25	0.30
/people/person/religion	1.00	0.03	0.05	0.86 ± 0.04	0.64 ± 0.17	0.72 ± 0.11	0.94	0.40	0.56
None	0.80	0.91	0.85	0.85 ± 0.04	0.78 ± 0.12	0.81 ± 0.05	0.79	0.95	0.86
macro avg	0.51	0.23	0.26	0.62 ± 0.05	0.58 ± 0.06	**0.52 ± 0.06**	0.74	0.38	0.45
micro avg	0.73	0.69	**0.71**	0.75 ± 0.04	0.66 ± 0.09	**0.70 ± 0.06**	0.74	0.74	**0.74**
weighted avg	0.72	0.69	0.67	0.79 ± 0.01	0.66 ± 0.09	**0.71 ± 0.05**	0.77	0.74	**0.72**

Table 18: Adapted on NYT with all biographical relations present in NYT. Evaluations on TACRED-RE. The labels, although borrowed from NYT dataset, reflect the shared labels between NYT and TACRED-RE. More fine-grained TACRED-RE were mapped to broader shared labels to enable cross-dataset evaluation comparison. The best results on full overlap are highlighted in bold.

Model	DeBERTa-v3-large 304M			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1
None	0.93	0.64	0.75	0.87 ± 0.10	0.49 ± 0.09	0.62 ± 0.04	0.90	0.68	0.78
/people/person/place_of_birth	0.89	0.73	0.80	0.84 ± 0.02	0.70 ± 0.08	0.76 ± 0.05	0.90	0.75	0.82
/people/person/children	0.44	0.44	0.44	0.14 ± 0.14	0.13 ± 0.14	0.14 ± 0.14	0.71	0.56	0.63
/people/deceased_person/place_of_death	0.95	0.33	0.49	0.87 ± 0.05	0.36 ± 0.05	0.51 ± 0.05	0.95	0.38	0.54
macro avg	0.80	0.54	0.62	0.68 ± 0.06	0.42 ± 0.04	0.51 ± 0.04	0.87	0.59	**0.69**
micro avg	0.91	0.63	0.74	0.84 ± 0.05	0.55 ± 0.01	0.66 ± 0.01	0.90	0.67	**0.77**
weighted avg	0.91	0.63	0.74	0.85 ± 0.06	0.55 ± 0.01	0.65 ± 0.01	0.90	0.67	**0.76**

Table 19: Adapted on TACRED-RE with all biographical relations present in TACRED-RE. Evaluations on Biographical with four biographical relations (full overlap between three datasets). The labels, although borrowed from NYT dataset, reflect the shared labels between NYT, TACRED-RE, and Biographical. The best results on full overlap are highlighted in bold.

Model	DeBERTa-v3-large 304M			LLaMA-3.1 8B 5-shot			LLaMA-3.1 8B fine-tuned		
	P	R	F1	P	R	F1	P	R	F1
/people/person/place_of_birth	0.75	0.28	0.41	0.82 ± 0.02	0.71 ± 0.14	0.75 ± 0.09	0.92	0.14	0.25
/people/person/children	0.80	0.44	0.57	0.19 ± 0.06	0.51 ± 0.06	0.27 ± 0.07	0.00	0.00	0.00
/people/deceased_person/place_of_death	0.72	0.12	0.21	0.76 ± 0.09	0.11 ± 0.08	0.19 ± 0.12	0.74	0.13	0.22
None	0.63	0.89	0.73	0.85 ± 0.09	0.62 ± 0.09	0.71 ± 0.06	0.57	0.96	0.72
macro avg	0.73	0.43	0.48	0.65 ± 0.02	0.49 ± 0.04	**0.48 ± 0.04**	0.56	0.31	0.30
micro avg	0.65	0.57	0.60	0.59 ± 0.05	0.59 ± 0.05	**0.59 ± 0.05**	0.55	0.55	0.55
weighted avg	0.69	0.57	0.55	0.82 ± 0.04	0.59 ± 0.05	**0.66 ± 0.04**	0.71	0.55	0.48

Table 20: Adapted on NYT with all biographical relations present in NYT. Evaluations on Biographical with four biographical relations (full overlap between three datasets). The labels, although borrowed from the NYT dataset, reflect the shared labels between NYT, TACRED-RE, and Biographical. The best results on full overlap are highlighted in bold.

Setting	Parameter	DeBERTa-v3-large Fine-tuned	LLaMA 3.1 8B Zero-Shot	LLaMA 3.1 8B Five-Shot	LLaMA 3.1 8B Fine-tuned
Common	# of Epochs	10	–	–	3
	seed	42	42	42	42
	Loss	Cross-Entropy Loss	–	–	Cross-Entropy Loss
	Optimiser	AdamW	–	–	AdamW
	Batch Size	8	–	–	4
	Gradient Accumulation	4	–	–	–
	Early Stopping Patience	2	–	–	2
	Temperature	–	0.1	0.1	–
	Nucleus Sampling	–	0.9	0.9	–
	Lora Settings†	–	–	–	8/32/0.1
	Train-dev-test split	70-20-10	–	–	70-20-10
TACRED-RE	Learning Rate	5×10^{-6} / 5×10^{-5}	–	–	5×10^{-5}
	Max Length	–	–	–	800/384
	Max New Tokens	–	40	256	–
	Cross-Validation	–/5-fold	–	–	–
NYT	Learning Rate	5×10^{-6}	–	–	1×10^{-4}
	Max Length	–	–	–	384
	Max New Tokens	–	256	256	–
	Cross-Validation	–/5-fold	–	–	–
Biographical	Learning Rate	5×10^{-6}	–	–	1×10^{-4}
	Max Length	–	–	–	384
	Max New Tokens	–	40	256	–

Table 21: Hyperparameter settings across datasets. Two values (x/y) indicate *All*/*Overlap* relation experiment settings respectively (if these differ), where *All* indicates experiments with the whole set of biographical relations in each dataset and *Overlap* uses only the intersection. Biographical experiments are performed only with the whole set of biographical relations. †Lora Settings: Rank/Alpha/Dropout.

	TACRED-RE		NYT		Biographical	
POS	Head Entity	Tail Entity	Head Entity	Tail Entity	Head Entity	Tail Entity
PROPN	77.4	55.6	98.4	98.8	87.7	60.5
PRON	16.8	6.2	–	–	0.1	0.2
NOUN	2.4	16.8	0.2	0.4	1.7	5.4
ADJ	1.2	4.5	0.2	–	0.7	0.9
ADP	0.7	1.6	0.2	0.3	0.3	1.1
NUM	0.0	8.5	–	–	6.4	25.5
DET	0.3	1.0	0.2	0.1	0.8	2.3
VERB	0.4	0.7	–	–	0.1	0.1

Table 22: (Top 8) POS Distribution Across TACRED-RE, NYT, and Biographical Test Sets with all Biographical Relations (%). POS tags are obtained with spaCy's transformer-based `en_core_web_trf` model.

G Misclassification Analysis

Issue	**Description**	**Representative Example**	**Misclassifications on**	**Models Affected**
Overpredicting 'None'	Overpredicting 'None' and struggling with even clear relations with cues like 'born' or'died'	"My name is <e1>Ruben</e1> and I am from <e2>Holland</e2>" (GT: *place_lived*, Pred: *None*; TACRED-RE sample ID: '098f6f318bc468878bbb')	TACRED-RE and Biographical	NYT- and Biographical-adapted models
Failure to Capture Implicit Relations	Models struggling to detect implicit relations requiring reasoning	"<e1>Gross</e1> [...] was sent to <e2>Cuba</e2> as a spy" (GT: *place_lived*, Pred: *None*; TACRED-RE sample ID: '098f6f318be29eddb182')	TACRED-RE	NYT-and Biographical-adapted models
Expected world knowledge	For NYT and Biographical this issue is also frequently paired with detatable ground truth labels	"<e1>Augustus</e1> also amassed an impressive art collection and built lavish baroque palaces in Dresden and <e2>Warsaw</e2>" (GT: *dplace_name*, Pred: *None*; Biographical sample ID: 'mS2/247724')	NYT, TACRED-RE, Biographical	models adapted on all 3 datasets affected
Relation Present in Sentence but Not Between Specified Entities	This issue raises concerns about the framing of the RE task itself	"Jan Malte, [...] resident of <e1>Bridgehampton</e1>, died [...] in <e2>San Francisco</e2>" (GT: *None*, Pred: *place_of_death*; NYT article ID: '/m/vinci8/data1/riedel/projects /relation/kb/nyt1/docstore/nyt-2005-2006.backup /1777142.xml.pb')	NYT, TACRED-RE, Biographical	models adapted on all 3 datasets affected
Debatable ground truth (GT) labels	Caused by distantly or semi-supervised manner in which NYT and Biographical were created	"<e1>Ida Freund</e1> was born in <e2>Austria</e2>" (GT: *Other*, Pred: *place_of_birth*; Biographical sample ID: 'mS10/37387826')	TACRED-RE, Biographical	NYT- and Biographical-adapted models
Single-Label Annotation Limitation	Sentences labeled with a single relation may contain additional relations that remain unlabeled	"<e1>Gross</e1>, who is himself Jewish [...] was sent to <e2>Cuba</e2>" (GT: *place_lived*, Pred: *None*; TACRED-RE sample ID: '098f6f318b69f98c850c')	NYT, TACRED-RE, Biographical	models adapted on all 3 datasets affected
Relation missing in annotation schema	Lack of granularity needed to fully capture an individual's biography	"Wen was detained in August and accused of protecting the gang operations masterminded by his sister-in-law, <e1>Xie Caiping</e1>, 46, known as the "godmother" of the <e2>Chinese</e2> underworld (GT: *place_lived*, Pred: *nationality*; TACRED-RE sample ID: '098f637935e6e6d1d093')	NYT, TACRED-RE, Biographical	—
Failure to Capture Relations in Long, Compound Sentences	Models struggling with long-term relational dependencies	"Ecoffey told jurors that he and another federal agent met with <e1>Graham</e1> in April 1994 in Yellowknife, the city in northwest <e2>Canada</e2> where Graham lived at the time" (GT: *place_lived*, Pred: *None*; TACRED-RE sample ID: '098f6f318b3ea9531448')	TACRED-RE	NYT-adapted models

Table 23: Common Misclassification Patterns Across TACRED-RE, NYT, and Biographical.

Relation	NYT	TACRED-RE	Biographical
None	year, york, united, mr, like, states, president, company, work, city	year, national, president, group, include, state, percent, million, american, china	release, contract, announce, song, star, award, series, role, sign, championship
children	father, son, higgins, clark, favre, richard, mary, daughter, carol, daley	son, daughter, grandchild, survive, wife, year, child, include, gude, jr	daughter, son, child, li, father, mother, wife, give, marry, actor
religion	islam, muhammad, prophet, religion, convert, leader, school, al, church, close	jewish, al, islam, shiite, christian, group, muslim, sunni, mohammed, tantawi	–
place_lived	senator, republican, state, year, representative, gov, democrat, city, john, mr	year, state, die, home, york, city, president, live, iran, old	–
place_of_birth	city, year, orleans, chicago, bear, bill, attorney, general, mr, california	bear, grow, family, child, york, year, native, july, old, son	bear, raise, née, grow, family, youth, york, california, city, mother
place_of_death	die, year, home, city, london, los, angeles, mr, yesterday, paris	die, home, hospital, cancer, paris, wednesday, sunday, find, early, dead	die, home, paris, age, california, near, october, london, live, move

Table 24: Top 10 tokens per overlapping relation in NYT, TACRED-RE, and Biographical datasets, following lemmatisation and stop word removal using spaCy's transformer-based `en_core_web_trf` model.