Counterfactual-based Cognitive Alignment In-Context Learning for Relation Extraction

Qibin Li[1*], **Shengyuan Bai**[1,2*†], **Nai Zhou**[3], **Nianmin Yao**[1†]

[1]School of Computer Science and Technology, Dalian University of Technology
[2]International Digital Economy Academy (IDEA)
[3]Quan Cheng Laboratory
liqibin@mail.dlut.edu.cn, jerry.sy.bai@gmail.com, zhounai1992@outlook.com, lucos@dlut.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable In-Context learning (ICL) capabilities for relation extraction (RE). While ICL has shown promise in RE tasks, current approaches face challenges in example selection and utilization. These challenges stem from the misalignment between example selection methods and LLMs' inherent cognitive processing mechanisms, particularly in pattern recognition and relational reasoning. To address these limitations, we propose Counterfactual Cognitive Alignment (CCA), a novel framework that systematically enhances ICL performance in RE by aligning example selection with cognitive principles underlying human relational reasoning. The framework incorporates a cognitive-inspired counterfactual generation mechanism that creates semantically diverse yet relationally coherent examples, mirroring human "what-if" reasoning processes. Additionally, it employs a cognitive alignment approach that integrates structural identification features with semantic understanding to better align with LLMs cognitive processing patterns. Extensive experiments across multiple RE benchmarks reveal the effectiveness of our cognitive alignment approach through the synergistic integration of counterfactual reasoning and cognitively-guided selection.

1 Introduction

Large Language Models (LLMs) have demonstrated significant capabilities across various NLP tasks (Wang et al. 2024; Xu et al. 2025), with their In-Context Learning (ICL) enabling zero-shot and few-shot learning capabilities (Müller, Hollmann, and Hutter 2024). The ICL approach has proven effective in Relation Extraction (RE), where models identify and classify semantic relationships between entities in text (Wan et al. 2023; Ma et al. 2023). While ICL offers a flexible solution for RE, its effectiveness depends on the selection of representative examples. By meticulously selecting and optimizing example inputs, ICL can activate relevant parameters through distribution alignment, thereby enhancing performance and adaptability for specific tasks (Detailed analysis of the mechanism in Appendix A).

Current approaches in ICL for relation extraction face fundamental limitations in example selection that stem from insufficient understanding of how LLMs process relational information. These methods typically focus on surface-level similarity metrics while neglecting the cognitive mechanisms underlying relational reasoning (An et al. 2023). This cognitive mechanism involves two aspects: distributional diversity and consistency, ensuring the example space covers diverse relation patterns to activate relevant parameters, and example selection effectiveness, ensuring chosen examples engage the model's semantic-structural understanding to capture deep relation patterns.

Psychological studies have shown that counterfactual reasoning—constructing "what-if" scenarios—is a fundamental mechanism underlying human causal inference and reasoning (Cui et al. 2024; Schubert et al. 2024). In relation extraction, this mechanism enables models to capture implicit causal and semantic associations by contrasting potential scenarios, thereby forming a more comprehensive distributional representation. However, existing ICL methods rely on static example pools, which fail to reflect such dynamic, counterfactual reasoning-based cognitive processes. Meanwhile, cognitive science research has revealed that relational reasoning involves both semantic comprehension and structural pattern recognition (Acharya, Jia, and Ginsburg 2024). Current ICL methods for relational reasoning, however, insufficiently capture this dual-process mechanism, focusing mainly on semantic similarity while overlooking structural recognition patterns that are essential for attention modulation and relational understanding in LLMs.

Motivated by these insights, we hypothesize that optimizing the sample distribution via counterfactual analysis and aligning example selection with human cognitive patterns for relational understanding can enhance LLM relational reasoning. To this end, we propose the Counterfactual Cognitive Alignment (CCA) framework, which integrates cognitive science principles with ICL optimization. Our framework aligns example selection with the dual-process mechanism observed in human relational reasoning—semantic understanding and structural pattern recognition—ensuring that chosen examples are not only semantically relevant but also exhibit structural patterns conducive to LLM relational comprehension.

Grounded in established theories of counterfactual reasoning in cognitive psychology, CCA introduces a principled approach to example space construction that mirrors

[*]These authors contributed equally.
[†]Corresponding author

human "what-if" reasoning processes. Unlike traditional data augmentation methods that focus on increasing quantity, our cognitive-inspired generation creates examples that are specifically designed to enhance the model's understanding of relational patterns through systematic exploration of counterfactual scenarios. Operationally, CCA implements this cognitive framework through three systematic generation methods that create relationally coherent examples while preserving semantic validity. Furthermore, we identify three key structural identification patterns that align with LLMs' attention mechanisms and develop a cognitive alignment mechanism that integrates these patterns with semantic understanding. This integration ensures that selected examples match both the semantic requirements and the cognitive processing patterns inherent in LLMs' relational reasoning capabilities. Our main contributions are as follows:

- We develop CCA, a novel cognitive alignment framework that enhances ICL performance by aligning example selection with established cognitive principles of relational reasoning. Our method creates a cognitively coherent example space that facilitates LLMs' natural processing patterns for relation extraction.
- We design an innovative dual-processing alignment mechanism that integrates semantic comprehension with structural pattern recognition, mirroring the cognitive mechanisms underlying human relational reasoning. This alignment enables more effective example selection by matching LLMs' inherent cognitive processing patterns.
- Our detailed analysis reveals how cognitive alignment principles enhance relational reasoning capabilities, demonstrating the effectiveness of bridging cognitive science with computational linguistics for improved ICL performance.

2 Related Work

2.1 In-Context Learning

Recent studies have shown that LLMs can effectively perform diverse tasks through few-shot learning with natural language prompts. In relation extraction, ICL achieves promising results via contextual example selection (Wan et al. 2023; Li et al. 2024) and instruction optimization (Ma, Li, and Zhang 2023). Example retrieval methods based on semantic similarity (Rubin, Herzig, and Berant 2022) and embedding optimization (Wan et al. 2023) have been shown to improve accuracy. However, incomplete semantic information and biased retrieved samples often introduce cognitive bias in LLMs. Recent findings reveal that corpus distribution, data characteristics, and sample difficulty strongly influence ICL performance (Wang et al. 2024; Zhang et al. 2024). These insights underscore the need to align example selection with LLMs' cognitive processing—integrating semantic understanding with structural pattern recognition through cognitively informed example construction.

2.2 Relation Extraction

Relation extraction (RE), which identifies and classifies semantic relations between entities, has long been a core challenge in information extraction. Earlier studies focused on effective relational encoding, such as tabular (Miwa and Sasaki 2014; Gupta, Schütze, and Andrassy 2016; Shang, Huang, and Mao 2022) and span-based representations (Zhong and Chen 2021; Ye et al. 2022; Li et al. 2025b,a). With the advancement of LLMs, research shifted toward fine-tuning decoder-based models for relation classification (Wang et al. 2023b; Bai et al. 2025). However, the high cost of fine-tuning has motivated growing interest in In-Context Learning (ICL) (Wan et al. 2023; Ma et al. 2023), which leverages few-shot examples to activate relational reasoning without parameter updates. In this work, we introduce a cognitively informed framework that aligns ICL example selection with human relational reasoning principles to enhance LLMs' capacity for relation extraction.

3 Preliminary

In this section, we introduce few-shot selection in ICL as the preliminary of our method.

3.1 Few-shot Selection in ICL

ICL requires selecting appropriate examples as demonstrations to guide LLMs (Dutta et al. 2024; Müller, Hollmann, and Hutter 2024). Formally, given a LLM f_θ with frozen parameters θ, the few-shot selection process in ICL can be formalized as:

$$\mathcal{D} = (x_i, y_i)_{i=1}^{k} \tag{1}$$

$$y^* = f_\theta(\mathcal{D}, x^*) \tag{2}$$

where $\mathcal{D}$ denotes the selected demonstrations consisting of k input-output pairs, (x_i, y_i) represents the i-th selected example and its corresponding label, x^* is the target instance to be predicted, and y^* is the LLM prediction.

Given an example space $\mathcal{E}$, the key challenge is selecting k demonstrations $\mathcal{D}$ that balance relevance to the target instance with diversity and representativeness. We explore optimizing few-shot selection to enhance ICL performance without model training or fine-tuning.

4 Method

In this section, we propose the **Counterfactual-based Cognitive Alignment (CCA)** framework to enhance ICL for relation extraction in LLMs. Figure 1 provides an overview of CCA. The method has three core parts: (1) Counterfactual Example Generation, (2) Cognitive Processing Pattern Extraction, (3) Cognitive Alignment Example Selection.

4.1 Counterfactual Example Generation

We introduce *Counterfactual Example Generation* into the CCA framework as its core cognitive foundation. This design is grounded in the psychological theory of counterfactual reasoning, where humans construct hypothetical scenarios to explore latent causal structures and deepen relational understanding. Existing ICL methods rely on static example pools, failing to capture this dynamic cognitive pattern. To address this limitation, CCA simulates the human counterfactual reasoning process to systematically expand the example space, optimize semantic distributions, and enhance the relational reasoning capability of LLMs.

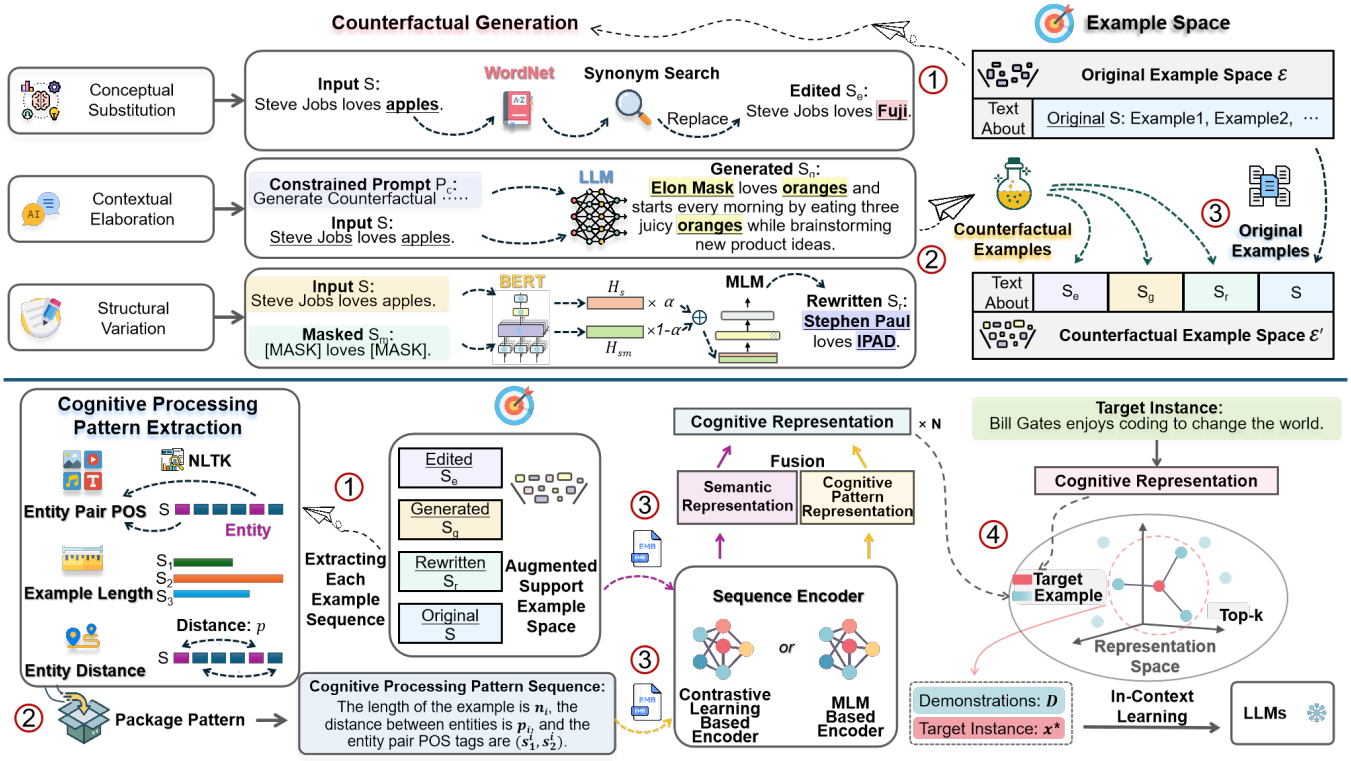

Figure 1: Overview of the Counterfactual Cognitive Alignment (CCA) framework for relation extraction. It comprises three components: (1) Counterfactual Generation, producing diverse examples via knowledge editing, constrained infilling, and semantic rewriting; (2) Cognitive Pattern Extraction, capturing structural patterns to augment the support example space; and (3) Cognitive Alignment Example Selection, integrating semantic and cognitive patterns into a unified representation for similarity-based example selection to improve ICL performance.

Our counterfactual generation module instantiates psychological theories of counterfactual reasoning via semantic transformations that preserve relational coherence while introducing alternative realizations. This mirrors human reasoning, which maintains core relational structure while varying surface forms. Formally, given an example pair $(x_i, y_i) \in \mathcal{E}$, a cognitively motivated transformation $T(\cdot)$ produces relationally coherent counterfactuals, forming an aligned example space $\mathcal{E}'$ that better supports LLMs' relational understanding.

$$\mathcal{E}' = \mathcal{E} \cup \{T(x_i, y_i) | (x_i, y_i) \in \mathcal{E}\} \tag{3}$$

To systematically operationalize counterfactual reasoning principles, we develop three cognitively motivated generation strategies that correspond to different aspects of human counterfactual thinking: (1) Conceptual Substitution, (2) Contextual Elaboration, (3) Structural Variation. Each strategy reflects a distinct human cognitive pathway for exploring counterfactuals in relational reasoning.

Counterfactual Conceptual Substitution

Conceptual Substitution simulates the cognitive process of considering alternative entities within the same relational framework. By leveraging `WordNet` (Miller 1992) to generate conceptually similar entities, it constructs counterfactual scenarios that preserve relational coherence while varying participants, thereby enhancing the model's understanding of relational invariance.

Counterfactual Contextual Elaboration

Contextual Elaboration captures the cognitive mechanism of exploring how relationships manifest in different contextual scenarios. This strategy systematically investigates how relational structures adapt to varying contextual conditions, enhancing the model's contextual sensitivity in relational understanding. By constraining relationship preservation while allowing contextual variation, we employ `DeepSeek-R1` (DeepSeek-AI 2025) to generate counterfactual elaborations that demonstrate relational robustness across different contexts. This approach strengthens the model's ability to recognize relational patterns despite contextual variations. Detailed prompting strategies are provided in Appendix B.2.

Counterfactual Structural Variation

Structural Variation captures the cognitive process of expressing the same relationship through alternative linguistic forms, enhancing robustness to relational expression diversity. Using masked language modeling (MLM)(Wolf et al. 2020), we generate structurally varied yet semantically equivalent alternatives by masking and reconstructing key elements while preserving core relational structure. Contextual representations H_s and H_{sm} from BERT(Devlin et al.

2019) are then fused via cognitively motivated interpolation:

$$H_F = \alpha \cdot H_s + (1 - \alpha) \cdot H_{sm} \tag{4}$$

where α controls the balance between original and masked information. The cognitively-fused embedding H_F is then processed through BERT's pre-trained MLM head to generate structurally varied alternatives that maintain relational coherence. This approach enables systematic exploration of linguistic variation while preserving the cognitive patterns underlying relational expression, thus enhancing the model's robustness to diverse relational manifestations.

4.2 Cognitive Processing Pattern Extraction

To select effective ICL examples, we leverage Cognitive Processing Patterns—patterns describing how LLMs perceive and process demonstrations, mirroring cognitive mechanisms in relational reasoning (Xiao et al. 2017). These patterns capture LLMs' ability to recognize structural and relational cues in $\mathcal{D}$, which strongly impacts performance on the target instance x^*. We therefore extract such patterns for each example to guide selection.

Given an example space $\mathcal{E}$, for each example $x_i \in \mathcal{E}$, we define its cognitive processing pattern through three key characteristics: (1) Example Length, (2) Entity Distance, (3) Entity Pair POS. CCA models the cognitive processing of language through three dimensions—Length, Entity Distance, and Part-of-Speech (POS)—representing Information Load, Attentional Span, and Structural Constraint. Example Length reflects information density and contextual integration (Sweller 2011), Entity Distance models attention decay during relation identification (Panichello and Buschman 2021), and POS combinations encode syntactic constraints on semantic inference. These dimensions form a hierarchical cognitive framework from discourse scope to syntactic function, capturing LLMs' relational perception while preserving interpretability by excluding overly complex structural features.

Example Length: The length n of an example serves as a crucial indicator of information density and context scope. Shorter examples tend to contain more concentrated information, while longer ones provide broader context but may include noise. This metric helps evaluate the information density during example selection.

Entity Distance: For an example containing entity pair (e_1, e_2) with positions (p_1, p_2), we calculate their distance $p = |p_1 - p_2|$. This physical distance metric reflects how LLMs perceive the relationship strength between entities, with closer entities typically indicating stronger semantic connections.

Entity Pair POS: Each entity e_i is associated with its part-of-speech tag s_i, obtained through `NLTK` processing. The POS pattern (s_1, s_2) provides syntactic cues about entity interactions, helping understand the grammatical nature of their relationships.

We combine these three aspects into a structured representation for each example x_i:

$$\mathcal{I}(x_i) = (n_i, p_i, (s_1^i, s_2^i)) \tag{5}$$

where $n_i \in \mathbb{N}$ represents the example length, $p_i \in \mathbb{N}$ denotes the entity distance, and (s_1^i, s_2^i) represents the entity pair POS tags.

Traditional encoding methods like one-hot encoding or training separate representation models for different feature types in $\mathcal{I}(x_i)$ face significant challenges in capturing the semantic relationships between these heterogeneous features. Instead, we propose a more natural and semantically-aware approach by packing the cognitive processing pattern through three key characteristics into a descriptive text format before encoding:

> **Cognitive Processing Pattern Sequence**: *The length of the example is $\underline{n_i}$, the distance between entities is $\underline{p_i}$, and the entity pair POS tags are $\underline{(s_1^i, s_2^i)}$.*

This text-based representation allows us to leverage LMs as cognitive processing pattern encoders, transforming the descriptive texts into continuous embeddings $H_{\mathcal{I}}$ while preserving the semantic relationships and interpretability of different features.

4.3 Cognitive Alignment Example Selection

Cognitive alignment in our framework incorporates two fundamental dimensions of information processing: *semantic understanding* and *cognitive processing pattern*. To effectively utilize both types of information, we propose a *Cognitive-Guided Example Selection* strategy that selects examples based on the integration of semantic and cognitive pattern representations. For each example in the augmented example space $\mathcal{E}'$, we employ a unified LM encoder f_{enc} to obtain dual representations:

$$H_S^i = f_{enc}(x_i) \tag{6}$$

$$H_{\mathcal{I}}^i = f_{enc}(\mathcal{I}(x_i)) \tag{7}$$

H_S^i and $H_{\mathcal{I}}^i$ are strategically integrated through mean pooling operations and concatenated to form a comprehensive cognitive representation:

$$H_C^i = [\text{meanpool}(H_S^i); \text{meanpool}(H_{\mathcal{I}}^i)] \tag{8}$$

where H_C^i serves as a holistic cognitive embedding that encapsulates both the semantic content and identification characteristics.

To facilitate accurate example selection, we implement a similarity-based retrieval process. The relevance between examples is quantified using cosine similarity:

$$\text{sim}(H_C^*, H_C^i) = \frac{H_C^* \cdot H_C^i}{|H_C^*| \cdot |H_C^i|} \tag{9}$$

The final demonstration set $\mathcal{D}$ is constructed by selecting the k examples that exhibit the highest similarity scores:

$$\mathcal{D} = \text{TopK}(\text{sim}(H_C^*, H_C^i) | x_i \in \mathcal{E}', k) \tag{10}$$

This cognitive alignment approach ensures that the selected examples are not only semantically relevant to the target instance x^* but also share similar identification patterns, thereby enhancing the effectiveness of ICL.

Models	Backbone	Method	SemEval				TACRED				SciERC			
			5-shot	10-shot	20-shot	50-shot	5-shot	10-shot	20-shot	50-shot	5-shot	10-shot	20-shot	50-shot
LLaMA2-7B	Random	ICL	48.40	49.11	49.65	49.31	24.17	23.69	24.66	24.21	10.37	11.43	11.62	11.44
		CCA	**49.32**	**49.17**	**50.49**	**51.94**	**24.13**	**24.15**	**24.97**	**25.24**	**11.34**	**12.18**	**12.79**	**12.64**
	SimCSE	ICL	57.33	59.13	62.49	64.26	27.48	28.64	30.08	27.81	13.48	13.89	14.62	14.79
		CCA	**60.97**	**62.77**	**64.55**	**66.11**	**29.78**	**31.95**	**32.03**	**33.26**	**18.48**	**18.13**	**19.24**	**20.88**
	BERT + PURE + ICL	ICL	58.68	64.90	65.67	72.32	26.11	26.35	31.15	33.35	14.75	15.43	16.92	16.77
		CCA	**63.44**	**67.64**	**71.51**	**76.77**	**30.36**	**32.41**	**33.94**	**35.87**	**20.42**	**20.14**	**21.75**	**22.24**
Qwen2.5-7B	Random	ICL	47.18	47.07	**48.91**	49.21	24.65	25.12	**24.93**	26.83	14.32	15.66	15.83	16.31
		CCA	**48.26**	**48.37**	48.68	**50.77**	**25.35**	**26.40**	24.23	**27.63**	**15.57**	**16.52**	**16.86**	**17.03**
	SimCSE	ICL	56.73	59.05	62.64	63.78	28.26	29.37	30.44	31.78	15.66	16.21	16.72	17.42
		CCA	**60.39**	**62.71**	**64.07**	**65.10**	**31.36**	**32.41**	**32.94**	**34.31**	**20.40**	**22.90**	**22.43**	**22.94**
	BERT + PURE + ICL	ICL	59.37	65.36	67.93	74.38	27.37	27.49	32.37	34.52	17.75	18.14	18.95	19.66
		CCA	**64.48**	**70.18**	**76.70**	**77.91**	**32.47**	**33.53**	**35.64**	**37.69**	**22.77**	**23.61**	**23.52**	**23.91**
Qwen2.5-72B	Random	ICL	58.99	**64.31**	68.27	70.33	31.21	32.46	33.38	34.49	18.63	18.12	18.44	18.62
		CCA	**59.82**	63.11	**70.42**	**71.10**	**31.45**	**32.77**	**34.10**	**34.68**	**18.71**	**18.93**	**19.61**	**19.44**
	SimCSE	ICL	66.07	67.94	70.12	73.21	35.86	37.94	37.73	39.08	21.60	21.46	22.87	21.61
		CCA	**68.13**	**69.23**	**72.51**	**77.09**	**37.73**	**39.51**	**39.40**	**39.95**	**22.58**	**23.60**	**24.35**	**23.96**
	BERT + PURE + ICL	ICL	68.93	71.29	74.51	77.17	35.90	37.14	38.13	39.14	21.64	22.52	21.73	22.81
		CCA	**70.26**	**73.55**	**76.25**	**80.15**	**37.88**	**38.19**	**39.21**	**40.00**	**22.42**	**22.81**	**23.52**	**23.77**

Table 1: Results (micro F1 scores) on SemEval, TACRED, and SciERC datasets in various few-shot settings. We use open-source LLMs: LLaMA2-7B-chat, Qwen2.5-7B/72B-Instruct. The best results are in **bold**.

5 Experiment and Evaluation

5.1 Datasets

We evaluate CCA on three standard RE datasets: SemEval (Hendrickx et al. 2010), TACRED (Zhang et al. 2017), and SciERC (Luan et al. 2018), under few-shot settings with $k = 5, 10, 20, 50$ examples per relation. Dataset details are provided in Appendix C.1.

5.2 ICL Foundation Models

We evaluate both open-source and closed-source models. For open-source models, we use LLaMA2-7B-chat (Touvron et al. 2023) and Qwen2.5-7B/72B-Instruct (Yang et al. 2024), covering diverse architectures and scales (7B–72B). All experiments use greedy decoding. Closed-source evaluations are conducted via GPT-3.5 and Claude-3.5-Sonnet APIs.

5.3 Backbones

Consistent with previous methods, we selected three different methods as the initial in-context demonstrations backbone of CCA. 1) Random: This method randomly selects the demonstration without LM. 2) SimCSE: The SimCSE method (Gao, Yao, and Chen 2021) is used to retrieve samples that are semantically similar to the test instance as the initial context demonstration. 3) BERT + PURE + ICL: This retriever uses PURE (Zhong and Chen 2021) retriever that has been trained on BERT with labeled samples (Wan et al. 2023). For more detailed information about backbones, see Appendix C.3.

5.4 Baselines

To assess the effectiveness of CCA on the ICL RE task, we conducted a comparative analysis between CCA and five baseline models: **R-Bert** (Wu and He 2019), **KnowPrompt** (Chen et al. 2022), **Self-Refine** (Madaan et al. 2023), **Self-Consistency** (Wang et al. 2023a), **I^2CL** (Li et al. 2024). Detailed descriptions of these baseline models are provided in Appendix C.4.

6 Results

6.1 Effectiveness on Open-Source LLMs

Table 1 reports the experimental results on the SemEval, TACRED, and SciERC datasets using different backbone strategies under open-source LLMs. The results indicate that CCA consistently outperforms the baseline across various scales of LLMs and different backbone strategies. Notably, the improvements are most significant under the conditions using SimCSE and BERT as backbones, with an average increase in micro F1 score of 3.35%. This can be attributed to CCA's thorough exploration of the sample space, enabling more precise context selection.

6.2 Effectiveness on Close-Source LLMs

Table 3 presents the experimental results of CCA on the SemEval, TACRED, and SciERC datasets using different backbone strategies in close-source LLMs. Due to cost considerations, the experiments were limited to 5-shot and 10-shot performance evaluations. The results demonstrate that CCA also surpasses the baseline in closed-source LLMs, clearly indicating that CCA is a versatile and universally applicable ICL method across various LLMs architectures.

Models	Backbones	Method	SemEval		TACRED		SciERC	
			5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
GPT-3.5	Random	ICL	57.38	60.03	30.36	31.06	17.73	17.08
		CCA	**57.80**	**60.71**	**31.21**	**31.41**	**18.76**	**18.64**
	SimCSE	ICL	65.14	65.83	31.90	32.01	20.03	21.16
		CCA	**67.06**	**68.94**	**32.58**	**34.88**	**21.14**	**22.69**
	BERT + PURE + ICL	ICL	67.26	70.58	32.46	33.38	20.45	21.66
		CCA	**69.45**	**72.32**	**34.33**	**34.57**	**21.15**	**22.16**
Claude 3.5	Random	ICL	59.21	**63.87**	**32.77**	33.32	19.44	19.52
		CCA	**59.28**	63.79	32.62	**34.01**	**19.79**	**19.66**
	SimCSE	ICL	67.83	68.94	36.31	38.04	22.07	22.68
		CCA	**68.14**	**70.86**	**38.40**	**39.15**	**23.27**	**23.18**
	BERT + PURE + ICL	ICL	68.55	72.89	36.51	37.79	22.77	23.04
		CCA	**70.78**	**74.51**	**38.50**	**38.71**	**23.04**	**23.62**

Table 2: Results (micro F1 scores) on relation extraction tasks across SemEval, TACRED, and SciERC datasets under few-shot settings. We evaluate using GPT-3.5 and Claude 3.5. Best scores shown in **bold**.

Method	SemEval		TACRED		SciERC	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
RB ♣	42.75	57.25	9.87	16.24	15.24	16.17
KP ♣	53.92	56.42	27.86	30.34	17.34	18.57
SR ♠	68.97	71.58	36.50	37.92	21.13	22.09
SC ♠	69.97	72.23	35.69	37.47	21.23	22.53
I^2CL ♠	69.31	72.53	36.74	37.65	21.11	21.90
CCA	**70.26**	**73.55**	**37.88**	**38.19**	**22.42**	**22.81**

Table 3: Experimental results compare the CCA framework with SOTA methods using micro F1 scores. Qwen2.5-72B serves as the LLM, with "BERT + PURE + ICL" as the ICL backbone. Baselines include R-BERT (RB), KnowPrompt (KP), Self-Refine (SR), and Self-Consistency (SC), where ♣ denotes BERT-based and ♠ LLM-based methods.

Method	SemEval		TACRED	
	5-shot	10-shot	5-shot	10-shot
CCA	**60.97**	**62.77**	**29.78**	**31.95**
w/o CS	60.27	62.09	28.52	29.33
w/o CE	59.19	62.13	29.59	29.93
w/o SV	59.03	61.23	28.77	30.16
w/o CS+SV	57.90	60.50	28.52	29.78
w/o CS+CE	59.07	61.78	29.31	29.17
w/o CE+SV	58.31	60.46	27.73	29.83
w/o CS+SV+CE	57.33	59.13	27.48	28.64

Table 4: Comparison of different counterfactual generation methods under LLaMA2-7B-chat using SimCSE as the backbone. The best results are in **bold**. CS: Counterfactual Conceptual Substitution, CE: Counterfactual Contextual Elaboration, SV: Counterfactual Structural Variation

6.3 Comparison with Existing Approaches

To accurately assess the performance gap between our method and state-of-the-art few-shot relation extraction techniques, we compared CCA with LLM-based methods Self-Refine, Self-Consistency, and bert-based methods R-Bert and KonewPrompt. The results in Table 3, achieved with a BERT + PURE + ICL backbone under Qwen2.5-72B, outperformed the baselines. This success is due to the advanced performance of LLMs and our novel demonstration selection strategy.

7 Ablation Study

7.1 Analysis of Individual Counterfactual Generation Method

This paper employs three distinct counterfactual generation methods: Counterfactual Conceptual Substitution, Counterfactual Contextual Elaboration and Counterfactual Structural Variation, to expand the sample space and assess their effectiveness. We conducted individual and combined tests of these methods across two datasets. The results in Table 4, illustrate the performance under the SimCSE framework with 5-shot and 10-shot configurations. The findings indicate that all counterfactual generation methods significantly enhance model performance, particularly Counterfactual Structural Variation. Integrating these methods boosts performance, as combining generation strategies enables CCA to produce higher-quality samples and optimize context learning.

7.2 Performance of Isolated Example Representation

CCA constructs a cognitive-guided strategy by integrating cognitive processing pattern (CPP) and semantic understanding (SU), enabling the model to synthesize multi-level information for more precise example selection. To validate the effectiveness of these features in example selec-

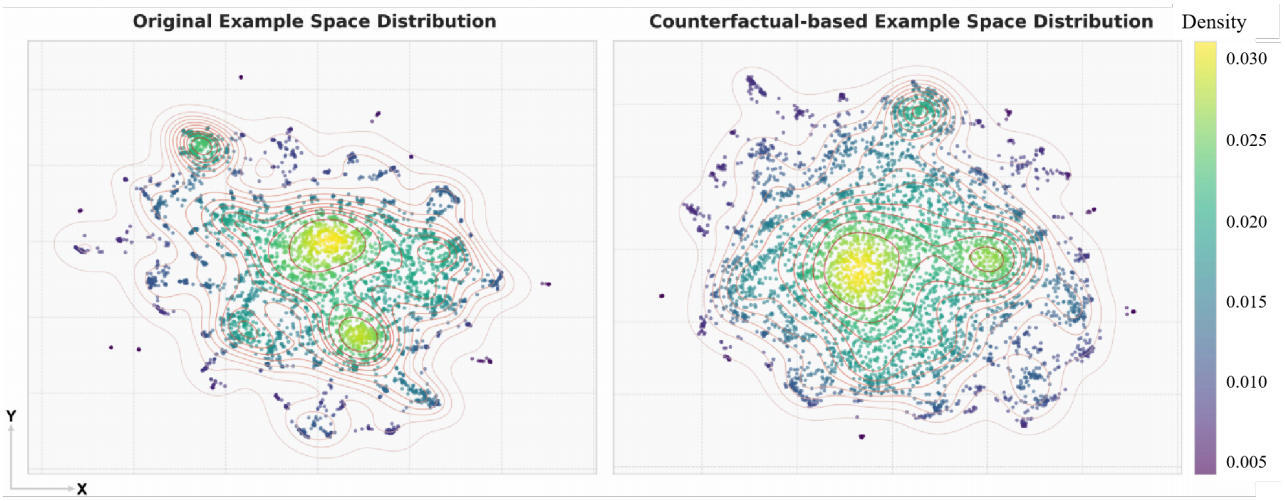

Figure 2: Visualization of example space distributions for original examples (left) and counterfactual-based examples (right). The density contours illustrate that counterfactual-based examples achieve a more diverse and balanced distribution. The color gradient from purple to yellow indicates increasing example density, with values ranging from 0.005 to 0.030.

Model	Method	SemEval		TACRED	
		5-shot	10-shot	5-shot	10-shot
LLaMA2-7B	Random	48.40	49.11	24.17	23.69
	SU	57.33	59.31	27.48	28.64
	CPP	55.17	56.28	27.00	26.14
	CG+CPP	57.13	59.45	27.53	27.97
	CG+SU	59.76	61.26	29.18	31.38
	CCA	**60.97**	**62.77**	**29.78**	**31.95**
Qwen2.5-7B	Random	47.18	47.07	24.65	25.12
	SU	56.73	59.05	28.26	29.37
	CPP	54.90	57.16	27.55	25.91
	CG+CPP	56.41	59.63	28.38	28.91
	CG+SU	59.38	62.00	30.42	31.92
	CCA	**60.39**	**62.71**	**31.36**	**32.41**

Table 5: Comparison of micro-F1 scores on SemEval and TACRED dataset using different sample representations using SimCSE as backbone. CG: Counterfactual Example Generation

tion, we conducted ablation experiments, with results presented in Table 5. Specifically, we conducted ablation studies under four configurations: using only CPP, using only SU, counterfactual example generation combined with cognitive processing pattern (CG+CPP), and counterfactual example generation combined with semantic understanding (CG+SU). Results show that CPP outperforms random sampling by capturing richer distributional and structural patterns, while SU surpasses CPP by providing deeper semantic cues for complex relations. Counterfactual generation (CG) further expands semantic coverage and enhances generalization, and its combination with CPP and SU delivers complementary benefits, significantly improving recognition and reasoning in relation extraction.

7.3 Manifold Analysis of Example Space Distribution

To understand how counterfactual generation improves ICL performance, we analyze the distribution of example spaces through manifold visualization (Figure 2). The visualization reveals that our counterfactual generation successfully optimizes the example space distribution in three aspects: (1) Enhanced density uniformity: The augmented space shows more continuous and uniform density distribution, effectively filling the gaps in the original space while maintaining essential semantic structures. (2) Improved semantic continuity: The smoother transitions between different regions indicate better semantic bridging between examples, facilitating more robust reasoning paths during ICL. (3) Balanced coverage expansion: While broadening the coverage of example space, our method preserves the original cluster structures, ensuring the generated examples remain task-relevant and semantically meaningful. These structural improvements directly support our method's effectiveness in enhancing ICL performance through better example space organization. Details and settings are in Appendix F.

8 Conclusion

In this paper, we investigate the enhancement of LLMs' performance on relation extraction through cognitive alignment in ICL. We propose the Counterfactual-based Cognitive Alignment (CCA) framework, which operates without the necessity of model finetuning or manual annotation, demonstrating substantial improvement in ICL's capacity for relation extraction across LLMs. Empirical evaluations indicate that the performance improvements achieved through CCA exhibit consistent robustness across diverse datasets. This advances cognitive-informed ICL and validates the effectiveness of counterfactual reasoning principles.

Acknowledgments

This work was supported in part by the Science and Technology Innovation Key R&D Program of Chongqing (Grant No. CSTB2024TIAD-STX0027) and the Natural Science Foundation of Shandong Province, China (Grant No. ZR2025QC1571).

References

Acharya, S.; Jia, F.; and Ginsburg, B. 2024. Star Attention: Efficient LLM Inference over Long Sequences. arXiv:2411.17116.

An, S.; Zhou, B.; Lin, Z.; Fu, Q.; Chen, B.; Zheng, N.; Chen, W.; and Lou, J.-G. 2023. Skill-Based Few-Shot Selection for In-Context Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13472–13492. Singapore: Association for Computational Linguistics.

Bai, S.; Li, Q.; Wang, Z.; Zhou, N.; and Yao, N. 2025. Enhancing NLU in Large Language Models Using Adversarial Noisy Instruction Tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22): 23451–23459.

Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2778–2788. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Cui, J.; Yu, M.; Jiang, B.; Zhou, A.; Wang, J.; and Zhang, W. 2024. Interpretable Knowledge Tracing via Response Influence-based Counterfactual Reasoning. arXiv:2312.10045.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dutta, S.; Pandita, D.; Weerasooriya, T. C.; Zampieri, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2024. ARTICLE: Annotator Reliability Through In-Context Learning. arXiv:2409.12218.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Gupta, P.; Schütze, H.; and Andrassy, B. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2537–2547.

Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Ó Séaghdha, D.; Padó, S.; Pennacchiotti, M.; Romano, L.; and Szpakowicz, S. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In Erk, K.; and Strapparava, C., eds., *Proceedings of the 5th International Workshop on Semantic Evaluation*, 33–38. Uppsala, Sweden: Association for Computational Linguistics.

Li, G.; Ke, W.; Wang, P.; Xu, Z.; Ji, K.; Liu, J.; Shang, Z.; and Luo, Q. 2024. Unlocking Instructive In-Context Learning with Tabular Prompting for Relational Triple Extraction. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 17131–17143. Torino, Italia: ELRA and ICCL.

Li, Q.; Bai, S.; Zhou, N.; and Yao, N. 2025a. Enhancing entity and relation extraction with dynamic hard negative augmentation framework. *Engineering Applications of Artificial Intelligence*, 161: 112211.

Li, Q.; Yao, N.; Zhou, N.; and Zhao, J. 2025b. Entity and relationship extraction based on span contribution evaluation and focusing framework. *Computer Speech & Language*, 90: 101744.

Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Ma, X.; Li, J.; and Zhang, M. 2023. Chain of Thought with Explicit Evidence Reasoning for Few-shot Relation Extraction. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2334–2352. Singapore: Association for Computational Linguistics.

Ma, Y.; Cao, Y.; Hong, Y.; and Sun, A. 2023. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10572–10601. Singapore: Association for Computational Linguistics.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651.

Miller, G. A. 1992. WordNet: A Lexical Database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Miwa, M.; and Sasaki, Y. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1858–1869.

Müller, S.; Hollmann, N.; and Hutter, F. 2024. Bayes' Power for Explaining In-Context Learning Generalizations. arXiv:2410.01565.

Panichello, M. F.; and Buschman, T. J. 2021. Shared mechanisms underlie the control of working memory and attention. *Nature*, 592(7855): 601–605. Epub 2021 Mar 31.

Rubin, O.; Herzig, J.; and Berant, J. 2022. Learning To Retrieve Prompts for In-Context Learning. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2655–2671. Seattle, United States: Association for Computational Linguistics.

Schubert, J. A.; Jagadish, A. K.; Binz, M.; and Schulz, E. 2024. In-context learning agents are asymmetric belief updaters. arXiv:2402.03969.

Shang, Y.-M.; Huang, H.; and Mao, X. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11285–11293.

Sweller, J. 2011. Cognitive Load Theory. *Psychology of Learning and Motivation*, 55: 37–76.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wan, Z.; Cheng, F.; Mao, Z.; Liu, Q.; Song, H.; Li, J.; and Kurohashi, S. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3534–3547. Singapore: Association for Computational Linguistics.

Wang, Q.; Wang, Y.; Wang, Y.; and Ying, X. 2024. Can In-context Learning Really Generalize to Out-of-distribution Tasks? arXiv:2410.09695.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023a. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wang, X.; Zhou, W.; Zu, C.; Xia, H.; Chen, T.; Zhang, Y.; Zheng, R.; Ye, J.; Zhang, Q.; Gui, T.; et al. 2023b. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. *arXiv preprint arXiv:2304.08085*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wu, S.; and He, Y. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, 2361–2364. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369763.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint Detection and Identification Feature Learning for Person Search. In *CVPR*.

Xu, Q.; Bai, S.; Chen, L.; Liu, Z.; and Li, Y. 2025. ChemLabs on ChemO: A Multi-Agent System for Multimodal Reasoning on IChO 2025. arXiv:2511.16205.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Ye, D.; Lin, Y.; Li, P.; and Sun, M. 2022. Packed Levitated Marker for Entity and Relation Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4904–4917. Dublin, Ireland: Association for Computational Linguistics.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35–45. Copenhagen, Denmark: Association for Computational Linguistics.

Zhang, Z.-Y.; Zhang, J.; Yao, H.; Niu, G.; and Sugiyama, M. 2024. On Unsupervised Prompt Learning for Classification with Black-box Language Models. arXiv:2410.03124.

Zhong, Z.; and Chen, D. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 50–61. Online: Association for Computational Linguistics.