Semantic Networks Extracted from Students' Think-Aloud Data are Correlated with Students' Learning Performance

Pingjing Yang[1] **Sullam Jeoung**[1] **Jennifer Cromley**[1] **Jana Diesner**[1,2]
[1]University of Illinois at Urbana-Champaign
[2]Technical University of Munich
py2@illinois.edu, sullamij@gmail.com
jcromley@illinois.edu, jana.diesner@tum.de

Abstract

When students reflect on their learning from a textbook via think-aloud processes, network representations can be used to capture the concepts and relations from these data. What can we learn from the resulting network representations about students' learning processes, knowledge acquisition, and learning outcomes? This study brings methods from entity and relation extraction using classic and LLM-based methods to the application domain of educational psychology. We built a ground-truth baseline of relational data that represents relevant (to educational science), textbook-based information as a semantic network. Among the tested models, SPN4RE and LUKE achieved the best performance in extracting concepts and relations from students' verbal data. Network representations of students' verbalizations varied in structure, reflecting different learning processes. Correlating the students' semantic networks with learning outcomes revealed that denser and more interconnected semantic networks were associated with more elaborated knowledge acquisition. Structural features such as the number of edges and surface overlap with textbook networks significantly correlated with students' posttest performance.

1 Introduction

As educational resources have become increasingly abundant and digitized, large amounts of data capturing both instructional materials (e.g., textbooks, slides) and students' learning processes have become available (Romero and Ventura, 2013). Large-scale online educational offerings, such as massive open online courses (MOOC), have also become accessible to hundreds of millions of learners, but efficiently evaluating these students' performance remains a crucial task for educators (Conijn et al., 2018). By advancing and automating the analysis of educational materials and students' learning of these materials, educators and (to some degree) automated systems can provide real-time, personalized education experiences that accommodate and support students' varying learning progress (Zhang et al., 2019); offering an added value to learners.

Computational techniques are increasingly used in educational research to evaluate both learning materials and students' responses (Charitopoulos et al., 2020). For example, Lucy et al. (2020) applied data science techniques to analyze U.S. history textbooks, revealing an underrepresentation of marginalized groups and systematic patterns in the portrayal of women and Black people. The Educational Testing Service (ETS) uses an e-rater engine to score students' writing skills on their standardized English tests (Burstein, 2003). Wang et al. (2022b) leveraged Bert models to score student essays and found a performance comparable to that of human graders. With the rapid development of computing technologies, educational evaluation has become more automated than ever before (Charitopoulos et al., 2020).

In the context of education science, verbal data, including transcripts of classroom discourse, small-group dialogues, and talk-aloud protocols from reasoning and problem-solving tasks, are increasingly used by researchers to understand and improve learning processes. Think-alouds are participant verbalizations while learning or solving problems that are audio-recorded and transcribed. Verbalizations can include content from the learning materials and/or verbalizations of learner strategies (e.g., self-generating a question). In this paper, we use semantic network analysis techniques to analyze the passage-related content that learners verbalized in the Think-Aloud data we used, excluding any reading or re-reading of the texts they were learning from. The meaningful analysis of such verbal data is not as simple as counting the occurrence of key concepts, but is a complex process that includes capturing linguistic features (Rogers, 2004), finding structures in text (Lemke, 1995), and applying

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25803–25816
November 4-9, 2025

statistical analysis techniques (Lemke, 2012). Semantic Network Analysis (SNA) transforms textual data into networks of concepts, enabling the integration of lexical, syntactic, and semantic analyses into a unified representation of meaning (Diesner and Carley, 2011b,a). In an SNA, relevant information from textual data sources, such as transcripts of verbal data, is first converted into graphs that consist of nodes representing key concepts defined by researchers and edges representing connections between these nodes. Subsequently, researchers can apply network metrics, such as betweenness centrality (Wagner and Priemer, 2023) and PageRank centrality (Bodin, 2012), or network algorithms such as community detections (Siew et al., 2019), to understand the structure and patterns of the generated networks and correlate them, e.g., with student performance measures.

Conducting SNA requires the construction or extraction of reliable, i.e., accurate with respect to the underlying data, relational data that represent the relevant (to the research question) information from verbal text data (Cromley et al., 2024). Classic approaches often use manual coding or simple heuristics based on co-occurrence in small context units (e.g., paper titles) (Henrique et al., 2014) to extract nodes and edges, and hence do not scale to large volumes of text data. For example, in one of our annotated textbook samples, manually converting two pages of information into relational data resulted in hundreds of nodes and edges. Also, as heuristics do not always capture the subtleties and variance in expressions of human language, inaccurate data can be anticipated, which can then lead to inaccurate results and conclusions drawn (Diesner, 2014). Thus, reliable and at least semi-automatic approaches to construct semantic network data from educational verbal text data are highly valuable to education science researchers, as such data allow for testing hypotheses and developing theories. Our study addresses this need.

This work addresses three research questions:

RQ1: How can we systematically identify and encode the nodes and edges that constitute a gold-standard semantic network from textbook data (Textbook)?

RQ2: How accurately can state-of-the-art techniques extract semantic networks from textbook data and transcripts of Think-Aloud data, and what errors are made?

RQ3: How do the structure and properties of semantic networks extracted from students' verbal data correlate with students' learning outcomes and knowledge structures?

We herein capitalize Textbook and Think-Aloud when referring to them as data. To answer RQ1, we collaborated with an educational expert to iteratively construct annotation guidelines (see Section 3) to generate a gold-standard semantic network from Textbook data. For RQ2, we trained and evaluated multiple relation extraction models, including smaller pre-trained language models (PLMs, such as LUKE) and large language models (LLMs, such as Llama), using Textbook and Think-Aloud data. For RQ3, we applied the best-performing models to student Think-Aloud transcripts, built semantic networks, and analyzed their correlation with posttest scores via network metrics and HI-MATT measures.

Our contributions include: (1) a domain-specific annotated dataset for educational relation extraction[1]; (2) empirical comparisons of relation extraction methods that show that BERT-based models outperform LLMs on this task; and (3) insights on using NLP-derived networks to model student learning in educational research, for example by relating network metrics, such as number of nodes, to students' learning outcomes.

2 Background

2.1 NLP in Educational Research

NLP has long been used for semantic analysis and network construction (Sowa et al., 1992; Woods, 1975), but its application in education remains relatively underexplored. Early tools like Why2-Atlas (VanLehn et al., 2002), Coh-Metrix (Graesser et al., 2004), and T-MITOCAR (Pirnay-Dummer, 2006) demonstrated how student text can be analyzed for conceptual structure. More recent efforts have applied NLP techniques such as classification and embeddings to understand student learning processes. For instance,Ostrovsky and Newell (2024) linked verbal and behavioral data to cognitive models, and Lu et al. (2019); Wang et al. (2022a) extracted prerequisite and ISA relations from textbooks. Such relations can be structured into semantic networks for quantitative analysis (Cela et al., 2015). T-MITOCAR's HIMATT framework provides metrics like Surface Matching and Graphical Matching to compare networks (Pirnay-Dummer, 2020). Network-based indicators such as central-

[1] https://github.com/david23145/thinkaloud-relation-data

ities and entropies have been linked to learning outcomes (Lim et al., 2018; Yang et al., 2025). Our work applies NLP methods to build network data and statistical methods to investigate the networks' correlations with student performance data.

2.2 Relation Extraction Models

Specific relation extraction (RE) tasks include 1) relation classification (RC), where entity spans (i.e., the nodes) are known and the goal is to classify their relation, and 2) joint entity and relation extraction (JRE), where models output full subject-predicate-object triples (Sarawagi et al., 2008) and the predicates become the relation. Recent RC models like LUKE (Yamada et al., 2020) and SpanBERT (Joshi et al., 2020) have shown strong results on benchmarks such as FewRel and Re-TACRED (Swarup et al., 2025). JRE models like SPN4RE (Sui et al., 2023), REBEL (Cabot and Navigli, 2021), and OneRel (Shang et al., 2022) can also handle triple extraction, despite the task's higher complexity.

LLMs have also been applied to RE (Wadhwa et al., 2023), and their performance varies depending on prompt engineering and model size. While models like GPT-3.5-Turbo and Llama 3.1 are more flexible and accessible, they don't consistently outperform fine-tuned PLMs (Swarup et al., 2025). In this study, we evaluate both PLMs and LLMs for RE on Textbook and student Think-Aloud data, comparing the models' effectiveness in extracting relations for educational applications.

3 Methodology

This study consists of three tasks: 1. Develop annotation guidelines based on textbook data for extracting relational data from Textbook and Think-Aloud data in the domain of education. 2. Develop and evaluate relation extraction models that extract relational data from a given set of Textbook and Think-Aloud data annotated in Task 1. 3. Apply the best performing models to students' Think-Aloud data to construct semantic networks. Figure 1 illustrates the overall study pipeline, which includes training data processing, manual steps, model development, and the final application of the models to students' Think-Aloud data. All experiments were conducted on a Linux server running Ubuntu 22.04 LTS with an AMD EPYC-Milan CPU (24 vCPUs), 226 GB RAM, and an NVIDIA A100 GPU with 80 GB memory.

3.1 Data

We extracted a semantic network from passages of a textbook in English for an introductory biology course (Textbook Data) (Sadava et al., 2009). This network served as the gold-standard reference network against which we compared students' semantic networks. We also reused verbal data collected at an in-lab Think-Aloud study conducted by one co-author with a background in educational psychology. In the think-aloud study, 77 students were given 40 minutes to study the same textbook passages while saying aloud everything they were thinking while reading. They were also provided with paper and pen to take notes. After the study session, the students moved to another room where they typed everything they could recall from the text. All Think-Aloud data about the textbook was used as the source for constructing individual students' semantic networks. Appendix A illustrates what a participant read and said during the study. In this research, the learning outcome was calculated and normalized based on a post-study written test to evaluate what the students learned from the provided textbook passages. The textbook passages were removed, and participants were asked to type everything they remembered from what they had just read. These typed free recalls were scored for main ideas (2 points each) and supporting facts (1 point each).

3.2 Constructing a Gold-standard Semantic Network from Textbook Passages

We first constructed a gold-standard dataset of semantic triples, namely *subject-verb-object*, by bottom-up coding of textbook data. This gold-standard network was then used to evaluate semantic networks built from the Think-Aloud data. We first conducted an open-coding approach to find all possible triples in the textbook passages. An educational expert then evaluated these extracted triples in terms of false positives, false negatives, and other adjustments needed. Figure 2 shows how a sentence from the textbook can be converted into relational data. Based on the extracted triples, we developed a codebook that summarizes our guidelines and standards for generating semantic triples, such as how to handle content in parentheses. After several rounds of coding and reconciliation among coders, we finalized the codebook. Applying the codebook to selected textbook passages resulted in an inter-coder agreement of 0.83.

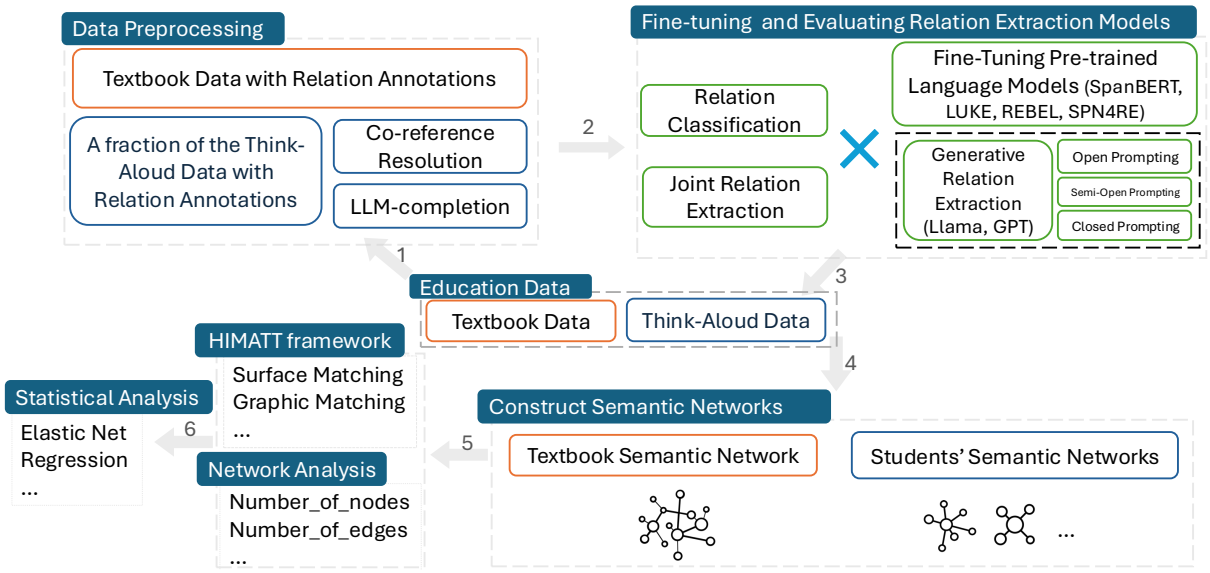


Figure 1: The overall pipeline of the study integrating both Relation Extraction and Semantic Network Analysis.

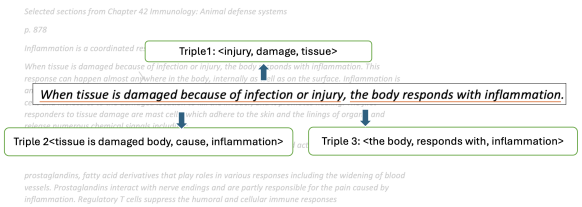


Figure 2: An illustration of how a sentence from the given textbook was converted into relational data.

	Textbook Data		Think-Aloud Data
	Original	Augmented	(Sample)
Sentences	56	1959	1147
Entities	226	310	1251
Relations	254	3058	1215

Table 1: Statistics of the annotated corpus for training and evaluating relation extraction models.

During the annotation, we identified 226 unique predicates from the corpus. Here, by predicate we mean a semantic relation that connects a subject and an object into a relational triple (e.g., inflammation, responds to, infection). These 226 predicates formed the basis for what we refer to as the original relation set. However, we saw overlapping cases and long-tail phenomena of these predicates. For example, some predicates were morphological variants of the same stem or phrases, e.g., "response" and "respond to"; some predicates were synonymous expressions with varying length, e.g., "accelerates" and "speeds up". To normalize these variations, the domain expert consolidated them by mapping them to a set of 46 stems, which the domain expert had also identified from the set of all predicates. We refer to these 46 predicates as the consolidated relations.

After manually extracting relational data to generate a gold-standard semantic network from the considered textbook passages, we augmented the Textbook data to enhance the size of our training data: a domain expert generated relevant synonyms for all annotated subjects, predicates, and objects. Using these synonyms, we created all combinations of subject–predicate–object triples. The resulting relational dataset consists of manually extracted relations and their expansions, totaling 3,058 relational triples, of which 254 are original. To improve the performance of relation extraction models on Think-Aloud data, we also annotated a random sample of the Think-Aloud data, resulting in 1,215 relational triples from 1,147 sentences, which were further used in a data ablation test. Table 1 presents the statistics of the training corpus.

We used the Augmented Textbook Relational dataset, which contains manually extracted and expanded relations from the textbook content, for model development and evaluation. This dataset was randomly split into training, validation, and test sets using an 8:1:1 ratio. This split was used to train and evaluate various relation extraction models. For the best-performing models trained on the Augmented Textbook Relational dataset, we further included the Think-Aloud data (sample) in a data ablation test to examine model performance in extracting relational data from Think-Aloud inputs.

3.3 Evaluating Relation Extraction Models

We systematically tested the performance of state-of-the-art models, from classic ones (e.g., SPN4RE, LUKE, SpanBERT) to recent LLMs, namely Llama and ChatGPT, on the RE task in the context of educational material. We chose SpanBERT and LUKE for relation classification and SPN4RE and REBEL for joint RE (JRE).

We applied various in-context learning strategies to LLM-based RE, and compared three conditions: zero-shot inference (Zero), random retrieval (RandomK), and KNN-based retrieval (TopK). Zero-shot inference provides only the test input without any training examples. RandomK randomly selects K training samples to use as in-context examples, and TopK selects K samples based on the semantic similarity between input and entries in training corpus. For TopK retrieval, we used the "all-mpnet-base-v2" model to generate text embeddings. To assess the impact of different retrieval sizes, we tested for K = 5, 10, and 20.

We designed prompts based on the richness of contextual information provided to the model: no additional information (open), only entity information (ent), only relation information (rel), and information on both entity and relation (ent-rel). Unlike previous work (Swarup et al., 2025), we applied these prompt strategies in both RC and JRE settings, aligning with educational research goals where specific relationships and entities can be of primary interest for assessing student learning outcomes. (see Appendix D)

3.4 Enhancing Think-Aloud Data

One challenge in applying RE to Think-Aloud data is the nature of conversational text: Think-aloud utterances are often grammatically incomplete and may omit key parts of a sentence. For example, a student might say, "Plays many defensive roles," without explicitly stating the subject. If RE models are applied directly, they may fail to identify a complete relation triple from such utterances.

To address this limitation, we improved the Think-Aloud data by using LLMs for coreference resolution (Otmazgin et al., 2022) and sentence completion. This step aimed to recover implicit or omitted information and make the data more suitable for RE. To validate the authenticity and accuracy of the enhanced data, a domain expert reviewed the processed corpus.

We will release the annotated corpus used in this study, including both the Textbook and Think-Aloud relational annotations (excluding the original textbook content due to copyright restrictions), to support future research in educational NLP.

4 Results

We first report on the manually extracted Textbook network that represents knowledge from learning materials. Next, we provide the accuracy results for using pre-trained and large LMs with varying parameter settings for RE from the Augmented Textbook Relation dataset. Finally, we report our findings from applying the best-performing models from the previous step to Think-Aloud data from educational experiments, and evaluate if the extracted relations correlate with students' learning outcomes.

4.1 RQ1: How can we systematically identify and encode the nodes and edges that constitute a gold-standard semantic network from textbook data?

The semantic network manually extracted from textbook passages has a strongly centralized structure with a giant component that consumes 37.56% of all nodes, forming a cohesive "core" cluster of related concepts. With 213 nodes and 226 edges, the network's density of 0.0050 underscores the selective nature of concept linking: most concepts connect to only a few others, resulting in a broadly distributed knowledge space (see Figure 3).

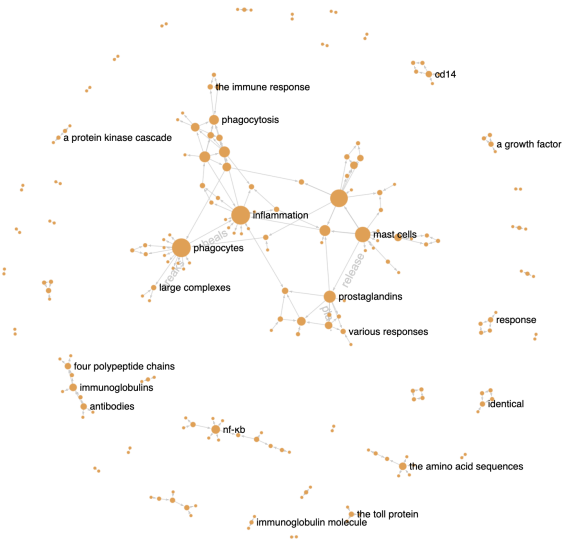


Figure 3: Textbook network representing expected semantic networks of students' Think-Aloud data.

Although nodes that represent key concepts,

such as “inflammation,” “mast cells,” and “phagocytes”, occupy more central network positions, the smooth distribution of centrality values indicates that no single concept monopolizes the network. This result also reflects our concept selection process, which removed terms that were not meaningful in the context of our task. To give an example, from the sentence that merely describe an illustration: "Here we show both diagrammatic (A) and space-filling (B) representations of immunoglobulin.", we removed the `<we-show-representations of immunoglobulin>` relational triple. High stationary entropy (7.25)-which measures how evenly probability mass is distributed across nodes-corroborates this evenness, while elevated transition entropy (1.54)-which captures the diversity of possible next-step transitions-reveals that, on average, each concept affords nearly three equally probable onward transitions—signaling a richly branching but non-hierarchical learning structure. We consider the Textbook network not only as a gold standard for relevant entities and typed relations that are represented in the underlying text material, but also as a point of expert knowledge to compare students’ understanding of the same underlying material to. By comparing students’ individual semantic networks to this expert baseline and correlating them with students’ post-study assessment scores, we can identify conceptual gaps in student learning and gain a better understanding of the relationship between relational data representations and student outcomes.

4.2 RQ2: How accurately can state-of-the-art techniques extract semantic networks from textbook data and transcripts of Think-Aloud data, and what errors are made?

We evaluated four pretrained RE models that include both smaller PLMs and LLMs under two evaluation settings: using 1) the original relation labels and 2) the consolidated relation set (Table 2). This allows us to test if consolidation as a form of data quality management is correlated with improved learning.

For relation classification (RC), SpanBERT and LUKE performed similarly and achieved near-perfect performance (F1 = 0.99) on the consolidated data. Under the original setting, their F1 scores were lower (F1 = 0.67–0.68), indicating reasonably good classification ability even with fine-grained labels. In contrast, REBEL, a generative sequence-to-sequence model, showed substantially lower performance (F1 = 0.55 original, 0.71 consolidated), highlighting the limitations of autoregressive models in classification-style RE tasks.

For the more complex joint relation extraction (JRE) task, SPN4RE consistently outperformed REBEL across both evaluation settings (F1 = 0.76 vs. 0.55 in original), benefiting from its span-based modeling capabilities. Notably, merging relations led to a slight drop in F1 for SPN4RE (0.74), primarily due to decreased recall. These results suggest that while merging simplifies label space, it may introduce ambiguity for JRE models that rely heavily on span boundaries.

We further assessed the generalizability of RE models by applying the best-performing BERT models to a sample of Think-Aloud data. Performance dropped modestly compared to Textbook-only test data (e.g., LUKE F1 = 0.95 on textbook+TA vs. 0.99 textbook-only), indicating the genre shift effect, but models remained robust.

Task	Corpus	Model	Original Relation			Consolidated Relation		
			P	R	F1	P	R	F1
RC	Textbook	SpanBERT	0.68	0.71	0.67	0.99	0.99	0.99
RC	Textbook	LUKE	0.68	0.68	0.68	0.99	0.99	0.99
JRE	Textbook	SPN4RE	0.77	0.75	0.76	0.81	0.68	0.74
JRE	Textbook	REBEL	0.55	0.54	0.55	0.71	0.71	0.71
RC	Textbook+TA Samples	LUKE	0.58	0.60	0.57	0.94	0.98	0.95
JRE	Textbook+TA Samples	SPN4RE	0.68	0.56	0.61	0.75	0.60	0.66

Table 2: Performance of fine-tuned BERT-style models on Relation Classification and Joint Relation Extraction tasks across Textbook and Think-Aloud Data

Appendix B gives performance numbers for two LLMs (Llama 3.1-8B and GPT-3.5-Turbo) across multiple in-context prompting conditions. For the RC task, Llama outperformed GPT-3.5-Turbo across most configurations. The best F1 for Llama (0.64) was obtained using top-k entity+relation information, and GPT-3.5 peaked at 0.47 under similar conditions. Interestingly, adding entity information sometimes reduced performance, suggesting that entity information can be a distraction for GPT-3.5-Turbo and may negatively affect its performance in classifying relations when the goal is to predict the relation between two known entities.

LLMs generally produced lower F1 scores on JRE, with most F1 scores below 0.10. The highest F1 score observed for Llama was 0.09 in a few cases; however, BERT-style models produced substantially higher scores, ranging from 0.66 to 0.81, indicating stronger performance in structured RE. These results show that, for JRE, it remains

challenging for LLMs to achieve performance comparable to more traditional training-based models.

4.3 RQ3: How do semantic relations extracted from students' verbal data correlate with students' learning outcomes and knowledge structure?

4.3.1 Correlations Between Networks and Learning Outcomes

To evaluate how the choice for RE models affects the correlation of network metrics with students' posttest scores, we conducted an ElasticNet regression analysis following the selection of network analysis metrics in Yang et al. (2025). This approach helps reduce multicollinearity and selects the most informative metrics. Table 3 reports the resulting coefficients from each model's best-performing configuration (as established in RQ2), covering both pretrained and large LMs.

A key finding is the consistent correlation of the `number_of_edges` feature across nearly all conditions, indicating that student-generated semantic networks with more connections are generally associated with better posttest outcomes. This suggests that a higher number of edges in expressed knowledge—possibly reflecting greater elaboration and integration—is correlated with student understanding.

However, model-specific patterns also emerged. For instance, LUKE and SPN4RE tended to emphasize simpler structural features, whereas Llama 3.1 and GPT-3.5-Turbo occasionally selected higher-order features such as `avg_pagerank` or entropy-based metrics. Still, the inconsistency in coefficient signs and selection across entropy-based measures (e.g., `stationary entropy`, `transition entropy`) implies they are not robust predictors. Moreover, the effect of TopK in LLM prompting is noteworthy: TopK=5 consistently yielded stronger coefficients and broader feature coverage, indicating that more tightly constrained retrieval settings help generate network structures that better align with posttest performance.

These findings underline the importance of decisions with respect to both RE model architecture and prompt engineering when applying network-based predictive analytics in educational NLP.

4.3.2 Effects on Students' Knowledge Structure Representations

We next examined how different RE models are related to the semantic and structural characteristics (using the definitions in HIMATT) of student knowledge networks by comparing students' networks to the Textbook-based gold-standard network. Our generated students' individual networks exhibited overlaps with the textbook-based gold-standard network, with node overlap ranging from 0 to 0.498 ($M = 0.146 \pm 0.082$) and edge overlap ranging from 0 to 0.143 ($M = 0.011 \pm 0.016$), respectively. Using metrics introduced in Hähnlein and Pirnay-Dummer (2024), we computed similarity scores, i.e., Surface Matching (SUR), Concept Matching (CONC), and Structural Matching (STRU), between the gold-standard network (based on textbook data) and each student network (based on Think-Aloud data).

As shown in Appendix C, the LLM-generated networks (Llama 3.1, GPT-3.5-Turbo) tend to have higher values in Graphical (GRA) and Gamma Matching (GAMMA), reflecting denser structures. In contrast, the traditional models (LUKE, SPN4RE) produced sparser networks with lower matching scores on structural dimensions like Proposition (PROP) and Concept Matching (CONC). This suggests that LLMs may generate more expansive—but potentially less precise—knowledge structures, while fine-tuned PLMs create narrower yet potentially more accurate networks.

Consolidated relation representations generally increased similarity scores, indicating improved alignment with expert knowledge. Still, without gold-standard networks for each student, these results speak more to differences in representational structure than to correctness per se.

We further examined how these structural measures relate to learning outcomes. Table 4 shows Spearman correlations between HIMATT metrics and posttest scores across models and configurations. Surface Matching (SUR) led to the strongest and most consistent positive associations, particularly for Llama 3.1 and GPT-3.5-Turbo, supporting the hypothesis that structural alignment with textbook knowledge reflects better comprehension.

Concept (CONC) and Graphical Matching (GRA) also showed significant positive correlations with posttest scores in several LLM and PLM configurations, especially for LUKE and Llama. Notably, Gamma Matching (GAMMA) was negatively correlated across many conditions, implying that overly dense or incoherent networks may not reflect meaningful learning. These patterns suggest that the students' semantic networks that balance

	SPN4RE	SPN4RE	LUKE	LUKE	Llama 3.1-8B			Llama 3.1-8B	Llama 3.1-8B			GPT3.5-Turbo		
					TopK=5	TopK=10	TopK=20		TopK=5	TopK=10	TopK=20	TopK=5	TopK=10	TopK=20
Relation Type	Original	Consolidated	Original	Consolidated	Original	Original	Original	Consolidated	Original	Original	Original	Consolidated	Consolidated	Consolidated
Task Type	JRE	JRE	RC	RC	JRE	JRE	JRE	JRE	RC	RC	RC	RC	RC	RC
Elastic net														
Best alpha	10.00	0.73	1.89	1.89	0.02	0.92	0.73	0.73	1.17	0.45	0.36	1.17	1.49	1.17
Best L1_ratio	0.50	0.10	0.10	0.10	0.10	1.00	1.00	1.00	0.10	1.00	1.00	0.10	1.00	1.00
Features														
number_of_nodes		0.73	0.62	0.61	-3.56						0.53	0.60		
number_of_edges		1.09	0.44	0.41	7.35	3.30	3.58	3.61	3.30	3.75	1.28	1.21	2.93	3.14
avg_deg_centrality		0.01	0.23	0.15	0.99				0.33	0.93	0.32	0.15		
avg_betweenness_centrality		0.81	-0.28	-0.33	-3.90		0.44	0.42			0.45	0.45		
avg_closeness_centrality		0.46	0.06	0.06	6.41			0.12	0.48		0.60	0.56		
avg_pagerank			-0.52	-0.57	-4.01					-0.11	-0.20	-0.53		
density		0.01	0.23	0.15	0.98				0.00	0.07	0.32	0.15		
reciprocity		-1.20	0.35	0.53	-1.13				0.04	0.63	0.51	0.54		
stationary entropy		-0.46	-0.02	-0.13	1.32						0.12	0.16		
transition entropy		0.74	0.12	0.01	-2.78		0.29				0.54	0.49		

Table 3: ElasticNet coefficients across models and topK configurations.

conceptual breadth with structural coherence are associated with better learning outcomes.

Overall, these analyses illustrate that RE models correlate substantially and systematically with the quality and educational validity of resulting semantic networks. While LLMs appear more capable of producing structures correlated with learning outcomes, PLMs still perform well in certain structural dimensions and may offer better precision. The choice of model, prompt, and merging strategy should thus be guided by the desired balance between coverage and specificity in modeling student knowledge.

5 Discussion

5.1 Advancing Educational Evaluation through Relation Extraction

This study advances work at the intersection of NLP and network analysis as methods brought to the domain of educational evaluation by examining how semantic networks extracted from educational text data can advance the analysis of student knowledge and learning outcomes. While previous efforts have used NLP in educational contexts (Shaik et al., 2022), few have used relation extraction to both Textbook and student Think-Aloud Data, and studied the relationship between these network representations. Our work addresses this gap by implementing and evaluating both pretrained and large LMs for extracting relational triples (subject-predicate-object), representing them as networks, comparing the networks generated from students' Think-Aloud graphs to the gold-standard network extracted from a textbook, and correlating the students' networks with student learning outcomes.

We show to what degree RE-based semantic networks can serve as representations of student understanding. Our regression results suggest that some metrics calculated on these networks—such as the number of edges or surface-level similarity to expert networks—are correlated with student learning outcomes. These findings support the vision of using NLP to model cognitive structures that reflect learning outcomes.

5.2 Model Performance and Its Downstream Impact

Although Llama 3.1 and GPT-3.5-Turbo showed lower RE accuracy rates than PLMs, they often generate denser networks that also better aligned with student posttest performance. This apparent paradox highlights a key insight: traditional evaluation metrics for RE (e.g., precision, Recall, F1) do not necessarily translate to downstream effectiveness. Semantic networks generated from models with lower classification scores still yielded strong associations with learning outcomes when analyzed via network metrics and semantic matching scores.

Additionally, consolidated relation labels enhanced correlation with learning measures across models. Our analysis highlights the importance of evaluating RE models in their application context—not solely based on their extraction accuracy, but also on the content validity of the semantic networks they produce.

5.3 Human-in-the-Loop Annotation and Data Efficiency

We incorporated expert input at multiple stages of the research design—initial relation annotation, augmentation of data with synonyms, and partial Think-Aloud data labeling—to mitigate data scarcity and domain complexity. Data ablation results confirmed that adding a small sample of annotated Think-Aloud data improves model generalizability. This observation emphasizes the value of targeted expert involvement in educational NLP pipelines, where annotated data is often costly and limited.

Moreover, our findings show how robust educa-

	SPN4RE	SPN4RE	LUKE	LUKE	Llama 3.1	Llama 3.1	Llama 3.1	GPT-3.5
Relation Type	Original	Consolidated	Original	Consolidated	Original	Consolidated	Original	Consolidated
Task Type	JRE	JRE	RC	RC	JRE	JRE	RC	RC
Surface Matching (SUR)	0.141	0.199	-0.137	-0.078	0.214**	0.147	0.185**	0.319***
Graphical Matching (GRA)	0.197	0.153	0.12	0.102	0.012	0.144	0.141*	0.196**
Gamma Matching (GAMMA)	0.069	0.123	-0.344**	-0.345**	-0.349***	-0.362**	-0.394***	-0.347***
Structural Matching (STRU)	0.083	0.081	0.05	0.07	0.07	0.072	0.093	0.002
Concept Matching (CONC)	0.098	0.152	0.178	0.168	0.174**	0.161	0.147*	0.217**
Propositional Matching (PROP)	0.101	0.079	-0.005	0.135	0.09	-0.013	-0.016	0.019
Balanced Semantic Matching (BSM)	0.097	0.085	-0.005	0.135	0.095	-0.001	0.001	0.036

Table 4: Spearman Correlation between structural + semantic measures and Posttest by model. Asterisks indicate the level of statistical significance: *p<.05, **p<.01, ***p<.001.

tional NLP tasks are to imperfect model outputs. Even when relation extraction models are noisy, downstream semantic network analyses remain informative, provided that the underlying structure reflects meaningful cognitive or conceptual relationships.

5.4 Implications for NLP in Education

This work offers methodological insights for future NLP applications in educational research. First, RE can bridge raw language and interpretable knowledge structures—particularly when coupled with network analysis. Second, network-based features like the number of edges or surface matching outperformed more abstract graph metrics in providing interpretable signals tied to educational theory. Third, the choice of model (PLM vs. LLM), prompt strategy, and preprocessing steps (e.g., consolidating relations) affects the quality of networks.

Overall, our results advocate for a shift in educational NLP evaluation—from a narrow focus on relation extraction performance to a broader assessment of the representational and explanatory power of generated networks. As RE models continue to improve, we anticipate that their integration into student modeling and learning analytics pipelines will become increasingly impactful.

6 Conclusion

This study demonstrates the feasibility and value of applying relation extraction to educational text data for constructing and analyzing semantic networks. By evaluating both traditional and large LMs on Textbook and Think-Aloud data, we show that model outputs can reveal meaningful insights into student learning outcomes and knowledge structures. Our findings highlight the promise of integrating NLP into educational evaluation while emphasizing the need for further refinement in model evaluation and domain adaptation. This work offers a foundation for building data-driven tools to support personalized learning and educational research.

Limitations

This study has several limitations. First, our evaluation metrics focused on precision, recall, and F-1 scores, while more recent methods, such as topic similarity, were not explored (Jiang et al., 2024). Future work could incorporate these alternatives for a more comprehensive assessment. Second, our pipeline was tested in a single subject domain; applying it to other educational contexts is needed to assess generalizability. Finally, although effective, the current workflow requires manual effort and domain expertise. Packaging the pipeline into an accessible toolkit could support broader adoption and help democratize the use of NLP in educational research.

Ethical Statement

This study was conducted using existing educational data collected in a controlled laboratory setting. All participants were undergraduate students who voluntarily took part in a Think-Aloud study approved by the university's Institutional Review Board (IRB). Informed consent was obtained from all participants prior to the study, and the students were informed that their verbal responses would be anonymized and used for research purposes only.

No personally identifiable information (PII) was used in any stage of data analysis. All student data were de-identified, and the relation extraction models were applied solely to the content of the Think-Aloud transcripts. The study does not involve any sensitive topics or interventions and poses minimal risk to participants.

The purpose of this research is to develop generalizable NLP methods for supporting educational assessment and improving our understanding of student learning processes. We are committed to upholding transparency, reproducibility, and responsi-

ble use of NLP in educational settings.

References

Madelen Bodin. 2012. Mapping university students' epistemic framing of computational physics using network analysis. *Physical Review Special Topics-Physics Education Research*, 8(1):010115.

Jill Burstein. 2003. The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective*, 113121.

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.

Karina L Cela, Miguel Ángel Sicilia, and Salvador Sánchez. 2015. Social network analysis in e-learning environments: A preliminary systematic review. *Educational psychology review*, 27:219–246.

Angelos Charitopoulos, Maria Rangoussi, and Dimitrios Koulouriotis. 2020. On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018. *International Journal of Artificial Intelligence in Education*, 30(3):371–430.

Rianne Conijn, Antoine Van den Beemt, and P Cuijpers. 2018. Predicting student performance in a blended mooc. *Journal of Computer Assisted Learning*, 34(5):615–628.

Jennifer G Cromley, Joseph F Mirabelli, and Andrea J Kunze. 2024. Three applications of semantic network analysis to individual student think-aloud data. *Contemporary Educational Psychology*, 79:102318.

Jana Diesner. 2014. Words and networks: How reliable are network data constructed from text data? In *Roles, trust, and reputation in social media knowledge markets: theory and methods*, pages 81–89. Springer.

Jana Diesner and Kathleen M Carley. 2011a. Semantic networks. *Encyclopedia of social networking*, pages 766–769.

Jana Diesner and Kathleen M Carley. 2011b. Words and networks. *Encyclopedia of social networking*, pages 958–961.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Inka Sara Hähnlein and Pablo Pirnay-Dummer. 2024. Promoting pre-service teachers' knowledge integration from multiple text sources across domains with instructional prompts. *Educational technology research and development*, 72(4):2159–2185.

Trazíbulo Henrique, Inácio de Sousa Fadigas, Marcos Grilo Rosa, and Hernane Borges de Barros Pereira. 2014. Mathematics education semantic networks. *Social Network Analysis and Mining*, 4:1–9.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. *arXiv preprint arXiv:2402.10744*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Jay L Lemke. 1995. Intertextuality and text semantics. *Advances in Discourse Processes*, 50:85–114.

Jay L Lemke. 2012. Analyzing verbal data: Principles, methods, and problems. *Second international handbook of science education*, pages 1471–1484.

Sunghoon Lim, Conrad S Tucker, Kathryn Jablokow, and Bart Pursel. 2018. A semantic network model for measuring engagement and performance in online learning platforms. *Computer Applications in Engineering Education*, 26(5):1481–1492.

Weiming Lu, Yangfan Zhou, Jiale Yu, and Chenhao Jia. 2019. Concept extraction and prerequisite relation learning from educational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9678–9685.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312.

Tehilla Ostrovsky and Ben R Newell. 2024. Verbal reports as data revisited: Using natural language models to validate cognitive models. *Decision*.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. *arXiv preprint arXiv:2209.04280*.

Pablo Pirnay-Dummer. 2020. Knowledge and structure to teach: A model-based computer-linguistic approach to track, visualize, compare and cluster knowledge and knowledge integration in pre-service teachers. In *International Perspectives on Knowledge Integration*, pages 133–154. Brill.

Pablo N Pirnay-Dummer. 2006. *Expertise und Modellbildung-MITOCAR*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg.

Rebecca Rogers. 2004. An introduction to critical discourse analysis in education. In *An introduction to critical discourse analysis in education*, pages 31–48. Routledge.

Cristobal Romero and Sebastian Ventura. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, 3(1):12–27.

David E Sadava, David M Hillis, and H Craig Heller. 2009. *Life: the science of biology*, volume 2. Macmillan.

Sunita Sarawagi and 1 others. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Thanveer Shaik, Xiaohui Tao, Yan Li, Christopher Dann, Jacquie McDonald, Petrea Redmond, and Linda Galligan. 2022. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *Ieee Access*, 10:56720–56739.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Cynthia SQ Siew, Dirk U Wulff, Nicole M Beckage, and Yoed N Kenett. 2019. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019.

John F Sowa and 1 others. 1992. Semantic networks. *Encyclopedia of artificial intelligence*, 2:1493–1511.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE transactions on neural networks and learning systems*.

Anushka Swarup, Tianyu Pan, Ronald Wilson, Avanti Bhandarkar, and Damon Woodard. 2025. Llm4re: A data-centric feasibility study for relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6670–6691.

Kurt VanLehn, Pamela W Jordan, Carolyn P Rosé, Dumisizwe Bhembe, Michael Böttner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Michael Ringenberg, Antonio Roque, and 1 others. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Intelligent Tutoring Systems: 6th International Conference, ITS 2002 Biarritz, France and San Sebastian, Spain, June 2–7, 2002 Proceedings 6*, pages 158–167. Springer.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566.

S Wagner and B Priemer. 2023. Assessing the quality of scientific explanations with networks. *International Journal of Science Education*, pages 1–25.

Haonian Wang, Xinyu Tang, Yurou Liu, and Zhichun Wang. 2022a. Extracting isa relations of concepts from books via weakly supervised learning. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pages 91–98.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022b. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

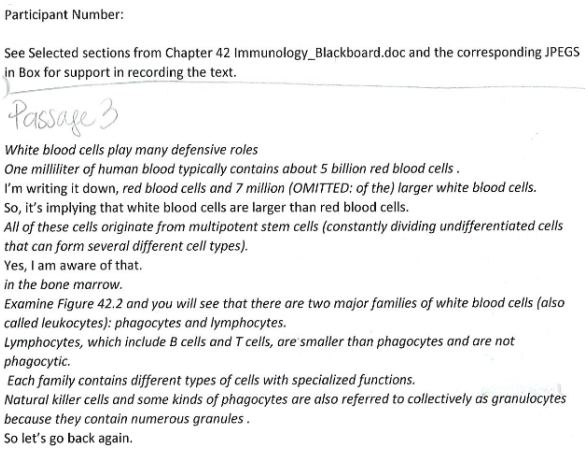
William A Woods. 1975. What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Pingjing Yang, Jennifer Cromley, and Jana Diesner. 2025. Change in eye gaze movement patterns is related to x-ray reading performance. In *International Conference on Artificial Intelligence in Education (**AIED**)*. To appear.

Ming Zhang, Jile Zhu, Zhuo Wang, and Yunfan Chen. 2019. Providing personalized learning guidance in moocs by multi-source data analysis. *World Wide Web*, 22:1189–1219.

A Example Think-Aloud Corpus



Participant Number:

See Selected sections from Chapter 42 Immunology_Blackboard.doc and the corresponding JPEGS in Box for support in recording the text.

Passage 3

White blood cells play many defensive roles
One milliliter of human blood typically contains about 5 billion red blood cells .
I'm writing it down, *red blood cells and 7 million (OMITTED: of the) larger white blood cells.*
So, it's implying that white blood cells are larger than red blood cells.
All of these cells originate from multipotent stem cells (constantly dividing undifferentiated cells that can form several different cell types).
Yes, I am aware of that.
in the bone marrow.
Examine Figure 42.2 and you will see that there are two major families of white blood cells (also called leukocytes): phagocytes and lymphocytes.
Lymphocytes, which include B cells and T cells, are smaller than phagocytes and are not phagocytic.
Each family contains different types of cells with specialized functions.
Natural killer cells and some kinds of phagocytes are also referred to collectively as granulocytes because they contain numerous granules .
So let's go back again.

Figure 4: An example of in-lab Think-aloud corpus.

B Large Language Model Results

Condition	Llama-3.1						GPT-3.5-Turbo					
	Original predictes			Consolidated predictes			Original predictes			Consolidated predictes		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RC_Ent Info_zero_0	0.35	0.36	0.36	0.32	0.35	0.33	0.34	0.40	0.37	0.05	0.06	0.05
RC_Ent Info_topk_Mean	0.49	0.64	0.56	0.24	0.28	0.26	0.40	0.47	0.43	0.17	0.20	0.18
RC_Ent Info_randomk_Mean	0.42	0.43	0.43	0.33	0.33	0.33	0.30	0.25	0.27	0.05	0.05	0.05
RC_Rel Info_zero_0	0.35	0.42	0.39	0.27	0.30	0.28	0.36	0.38	0.37	0.08	0.06	0.07
RC_Rel Info_topk_Mean	0.47	0.66	0.55	0.14	0.18	0.16	0.34	0.46	0.39	0.55	0.48	0.51
RC_Rel Info_randomk_Mean	0.37	0.43	0.40	0.17	0.19	0.18	0.35	0.23	0.28	0.15	0.08	0.11
RC_Ent+Rel Info_zero_0	0.30	0.35	0.32	0.21	0.27	0.24	0.27	0.36	0.31	0.07	0.09	0.08
RC_Ent+Rel Info_topk_Mean	0.54	0.63	0.58	0.16	0.17	0.17	0.35	0.47	0.40	0.28	0.29	0.28
RC_Ent+Rel Info_randomk_Mean	0.42	0.41	0.41	0.22	0.22	0.22	0.31	0.28	0.29	0.08	0.06	0.07
RC_No Info_zero_0	0.34	0.38	0.36	0.35	0.38	0.36	0.37	0.30	0.33	0.04	0.03	0.04
RC_No Info_topk_Mean	0.48	0.64	0.55	0.19	0.23	0.21	0.44	0.44	0.44	0.32	0.29	0.31
RC_No Info_randomk_Mean	0.38	0.41	0.40	0.30	0.33	0.31	0.38	0.28	0.32	0.06	0.05	0.05
JRE_Ent Info_zero_0	0.07	0.17	0.10	0.06	0.19	0.09	0.07	0.20	0.10	0.01	0.02	0.01
JRE_Ent Info_topk_Mean	0.11	0.24	0.15	0.06	0.18	0.09	0.09	0.24	0.13	0.02	0.05	0.02
JRE_Ent Info_randomk_Mean	0.09	0.19	0.12	0.06	0.17	0.09	0.07	0.19	0.10	0.01	0.03	0.01
JRE_Rel Info_zero_0	0.06	0.18	0.09	0.06	0.17	0.09	0.07	0.20	0.10	0.01	0.03	0.02
JRE_Rel Info_topk_Mean	0.10	0.25	0.14	0.04	0.12	0.06	0.08	0.22	0.12	0.03	0.07	0.04
JRE_Rel Info_randomk_Mean	0.07	0.17	0.10	0.05	0.13	0.07	0.07	0.17	0.10	0.02	0.04	0.02
JRE_Ent+Rel Info_zero_0	0.09	0.21	0.13	0.05	0.17	0.08	0.09	0.26	0.13	0.01	0.02	0.01
JRE_Ent+Rel Info_topk_Mean	0.13	0.22	0.16	0.04	0.13	0.07	0.10	0.29	0.15	0.03	0.07	0.04
JRE_Ent+Rel Info_randomk_Mean	0.11	0.17	0.13	0.06	0.16	0.08	0.08	0.21	0.11	0.02	0.04	0.02
JRE_No Info_zero_0	0.07	0.15	0.10	0.05	0.17	0.08	0.05	0.13	0.07	0.01	0.02	0.01
JRE_No Info_topk_Mean	0.12	0.21	0.15	0.06	0.18	0.09	0.08	0.21	0.11	0.01	0.04	0.02
JRE_No Info_randomk_Mean	0.08	0.15	0.11	0.05	0.14	0.07	0.05	0.15	0.08	0.01	0.03	0.02

Table 5: Performance of LLaMA 3.1 and GPT-3.5-Turbo under various in-context prompting conditions on Relation Extraction tasks.

C Semantic Network Metrics

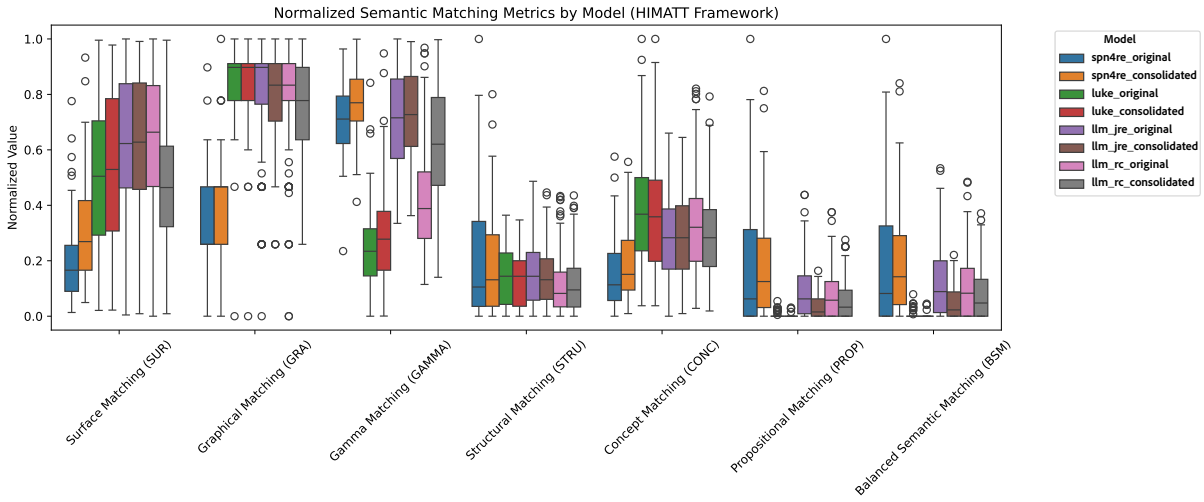


Figure 5: Normalized structural and semantic matching scores (e.g., SUR, GRA, CONC) across all model configurations using the HIMATT framework.

D Prompt templates

You are an expert in joint entity and relation extraction. Your task is to:

1. Identify all entity pairs mentioned in the sentence.
2. For each entity pair, classify the relation between them.

If no relation exists between a pair, label it as "No Relation".

Here are all the annotated entities/relations/entities and relations from training corpus:
$NONE | ENTITIES | RELATIONS | ENTITIES AND RELATIONS$ <- meta information

Here are some examples for your consideration.
$EXAMPLES$ <- this can be 0,5, 10, or 20 examples based on different in-context learning strategies

Here is your task:

Context: **$TEXT$** <- include sentence
Given the context, the entity and relation triplets are:

Figure 6: Prompt template used for joint relation extraction (JRE) with varying levels of entity and relation information and different in-context learning strategies.

You are an expert in relation classification. Given a sentence and two entities, identify the relationship between them from the predefined relation types. If no relation exists, output "No Relation".

Here are all the annotated entities/relations/entities and relations from training corpus:
$NONE | ENTITIES | RELATIONS | ENTITIES AND RELATIONS$ <- meta information

Here are some examples for your consideration.
$EXAMPLES$ <- this can be 0, 5, 10, or 20 examples based on different in-context learning strategies

Here is your task:

Context: **$TEXT$** <- include sentence and entities
Given the context, the entity and relation triplets are:

Figure 7: Prompt Template for Relation Classification (RC) with varying levels of entity and relation information and different in-context learning strategies.