A Variational Approach for Mitigating Entity Bias in Relation Extraction

Samuel Mensah[1] **Elena Kochkina**[1] **Jabez Magomere**[1,2*] **Joy Prakash Sain**[1]
Simerjot Kaur[1] **Charese Smiley**[1]

[1]JP Morgan AI Research [2]University of Oxford
`{name}.{surname}@jpmorgan.com` `jabez.magomere@keble.ox.ac.uk`

Abstract

Mitigating entity bias is a critical challenge in Relation Extraction (RE), where models often rely excessively on entities, resulting in poor generalization. This paper presents a novel approach to address this issue by adapting a Variational Information Bottleneck (VIB) framework. Our method compresses entity-specific information while preserving task-relevant features. It achieves state-of-the-art performance on relation extraction datasets across general, financial, and biomedical domains, in both in-domain (original test sets) and out-of-domain (modified test sets with type-constrained entity replacements) settings. Our approach offers a robust, interpretable, and theoretically grounded methodology.[1]

1 Introduction

Relation Extraction (RE) aims to identify and classify semantic relationships between entities in text. For example, to identify an *"investor"* relationship between the entities *"Microsoft"* and *"OpenAI"* in *"Microsoft invests $10 billion in ChatGPT maker OpenAI"*. By extracting structured relational information from unstructured data, RE serves as a critical enabler for downstream tasks such as knowledge graph construction (Distiawan et al., 2019), question answering (Li et al., 2019), and retrieval-augmented generation (Lewis et al., 2020).

While large language models (LLMs), such as LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI, 2023), have been explored for RE tasks (Wei et al., 2024; Li et al., 2023a; Zhang et al., 2023b), fine-tuned pretrained language models (PLMs) achieve state-of-the-art performance (Gutiérrez et al., 2022; Li et al., 2023b; Zhang et al., 2023a; Wan et al., 2023), particularly in specialized domains like biomedicine (Gutiérrez et al., 2022) and finance (Li et al., 2023b).

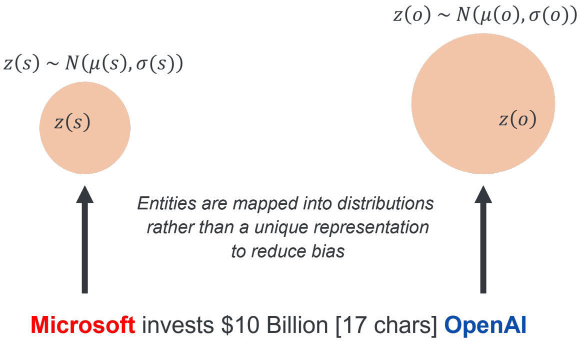

$z(o) \sim N(\mu(o), \sigma(o))$
$z(s) \sim N(\mu(s), \sigma(s))$
$z(s)$
$z(o)$
Entities are mapped into distributions rather than a unique representation to reduce bias
Microsoft invests $10 Billion [17 chars] OpenAI

Figure 1: Microsoft, the subject entity s and OpenAI the object entity o are both mapped into stochastic encodings $z(s)$ and $z(o)$ via VIB. The learned variance of the distribution control the variability to reduce bias.

Despite their success, PLMs often suffer from entity bias (Zhang et al., 2017a), where models overly rely on entity-specific information rather than contextual or relational cues. To mitigate the bias, previous work has explored various solutions, including entity masking (Zhang et al., 2017a, 2018), contrastive pre-training (Peng et al., 2020), counterfactual analysis (Wang et al., 2022, 2023b) and generation (Modarressi et al., 2024).

The current state-of-the-art method is a Structured Causal Model (SCM) (Wang et al., 2023a) that reduces entity bias by constructing a convex hull around an entity's neighbors and using its center to replace the entity embeddings. In contrast to SCM, we draw parallels with variational information bottleneck (VIB) (Alemi et al., 2022) and propose adapting VIB to map entities to a probabilistic distribution $\mathcal{N}(\mu, \sigma)$, where the variance σ^2 explicitly quantifies the model's reliance on entities versus contextual cues. For example, in Fig. 1, the entity *Microsoft* is mapped into a tighter distribution compared to *OpenAI*. The larger distribution for *OpenAI* indicates that the model knows less about it. This helps to debias entities by preventing overconfident assumptions while relying more on

[*] Work done during an internship at JPMorgan AIR
[1] Code available upon request

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 676–684
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

the context. Thus this approach not only mitigates entity bias but also enhances interpretability.

Our contributions are as follows:

- We propose a novel method for relation extraction, a principled, interpretable, variational framework to reduce entity bias in PLMs, specifically RoBERTa-Large (Liu et al., 2019) and LUKE-Large (Yamada et al., 2020).
- We demonstrate the presence of entity bias in both financial and biomedicine relation extraction domains and compare it with the general domain.
- Our method achieves state-of-the-art performance on both general and specialized domains.
- Our approach's interpretability is shown through variance analysis, where low variance reflects reliance on entity information, while high variance indicates greater use of context.

2 Background

Several works address entity bias in RE through diverse techniques. Entity masking (Zhang et al., 2017a, 2018), forces models to focus more on context by replacing entities with generic tokens (e.g., `[subj-person]`). Entity substitution approaches have been explored to test robustness against entity-based knowledge conflicts in question answering (Longpre et al., 2021) and to mitigate factual bias in document-level RE (Modarressi et al., 2024). Peng et al. (2020) propose to mask entity mentions during pre-training to encourage models to focus on context and type information. Wang et al. (2022) perform counterfactual analysis on a causal graph, to guide models to focus on context without losing entity information. The current SOTA method (Wang et al., 2023a) proposes to perturb original entities with neighboring ones to reduce biasing information while preserving context.

Several studies have introduced bias mitigation techniques for black-box large language models (LLMs) that do not require full access to the underlying models (e.g., GPT-4). For instance, Li et al. (2024) demonstrated that LLMs often rely on shortcuts, such as semantic associations or inherent entity biases. Wang et al. (2023a) proposed using an LLM to identify neighboring entities to debias target entities during inference. Similarly, Zhou et al. (2023) showed that employing opinion-based prompts and counterfactual demonstrations can enhance an LLM's contextual faithfulness. In addition, Zhang et al. (2024) introduced a causal prompting framework leveraging front-door adjustment to mitigate biases, while Wu et al. (2024) developed a method for debiasing chain-of-thought reasoning through causal interventions.

Our approach follows the whitebox settings for PLMs. Different from current SOTA method (Wang et al., 2023a), we debias entities through a probabilitic framework, allowing us to estimate the extent to which we use entity versus contextual information.

3 Variational Approach

Our goal is to learn a latent representation Z that preserves the semantic meaning of the input word embeddings X, while minimizing the influence of entity information E. The variational approach (VIB) (Alemi et al., 2022) provides a principled method to achieve this through the mutual information $I(X;Z|E)$, defined as:[2]

$$\begin{aligned} I(X;Z|E) &= \int dx\,dz\,de\,p(x,z,e)\log\frac{p(z|x,e)}{p(z|e)} \\ &\leq \int dx\,dz\,de\,p(x,z,e)\log\frac{p(z|x,e)}{r(z|e)} \end{aligned}$$

where $r(z|e)$ is a variational approximation to $p(z|e)$, inducing an upper bound on $I(X;Z|E)$. We can now interpret the upper bound as a KL divergence (Kullback and Leibler, 1951), so that the upper bound of $I(X;Z|E)$ becomes the expected KL divergence given by:

$$\begin{aligned} I(X;Z|E) \leq& \mathbb{E}_{p(x,z,e)}[\mathrm{KL}(p(z|x,e)||r(z|e))] \\ =& L_{\mathrm{VIB}} \end{aligned}$$

This bound forms the basis of the VIB loss L_{VIB}, where the bottleneck is enforced by minimizing the KL divergence, restricting $p(z|x,e)$ to stay close to $r(z|e)$. In practice, $p(z|x,e)$ is modelled as a Gaussian distribution $\mathcal{N}(\mu,\sigma)$, where the mean μ and standard deviation σ are parametrized by single-layer perceptrons (SLP) while $r(z|e)$ is modelled as a standard normal distribution $\mathcal{N}(0,I)$.

Compressing Entity Representations. Since the goal is to limit entity-specific information in Z while preserving the semantic meaning in X, VIB is applied selectively to entities using a binary entity mask M that identifies the position of entity tokens. To enable efficient and differentiable optimization, we sample z be from $\mathcal{N}(\mu,\sigma)$ using the

[2] In this work, X, Z, E, H are random variables, and x, z, e, h are instances of these random variables.

reparameterization trick (Kingma, 2013), that is, $z = \mu + \epsilon \cdot \sigma$ and $\epsilon \sim \mathcal{N}(0, 1)$. σ helps control how much information about the input is retained in z. Smaller σ values lead to tighter, more deterministic representations, while larger σ values encourage more stochastic exploration, which is essential for mitigating entity bias and learning better context-based representations.[3]

To encourage retaining the context of non-entity tokens and reducing entity-specific details, we selectively blend the original embeddings x with z using M and a blending factor β. Non-entity tokens ($M = 0$) retain their original embeddings, while entity tokens ($M = 1$) are represented as a weighted combination of x and z.

$$x' = x \cdot (1 - M) + x \cdot M \cdot (1 - \beta) + z \cdot M \cdot \beta$$

This formulation ensures the final embeddings x' reduce entity-specific details while preserving task-relevant features.

Classification and Training Objective. Given x', we apply a pretrained PLM encoder to obtain contextualized embeddings $h = \text{PLM}(x')$. We then extract and concatenate the representations of special tags $[h_s]$ and $[h_o]$ which mark the subject and object entities, and feed this joint representation $[h_s; h_o]$ to a fully connected layer and softmax for classification. The total loss combines the cross-entropy (CE) loss L_{CE} for relation classification and the VIB loss L_{VIB}.

$$\mathcal{L} = L_{\text{CE}} + \alpha L_{\text{VIB}}$$

where α is an adaptive weight, computed as a ratio between the CE and VIB loss. This ensures a balanced contribution of both loss terms.

4 Experiments

We conduct experiments on three large relation extraction datasets: TACRED (Zhang et al., 2017b) (general domain), REFinD (Kaur et al., 2023) (financial domain) and BioRED (Luo et al., 2022) (biomedical domain).[4] Evaluation follows previous work (Wang et al., 2023a) using `entity_marker_punctuation` (Zhou and Chen, 2022) to mark entities, and Micro-F1 as the metric on both in-domain (ID) and out-of-domain (OOD) test sets. Here, in-domain refers to data where entities align with those in the train set, allowing for overlapping entity mentions. Meanwhile, out-of-domain data where entities are replaced to eliminate overlap with the train set. We generate OOD test sets following the approach by Wang et al. (2023c), using entities from Wikepedia dumps.[5] We experiment with LUKE-Large (Yamada et al., 2020) and RoBERTa-Large (Liu et al., 2019) as PLM backbones.[6]

5 Main Results

The results in Table 1 highlight the performance of LUKE-Large and RoBERTa-Large backbone models. We find that traditional methods like Entity Masking (Zhang et al., 2017a) and Entity Substitution (Longpre et al., 2021) show underperformance, highlighting the importance of retaining some information about the original entity. Both SCM and VIB retain some information about the original entity, leading to their stronger performance compared to early methods.

For LUKE-Large in ID settings, VIB achieves Micro-F1 scores: 70.4% on TACRED, 75.4% on REFinD and 61.2% on BioRED, outperforming SCM by about 1.8%, 0.9% and 2.9%, respectively. Under entity-replaced conditions (OOD), VIB consistently shows competitive or better performance compared to SCM. Specifically, VIB achieves Micro F1 scores: 66.5% on TACRED, 74.8% on REFinD, and 58.7% on BioRED, outperforming SCM by about 1.7%, 1% and 5.3%, respectively. For the RoBERTa-Large backbone, SCM and VIB achieve comparable performance, with SCM slightly outperforming VIB in OOD TACRED (67.5% vs. 67.2%) and OOD BioRED (52.5% vs. 52.5%). VIB has an edge in ID and OOD REFinD, and ID for both TACRED and BioRED.

Comparing the results of VIB across the backbones, LUKE's knowledge-based entity representations appear to amplify VIB's ability to effectively balance the utilization of entities and context. This highlights VIB's strength in leveraging entity-rich backbones for improved generalization, especially in domain-specific datasets like REFinD.

[3] To ensure $\sigma > 0$, we apply a softplus activation function on the raw ouput σ' of the SLP, i.e., $\sigma = \text{softplus}(\sigma')$

[4] Dataset statistics can be obtained from the original papers.

[5] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[6] We adapted the source code provided by Wang et al. (2023a) under the Apache-2.0 license. Source code available at `https://github.com/luka-group/Causal-View-of-Entity-Bias/` Hyperparameters were tuned for β in $\{0.1, 0.2 \ldots, 1\}$, Learning rates in {1e-5, 1e-4, 1e-3}, using Adam as the optimizer. Best hyperparameter $\beta = 0.5$ with learning rate $lr =$ 1e-3. All experiments were conducted on an AWS g5.24xlarge.

	TACRED		REFinD		BioRED	
Method	**ID**	**OOD**	**ID**	**OOD**	**ID**	**OOD**
LUKE-Large (Yamada et al., 2020)	71.1$_{\pm 0.3}$	63.8$_{\pm 1.5}$	75.0$_{\pm 0.2}$	73.4$_{\pm 0.3}$	56.9$_{\pm 0.7}$	51.8$_{\pm 1.2}$
w/ Ent. Mask (Zhang et al., 2017a)	63.6$_{\pm 0.1}$	61.7$_{\pm 1.2}$	71.4$_{\pm 0.4}$	71.4$_{\pm 0.9}$	53.2$_{\pm 0.6}$	40.2$_{\pm 1.1}$
w/ Ent. Substitution (Longpre et al., 2021)	66.6$_{\pm 0.3}$	60.3$_{\pm 0.6}$	74.3$_{\pm 0.5}$	72.9$_{\pm 1.2}$	56.2$_{\pm 0.4}$	46.7$_{\pm 1.0}$
w/ SCM (Wang et al., 2023a)	68.6$_{\pm 0.2}$	64.8$_{\pm 0.4}$	74.5$_{\pm 0.6}$	73.8$_{\pm 0.6}$	58.3$_{\pm 1.7}$	53.4$_{\pm 1.7}$
w/ VIB ($\beta = 0.5$)	**70.4$_{\pm 0.4}$**	**66.5$_{\pm 0.4}$**	**75.4$_{\pm 0.2}$**	**74.8$_{\pm 1.5}$**	**61.2$_{\pm 0.8}$**	**58.7$_{\pm 0.6}$**

RoBERTa-Large (Liu et al., 2019)	70.8$_{\pm 0.1}$	61.5$_{\pm 0.9}$	75.1$_{\pm 0.2}$	72.7$_{\pm 0.1}$	57.7$_{\pm 1.9}$	47.9$_{\pm 2.3}$
w/ Entity Mask (Zhang et al., 2017a)	62.0$_{\pm 0.7}$	60.6$_{\pm 0.8}$	70.4$_{\pm 1.5}$	71.2$_{\pm 1.0}$	55.2$_{\pm 1.9}$	45.7$_{\pm 1.1}$
w/ Entity Substitution (Longpre et al., 2021)	67.1$_{\pm 0.3}$	61.2$_{\pm 1.1}$	73.5$_{\pm 0.9}$	71.9$_{\pm 0.2}$	56.9$_{\pm 1.1}$	46.8$_{\pm 3.7}$
w/ Structured Causal Model (Wang et al., 2023a)	70.5$_{\pm 0.6}$	**67.5$_{\pm 0.3}$**	74.9$_{\pm 1.0}$	73.7$_{\pm 1.1}$	57.3$_{\pm 3.3}$	**52.5$_{\pm 3.3}$**
w/ VIB ($\beta = 0.5$)	**70.7$_{\pm 0.3}$**	67.2$_{\pm 0.3}$	**75.4$_{\pm 0.1}$**	**74.4$_{\pm 0.2}$**	**63.0$_{\pm 2.3}$**	52.5$_{\pm 3.6}$

Table 1: **Main Results**: Micro-F1 scores of compared methods with the RoBERTa-Large and LUKE-Large backbones on the TACRED, REFinD, and BioRED datasets, evaluated in both in-domain and out-of-domain settings. Results are averaged over 3 runs, with standard deviations reported.

	Var. Bin	**Prop.**	**Dominant Relations (Correct Predictions / Total Gold)**
LUKE-Large w/ VIB	**In-Domain**		
	0.0-0.1	4.6%	pers:title:title (43/71), org:gpe:headquartered_in (11/11), org:money:revenue_of (9/10)
	0.1-0.2	85.8%	pers:title:title (503/600), org:gpe:operations_in (419/536), pers:org:employee_of (329/352)
	0.2-0.3	9.6%	org:date:formed_on (73/78), org:gpe:operations_in (55/60), org:org:subsidiary_of (4/6)
	0.3-0.4	0.1%	org:date:formed_on (3/3)
	Out-of-Domain		
	0.0-0.1	13.2%	pers:title:title (59/107), pers:org:employee_of (49/89), org:gpe:operations_in (19/30)
	0.1-0.2	82.8%	pers:title:title (463/564), org:gpe:operations_in (433/550), pers:org:employee_of (259/283)
	0.2-0.3	3.8%	org:date:formed_on (66/68), org:gpe:operations_in (22/25), pers:org:employee_of (2/2)
	0.3-0.4	0.2%	org:date:formed_on (8/8)
RoBERTa-Large w/ VIB	**In-Domain**		
	0.0-0.1	13.9%	pers:title:title (476/479), pers:org:employee_of (31/40), org:gpe:operations_in (12/14)
	0.1-0.2	43.4%	pers:org:employee_of (287/297), org:gpe:operations_in (280/332), pers:title:title (75/181)
	0.2-0.3	35.0%	org:gpe:operations_in (157/222), pers:org:employee_of (34/37), org:org:agreement_with (24/74)
	0.3-0.4	6.4%	org:date:formed_on (57/60), org:gpe:operations_in (25/34), org:org:subsidiary_of (5/10)
	0.4-0.5	1.0%	org:date:formed_on (15/15), org:gpe:headquartered_in (1/1), org:gpe:operations_in (1/2)
	0.5-0.6	0.1%	org:date:formed_on (1/1), org:gpe:operations_in (1/1), org:org:subsidiary_of (1/1)
	Out-of-Domain		
	0.0-0.1	15.0%	pers:title:title (424/442), pers:org:employee_of (49/75), org:gpe:operations_in (33/40)
	0.1-0.2	56.3%	org:gpe:operations_in (299/397), pers:org:employee_of (254/277), pers:title:title (102/221)
	0.2-0.3	24.8%	org:gpe:operations_in (114/151), pers:org:employee_of (19/21), org:date:formed_on (16/20)
	0.3-0.4	3.2%	org:date:formed_on (53/55), org:gpe:operations_in (12/17), org:money:revenue_of (1/2)
	0.4-0.5	0.7%	org:date:formed_on (12/12), no_relation (0/0), org:org:shares_of (0/4)
	0.5-0.6	0.0%	org:date:formed_on (1/1)

Table 2: Variance analysis of REFinD ID and OOD test sets, categorized by variance bins (Var. Bin). The table highlights the proportion of samples (Prop.) within each bin and identifies dominant relations based on correct predictions versus total gold labels. Results are presented for both LUKE- and RoBERTa-Large w/VIB models.

5.1 Variance Analysis

During inference, σ is predicted by the learned SLP that parameterizes the distribution $\mathcal{N}(\mu, \sigma)$, with variance computed as σ^2. Variance σ^2 reflects the model's reliance on entities versus context, where low variance indicates stronger reliance on entities and high variance reflects greater use of contextual cues. We analyze Micro-F1 performance and data distribution across in-domain and out-of-domain settings on REFinD to understand this balance in the financial domain.

Micro-F1 Distribution Across Variance Figure 2 illustrates VIB's Micro-F1 scores for in-domain and out-of-domain datasets. Samples are grouped by ascending mean variance, where subsets with lower percentages (e.g., 10%) corre-

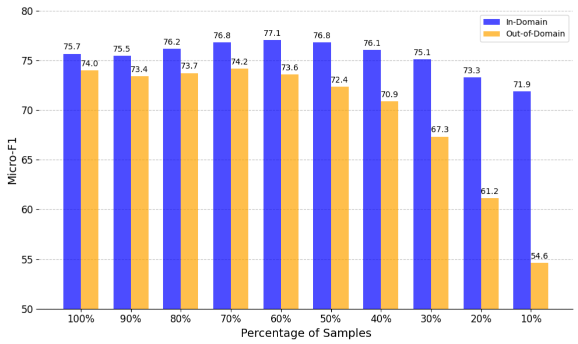


Figure 2: Micro-F1 scores across sample subsets (sorted by variance) for ID and OOD on REFinD.

spond to the highest variance. For in-domain data, the scores remain stable across subsets (75.7% to 71.9%), indicating that VIB effectively mitigates entity bias while leveraging both entity and contextual information. In out-of-domain data, however, the sharper decline in Micro-F1 scores (74.0% to 54.6%) indicates that while VIB reduces over-reliance on entities by mapping entities into distributions of high variances, contextual signals may not always provide strong predictive cues. This underscores the importance of robust context-entity interaction for generalization.

Data Distribution Across Variance In Table 2, we group instances into bins based on mean variance - average variance across entity tokens. Each bin reports the proportion of samples, and the dominant relations, highlighting the most accurately predicted relations. The results highlight VIB's ability to balance entity and contextual information while adapting to varying data distributions in in-domain and out-of-domain settings. For in-domain data, most samples (85.8%) fall into the 0.1–0.2 variance bin, dominated by relations like `pers:title:title` and `org:gpe:operations_in`, with smaller proportions in lower (0.0–0.1, 4.6%) and moderate (0.2–0.3, 9.6%) variance bins. This concentrated distribution explains the stability of Micro-F1 scores observed in the bar graph, as removing high-variance samples has minimal impact on performance. In contrast, out-of-domain data shifts more samples into the lowest variance bin (0.0–0.1, 13.2%), reflecting stronger reliance on entities; however, entity replacements disrupt predictive utility, leading to lower performance in the bar graph. Additionally, sparsely populated high-variance bins (e.g., 0.2–0.3, 3.8%) correspond to sharp performance drops (e.g., 30%–10%), highlighting challenges with relying on context alone and the need for stronger contextual adaptability in out-of-domain scenarios (see examples in A.2).

6 Conclusions

We proposed a novel robust, interpretable, and theoretically grounded method for mitigating entity bias in relation extraction. We evaluated this approach on general and domain-specific datasets, TACRED, REFinD and BioRED, and showed that it achieves state-of-the-art results on each.

Limitations

In this study, we focus on the application of PLMs, acknowledging that our work does not easily extend to LLMs, which have become increasingly significant in recent advancements. Future research will aim to expand our VIB method to encompass generative models such as T5 and LLMs, potentially uncovering new insights and applications.

Furthermore, our research is conducted solely in the English language, which may limit its relevance to non-English contexts. Language-specific challenges and nuances could influence the performance of PLMs, and future studies should consider incorporating multiple languages to enhance the generalizability and impact of our findings.

Acknowledgments

We thank Arturo Oncevay, Toyin Aguda, Keshav Ramani and Xiaomo Liu for insightful discussions.

Disclaimer This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2022. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3054–3063.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? 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 7668–7681.

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.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023b. Are chatgpt and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*, pages 408–422. Association for Computational Linguistics.

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

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*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Ali Modarressi, Abdullatif Köksal, and Hinrich Schütze. 2024. Consistent document-level relation extraction via counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11501–11507.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5679–5688.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023a. A causal view of entity bias in (large) language models. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pages 15173–15184.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob Gardner, Dan Roth, and Muhao Chen. 2023b. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 541–553.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081.

Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. 2023c. How fragile is relation extraction under entity replacements? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 414–423.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2024. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *arXiv preprint arXiv:2403.02738*.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023a. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023b. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017a. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 35–45. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017b. Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*.

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.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

A Appendix

A.1 Performance Across Different Relations

REFinD Table 3 shows a detailed comparison of SCM and VIB performance across multiple relations in REFinD, evaluated in both ID and OOD settings. VIB outperforms SCM in many cases, particularly for relations like `org:org:agreement_with` and `pers:org:member_of`, where contextual cues are critical, demonstrating its ability to effectively balance entity and context. For high-frequency relations such as `no_relation` and `org:gpe:operations_in`, VIB also shows slight but consistent improvements. However, SCM often matches or outperforms VIB on relations like `org:money:loss_of` and `org:gpe:headquartered_in`. While both methods achieve comparable overall performance, VIB provides the added advantage of quantifying the reliance on entity versus context information, making it more insightful for understanding model behavior.

TACRED Table 4 compares SCM and VIB performance on TACRED in both ID and OOD settings, across various relation types. While SCM excels in high-frequency relations like `no_relation` and `per:title`, VIB consistently outperforms SCM in challenging relations such as `per:employee_of` and `org:top_members/employees`, particularly under entity replacement in OOD. VIB's strength lies in leveraging contextual information effectively when entity reliability diminishes. For rare relations like `per:city_of_death` and `org:dissolved`, VIB often surpasses SCM, though both methods struggle with extremely sparse relations like `org:shareholders`. Overall, VIB demonstrates strong generalization under entity replacement, making it a robust approach for mitigating entity bias while maintaining competitive performance across diverse relations.

	REFinD-ID		**REFinD-OOD**	
Relation	**SCM**	**VIB**	**SCM**	**VIB**
no_relation	85.01	**86.91**	85.01	**86.91**
pers:title:title	77.79	77.79	77.79	77.79
org:gpe:operations	76.20	**78.35**	76.20	**78.35**
pers:org:employee	**93.05**	82.89	**93.05**	82.89
org:org:agrmnt	26.95	**35.46**	26.95	**35.46**
org:date:formed	86.32	**87.37**	86.32	**87.37**
pers:org:member	9.47	**15.79**	9.47	**15.79**
org:org:subsidiary	38.55	**49.40**	38.55	**49.40**
org:org:shares	**27.87**	6.56	**27.87**	6.56
org:money:revenue	74.47	**82.98**	74.47	**82.98**
org:money:loss	**96.77**	90.32	**96.77**	90.32
org:gpe:headqtr	79.31	79.31	79.31	79.31
org:date:acquired	**54.17**	37.50	**54.17**	37.50
pers:org:founder	**40.00**	30.00	**40.00**	30.00
org:gpe:formed	23.53	**64.71**	23.53	**64.71**
pers:univ:employee	58.33	**66.67**	58.33	**66.67**
org:org:acquired	**18.18**	0.00	**18.18**	0.00
pers:gov:member	**12.50**	0.00	**12.50**	0.00
pers:univ:attended	85.71	85.71	85.71	85.71
org:money:profit	80.00	80.00	80.00	80.00
pers:univ:member	**60.00**	40.00	**60.00**	40.00
org:money:cost	**75.00**	0.00	**75.00**	0.00

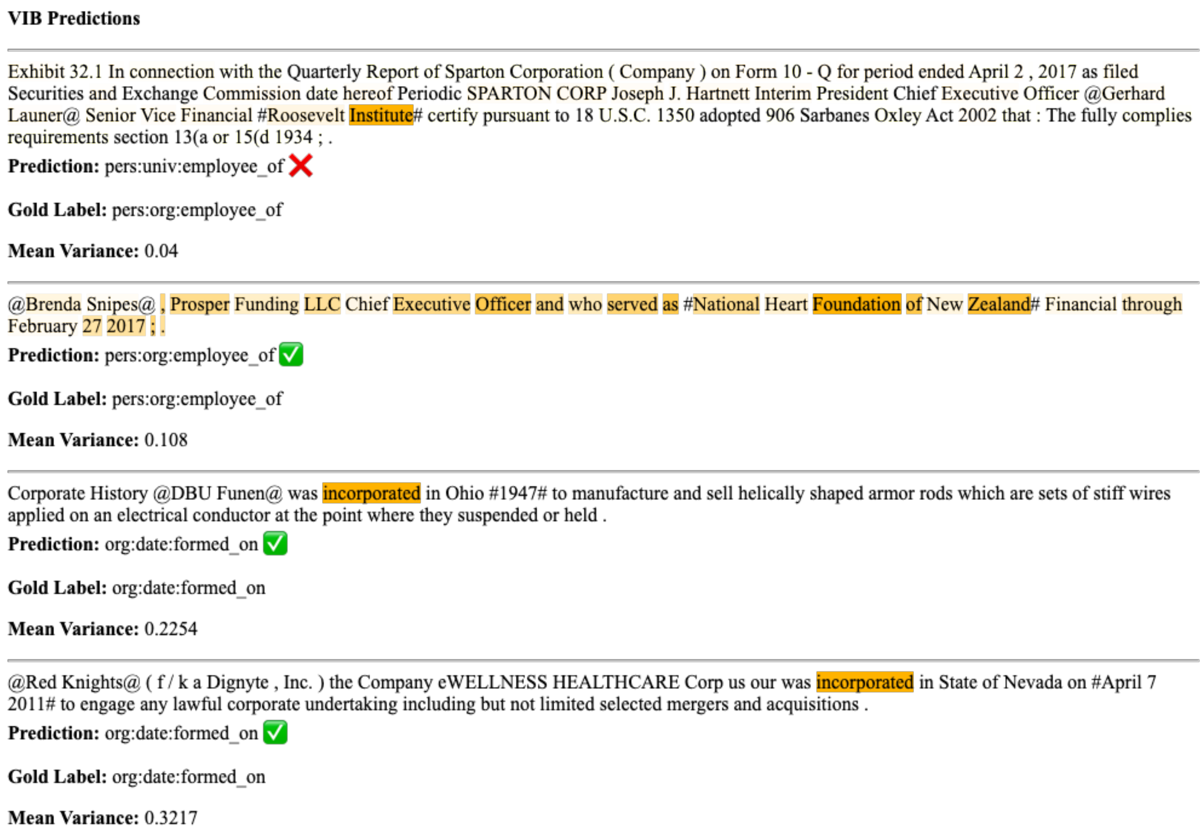
Table 3: **LUKE-Large** Performance of SCM and VIB models on various relations within the **REFinD dataset**, evaluated in both in-domain and out-of-domain settings. Relations are ordered by their frequency in the dataset, with the most frequent at the top (i.e., no_relation). Bolded values indicate the best performance for a relation in either ID or OOD settings.

A.2 Mask Experiment Vrs Variance Analysis

The mask experiment, as proposed by (Sun et al., 2019), evaluates token-level relevance by measuring the contribution of individual tokens to the final relation representation, effectively visualizing the model's attention patterns. Complementary to this, variance analysis provides a quantitative measure of reliance on entity or contextual information, where low variance indicates strong reliance on entity tokens and high variance reflects greater dependence on contextual cues. As demonstrated in Figure 3, entities in relations like `pers:org:employee_of` exhibit low mean vari-

	TACRED-ID		**TACRED-OOD**	
Relation	**SCM**	**VIB**	**SCM**	**VIB**
no_relation	**93.71**	92.61	**93.60**	92.26
per:title	**94.20**	89.60	**93.90**	92.20
org:top_memb/empl	80.64	**82.66**	74.28	**76.30**
per:employee	53.41	**71.59**	38.64	**55.30**
org:alt_names	91.08	**91.55**	**80.75**	78.87
per:age	**96.00**	95.50	**97.63**	97.63
per:cities_res	48.68	**57.14**	40.31	**54.26**
per:countries_res	5.41	**43.92**	5.77	**48.08**
per:origin	**65.91**	48.48	**65.79**	54.39
org:country_of_hq	37.04	**58.33**	30.84	**38.32**
per:charges	88.35	**93.20**	**91.25**	91.25
per:parents	**79.55**	79.55	**79.52**	78.31
org:city_hq	**73.17**	67.07	**67.07**	65.85
per:state_res	49.38	**59.26**	45.61	**49.12**
org:founded_by	**86.76**	85.29	**82.35**	80.88
per:spouse	50.00	**78.79**	46.77	**75.81**
org:parents	35.48	**43.55**	20.97	**24.19**
per:other_fam	**51.67**	51.67	**58.82**	56.86
per:siblings	67.27	**76.36**	72.55	**76.47**
per:date_death	18.52	**44.44**	23.08	**58.97**
per:cause_death	40.38	**48.08**	42.50	**50.00**
org:state_hq	**76.47**	74.51	**74.51**	72.55
per:religion	**42.55**	40.43	**51.61**	48.39
org:subsidiaries	40.91	**45.45**	**36.36**	34.09
org:founded	**83.78**	83.78	**86.11**	80.56
per:children	40.54	**43.24**	40.63	**50.00**
org:members	0.00	0.00	0.00	**3.23**
per:sch_attended	60.00	**83.33**	46.67	**70.00**
per:city_death	0.00	**39.29**	0.00	**52.17**
org:website	**84.62**	84.62	41.67	**79.17**
org:num_empl/memb	**68.42**	57.89	**70.59**	54.90
org:member	0.00	0.00	0.00	0.00
per:state_death	0.00	**42.86**	0.00	**42.42**
org:shareholders	0.00	0.00	0.00	0.00
per:alt_names	**27.27**	18.18	**21.21**	6.06
org:pol/relig_affil	**40.00**	40.00	**48.15**	44.44
per:date_birth	**77.78**	77.78	**75.00**	75.00
per:country_death	0.00	0.00	0.00	0.00
per:state_birth	25.00	**50.00**	27.78	**50.00**
per:country_birth	0.00	**20.00**	0.00	**26.67**
per:city_birth	20.00	**40.00**	20.00	**40.00**
org:dissolved	0.00	**50.00**	0.00	**50.00**

Table 4: **Luke-Large** Performance of models SCM and VIB on various relations within **TACRED dataset**, evaluated in both in-domain and out-of-domain settings. Relations are ordered by their frequency in the dataset, with the most frequent at the top (i.e., no_relation). Bolded values indicate the best performance for a relation in either ID or OOD settings.



VIB Predictions

Exhibit 32.1 In connection with the Quarterly Report of Sparton Corporation (Company) on Form 10 - Q for period ended April 2 , 2017 as filed Securities and Exchange Commission date hereof Periodic SPARTON CORP Joseph J. Hartnett Interim President Chief Executive Officer @Gerhard Launer@ Senior Vice Financial #Roosevelt Institute# certify pursuant to 18 U.S.C. 1350 adopted 906 Sarbanes Oxley Act 2002 that : The fully complies requirements section 13(a or 15(d 1934 ; .

Prediction: pers:univ:employee_of ❌

Gold Label: pers:org:employee_of

Mean Variance: 0.04

@Brenda Snipes@ , Prosper Funding LLC Chief Executive Officer and who served as #National Heart Foundation of New Zealand# Financial through February 27 2017 ; .

Prediction: pers:org:employee_of ✅

Gold Label: pers:org:employee_of

Mean Variance: 0.108

Corporate History @DBU Funen@ was incorporated in Ohio #1947# to manufacture and sell helically shaped armor rods which are sets of stiff wires applied on an electrical conductor at the point where they suspended or held .

Prediction: org:date:formed_on ✅

Gold Label: org:date:formed_on

Mean Variance: 0.2254

@Red Knights@ (f / k a Dignyte , Inc.) the Company eWELLNESS HEALTHCARE Corp us our was incorporated in State of Nevada on #April 7 2011# to engage any lawful corporate undertaking including but not limited selected mergers and acquisitions .

Prediction: org:date:formed_on ✅

Gold Label: org:date:formed_on

Mean Variance: 0.3217

Figure 3: Visualization of attention and prediction results for VIB. Subject and object entities are marked with @ and # respectively. Low mean variance indicates strong reliance on entity tokens, while high mean variance reflects a shift toward contextual cues. Highlighted tokens show entity-focused attention, visualized using the mask method (Sun et al., 2019)

ance (e.g., 0.108), aligning with the mask experiment's focus on entity tokens such as "@Brenda Snipes@" and "#Prosper Funding LLC#". Conversely, for relations like `org:date:formed_on`, higher mean variance (e.g., 0.2254 and 0.3217) suggests greater reliance on context, consistent with the mask experiment, where contextual words like "incorporated" contribute prominently. This alignment between variance analysis and the mask experiment highlights the model's ability to balance entity and contextual cues, reinforcing interpretability.