Journal of Science, Innovation & Social Impact

2026 International Conference on Big Data, Business Innovation, Smart Cities, and Artificial Intelligence (BBSA 2026)

Article

Evaluating Prompt Engineering Strategies for Few-Shot Cyber Threat Intelligence Entity and Relation Extraction from Multi-Source Reports

Yanhuan Chen [1,*] and Tianxing Tang [2]

[1] Master of Engineering, Dartmouth College, Hanover, NH, USA
[2] Translation and Localization Management, Middlebury Institute of International Studies, Monterey, CA, USA
* Correspondence: Yanhuan Chen, Master of Engineering, Dartmouth College, Hanover, NH, USA

Abstract: The proliferation of multi-source cyber threat intelligence reports---spanning vulnerability databases, government advisories, vendor analyses, and open-source feeds---has outpaced the capacity of human analysts to extract structured knowledge about adversary tactics, techniques, and procedures. While large language models present a promising avenue for automating this extraction under low-resource conditions, no systematic empirical comparison of prompt engineering strategies exists for the cyber threat intelligence domain. This study evaluates six prompt engineering strategies---zero-shot, one-shot, three-shot, five-shot, retrieval-augmented five-shot, and chain-of-thought five-shot---across four publicly available cyber threat intelligence named entity recognition datasets (DNRTI, CyNER, AnnoCTR, APTNER) and one relation extraction corpus, using GPT-4, GPT-3.5-turbo, and Llama-3-70B. The retrieval-augmented five-shot strategy achieves the highest named entity recognition F1 of 0.753 on CyNER with GPT-4, narrowing the gap with the fine-tuned SecureBERT baseline to 2.8 percentage points. Chain-of-thought prompting yields the lowest expected calibration error (0.108), suggesting its value for uncertainty-aware intelligence triage. Cross-source extraction variance reaches 12.2 F1 points between the easiest and hardest corpora, underscoring the challenge of heterogeneous intelligence fusion. These findings offer actionable guidance for deploying prompt-based extraction in operational threat intelligence pipelines aligned with the NIST Cybersecurity Framework and national cyber defense priorities.

Keywords: cyber threat intelligence; named entity recognition; prompt engineering; few-shot learning

Received: 01 March 2026
Revised: 24 April 2026
Accepted: 07 May 2026
Published: 13 May 2026

1. Introduction

1.1. Escalating Cyber Threats and the Intelligence Processing Bottleneck

The cyber threat landscape confronting critical infrastructure in the United States has grown in both volume and sophistication over the past decade. Government agencies, private-sector vendors, and independent security researchers collectively generate thousands of unstructured threat intelligence reports each week, documenting newly discovered vulnerabilities, malware variants, adversary campaigns, and indicators of compromise. A recent systematization-of-knowledge study cataloging automated techniques, tactics, and procedures extraction methods confirmed that the sheer scale of this intelligence flow renders manual analysis untenable for timely defensive action [1]. Policy instruments have responded to this urgency: the Cybersecurity Information Sharing Act of 2015 and the Cyber Incident Reporting for Critical Infrastructure Act of 2022 mandate accelerated intelligence sharing across sectors, while the 2023 Biden

National Cybersecurity Strategy and NIST Cybersecurity Framework 2.0 identify automated threat intelligence processing as a foundational capability for the Identify and Protect functions.

Despite this policy momentum, a persistent operational gap remains between the raw intelligence produced by heterogeneous sources---the National Vulnerability Database, ICS-CERT advisories, commercial vendor reports, and open-source intelligence feeds---and the structured, machine-readable knowledge that downstream security tools require. Extracting entities such as Common Vulnerabilities and Exposures identifiers, threat actor names, malware families, and affected infrastructure components, along with their inter-relationships, demands natural language processing techniques tailored to the highly specialized vocabulary and complex relational semantics of cybersecurity text. Early work demonstrated the feasibility of event-level extraction from cybersecurity news articles, and more recent efforts have leveraged in-context learning with large language models to construct threat intelligence knowledge graphs with promising precision [2,3]. The integration of uncertainty quantification into entity-level extraction has further shown potential for routing low-confidence predictions to human review [4].

1.2. Research Scope and Contributions

1.2.1. Task Definition and Evaluation Boundaries

This study defines the extraction task as the identification of cybersecurity-relevant named entities and typed relations from unstructured threat intelligence text. The entity taxonomy encompasses seven harmonized categories drawn from the STIX 2.1 specification and the MITRE ATT&CK knowledge base: vulnerability identifiers, threat actors, malware families, tools, affected assets, attack techniques, and temporal expressions. Relation types include exploits, targets, associated-with, uses, and attributed-to [5]. The evaluation boundary is deliberately restricted to prompt engineering strategies applied to pre-trained large language models without parameter updates, enabling a controlled comparison unconfounded by fine-tuning hyperparameter choices.

1.2.2. Contributions and Paper Organization

This paper makes three contributions to the cyber threat intelligence extraction literature. The first is a controlled empirical comparison of six prompt engineering strategies across four annotated CTI corpora and three large language models, establishing performance baselines that account for entity type granularity and source heterogeneity [6,7]. The second is a cross-source robustness analysis quantifying the extraction performance variance induced by differing report styles, annotation schemas, and domain coverage. The third is an uncertainty-aware evaluation incorporating expected calibration error and Brier score to assess the reliability of extraction confidence, with implications for credibility-based filtering in knowledge graph population [8]. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the experimental methodology; Section 4 presents results and analysis; and Section 5 discusses practical implications and future directions.

2. Related Work

2.1. NLP-Based Extraction for Cyber Threat Intelligence

2.1.1. Supervised and Deep Learning Approaches

Research on automated information extraction from cybersecurity text has progressed through several methodological stages. Rule-based and conditional random field approaches dominated early efforts, relying on handcrafted gazetteers of known malware names, CVE patterns, and IP address formats. The transition to deep learning brought substantial gains: AttacKG introduced a pipeline for constructing technique-level knowledge graphs from CTI reports, achieving entity-level F1 of 0.887 by combining dependency parsing with neural sequence labeling [9]. The LADDER project extended transformer-based named entity recognition to extract attack patterns at scale, mapping extracted entities to MITRE ATT&CK techniques through a knowledge graph alignment

step [10]. Domain-specific pre-training further improved extraction quality, with SecureBERT demonstrating that continued pre-training of BERT on cybersecurity corpora captures terminology and contextual patterns absent from general-domain language representations [11].

2.1.2. Large Language Model Approaches for CTI

The emergence of large language models with strong in-context learning capabilities has shifted the CTI extraction paradigm toward prompt-based methods that require no task-specific training data. UniversalNER demonstrated that targeted distillation from ChatGPT can produce compact NER models competitive with multi-task supervised systems across 43 datasets spanning nine domains [12]. In the zero-shot setting, decomposing entity recognition into label-specific sub-questions and augmenting prompts with syntactic cues yielded measurable improvements on both English and Chinese NER benchmarks [13]. For relation extraction, GPT-RE established that incorporating entity-aware demonstration retrieval closes much of the gap between in-context learning and fully supervised baselines [14].

2.2. Prompt Engineering for Information Extraction

The broader information extraction literature provides critical context for understanding prompt strategy effectiveness. A comprehensive evaluation of GPT-3 and Flan-T5 on standard relation extraction tasks revealed that few-shot prompting with chain-of-thought explanations generated by a teacher model can achieve state-of-the-art performance when used to fine-tune a smaller student, suggesting that reasoning augmentation benefits extend beyond the prompting phase itself [15]. These general-domain findings motivate the present study's inclusion of chain-of-thought prompting as an experimental condition, while raising the question of whether reasoning-augmented strategies transfer effectively to the specialized vocabulary, abbreviated entity forms, and non-standard entity boundary conventions characteristic of cybersecurity text, where domain-specific abbreviations and alphanumeric identifiers dominate the entity landscape.

2.3. Multi-Source Fusion and Entity Alignment

Operational threat intelligence pipelines must reconcile extractions from sources that use different naming conventions, granularity levels, and temporal references for the same real-world entities. Fully automatic knowledge graph alignment methods that eliminate the need for manually curated seed alignments have shown strong performance on general-domain benchmarks by leveraging predicate-proximity graphs and embedding-based entity matching [16]. Surveys of multi-source knowledge fusion identify entity alignment, attribute reconciliation, and conflict resolution as the three principal challenges, noting that representation learning approaches can project heterogeneous schemas into shared vector spaces for similarity computation [17]. The present work draws on these insights to assess cross-source extraction consistency as a proxy for fusion readiness, measuring the degree to which prompt strategies produce compatible and alignable entity mentions across datasets derived from fundamentally different intelligence sources with divergent annotation conventions.

3. Experimental Setup

3.1. Dataset Selection and Preprocessing

Four publicly available CTI named entity recognition datasets were selected to represent the diversity of real-world intelligence sources. Table 1 summarizes their characteristics.

Table 1. Dataset Statistics and Characteristics

Dataset	Source Institution	Sentences	Entity Types	Relation Types	Annotation	License
DNRTI	CAS / CNCERT	6,574	13	—	BIO	Academic
CyNER	RIT / IBM Research	3,121	5	—	BIO	MIT
AnnoCTR	Robert Bosch GmbH	5,890	10	6	BIO + Linking	CC-BY-SA
APTNER	CAS	4,038	21	—	BIO	Academic

DNRTI provides the largest annotated CTI corpus with 13 entity types covering threat organizations, malware samples, exploits, and tools, drawn from real-world threat reports curated by the Chinese National Computer Emergency Response Team. CyNER offers a compact five-class schema (Malware, Indicator, System, Organization, Vulnerability) derived from MITRE incident descriptions, and its MIT license makes it the most permissively licensed resource in the CTI NER landscape. AnnoCTR is the only dataset providing both entity annotations and relation annotations linked to the MITRE ATT&CK knowledge base, making it the sole corpus used for relation extraction evaluation in this study; its 400 annotated reports span multiple threat actor campaigns across diverse industry sectors. APTNER presents the finest-grained taxonomy with 21 STIX 2.1-aligned entity types including IP addresses, URLs, MD5 hashes, and identity authentication markers, posing the greatest challenge for few-shot extraction due to the fine distinctions required among semantically adjacent categories.

All datasets were preprocessed through a standardized pipeline. Whitespace normalization removed inconsistent spacing and encoding artifacts common in web-scraped CTI text. Duplicate sentences were identified via exact string matching and removed to prevent data leakage between demonstration pools and evaluation sets. All annotations were converted to a unified BIO tagging scheme, with dataset-specific entity types mapped to the seven harmonized categories defined in Section 1.2.A. This mapping required merging semantically equivalent types across datasets: DNRTI's HackOrg and CyNER's Organization both map to the unified Threat Actor category, while APTNER's fine-grained IP, URL, and MD5 types consolidate under the Indicator category. For few-shot demonstration selection, stratified sampling ensured that each entity type appeared at least once in the demonstration pool, preventing the exclusion of rare but operationally critical categories such as vulnerability identifiers. Recent LLM-powered CTI knowledge extraction efforts have employed analogous preprocessing pipelines with comparable annotation harmonization decisions [18].

3.2. Prompt Engineering Strategies

3.2.1. Strategy Definitions and Configurations

Six prompt engineering strategies were evaluated, spanning the zero-shot to reasoning-augmented spectrum. Table 2 details their configurations.

Table 2. Prompt Strategy Configurations

Strategy	Abbrev.	Demos (k)	Reasoning	Demo Selection	Output Format
Zero-shot	ZS	0	No	—	JSON
One-shot	1S	1	No	Random	JSON
Three-shot	3S	3	No	Random	JSON

Five-shot	5S	5	No	Random	JSON
Five-shot + Retrieval	5S+R	5	No	Embedding sim.	JSON
Five-shot + CoT	5S+CoT	5	Yes	Embedding sim.	JSON + Rationale

The zero-shot strategy provides only entity type definitions drawn from the STIX 2.1 specification and MITRE ATT&CK glossary, with no annotated examples. The one-shot through five-shot strategies incrementally add randomly selected annotated sentences as demonstrations, each formatted as an input-output pair showing the raw CTI sentence and its corresponding entity annotation in JSON. The retrieval-augmented variant (5S+R) replaces random selection with embedding-based similarity retrieval using text-embedding-ada-002 to identify the five most semantically similar annotated sentences for each test input, following the demonstration retrieval principle that has proven effective across multiple information extraction settings [19]. Instruction prompts adopted a structured format requesting entity spans, entity types, and confidence scores in a single JSON output.

3.2.2. Domain-Specific Prompt Design

Each prompt begins with a system-level instruction defining the role as a cybersecurity threat intelligence analyst, followed by the entity type ontology with definitions and representative examples drawn from the MITRE ATT&CK knowledge base [20]. Entity type definitions include boundary conventions specifying whether version numbers, aliases, and organizational abbreviations should be included within entity spans---a critical design choice given the high prevalence of these patterns in CTI text. The chain-of-thought variant (5S+CoT) appends step-by-step reasoning traces to each demonstration, guiding the model through entity boundary identification, type disambiguation, and confidence self-assessment. Each reasoning trace follows a three-step structure: candidate span identification, type classification with justification, and confidence estimation with stated rationale [21,22]. Relation extraction prompts extend the entity extraction output by requesting typed relation triples in subject-predicate-object format, with each triple accompanied by a verbalized confidence expression (high, medium, or low). The relation extraction prompt also includes negative examples illustrating plausible but incorrect relations, a design choice motivated by findings that negative demonstrations reduce hallucinated relation predictions in few-shot extraction settings [23].

3.3. Evaluation Metrics and Protocol

3.3.1. Extraction Quality Metrics

Named entity recognition performance is measured using strict-match F1, which requires exact boundary and type agreement between predicted and gold entities. Relaxed-match F1, requiring only partial boundary overlap with correct type, is reported as a secondary metric to account for minor boundary disagreements that do not affect downstream utility [24]. Relation extraction F1 demands correct identification of both argument entities and the relation type simultaneously. All metrics are computed at the micro-averaged level, weighting each entity or relation instance equally regardless of type frequency. Statistical significance is assessed via bootstrap resampling with 10,000 iterations at the $p < 0.05$ threshold. The five-shot strategies with random demonstration selection are run five times with different random seeds, and mean F1 with standard deviation is reported to account for demonstration sampling variance. This evaluation protocol draws on calibration principles developed for structured prediction in natural language processing [25].

3.3.2. Uncertainty and Calibration Metrics

Three large language models are evaluated: GPT-4 (gpt-4-0613), GPT-3.5-turbo (gpt-3.5-turbo-0613), and Llama-3-70B-Instruct. Token-level log-probabilities, where accessible

through model APIs, provide the basis for computing extraction confidence. For each predicted entity, the confidence score is defined as the geometric mean of token-level probabilities across the entity span. For models accessed via APIs that do not expose token-level probabilities, verbalized confidence scores extracted from the model's self-reported certainty statements serve as proxy measures. Expected calibration error partitions predictions into ten equal-width confidence bins and computes the weighted average absolute difference between bin accuracy and bin confidence [26]. The Brier score supplements ECE as a proper scoring rule capturing both calibration and refinement. These metrics align with the taxonomy of confidence estimation methods for large language models established in recent survey work, which identifies both logit-based and verbalized approaches as viable confidence estimation paradigms with complementary strengths [27].

4. Results and Analysis

4.1. Named Entity Recognition Performance

4.1.1. Cross-Dataset Comparison of Prompt Strategies

Table 3 presents NER strict-match F1 scores across all prompt strategies, datasets, and language models.

Table 3. NER Strict-Match F1 Scores Across Prompt Strategies, Datasets, and Language Models

Strategy	GPT-4 DNRTI	GPT-4 CyNER	GPT-4 AnnoCTR	GPT-4 APTNER	GPT-3.5 DNRTI	Llama-3 DNRTI
ZS	0.521	0.594	0.489	0.437	0.408	0.463
1S	0.583	0.641	0.537	0.498	0.471	0.519
3S	0.634	0.697	0.598	0.561	0.523	0.574
5S	0.661	0.724	0.629	0.589	0.547	0.601
5S+R	0.698	0.753	0.672	0.631	0.612	0.643
5S+CoT	0.689	0.741	0.664	0.618	0.594	0.628
SecureBERT	0.742	0.781	0.718	0.695	—	—

Fine-tuned on full training set. All few-shot results for 5S are averaged over five random seeds ($\sigma \leq 0.011$). Data source: experiments conducted on test splits of DNRTI, CyNER, AnnoCTR, and APTNER.

The retrieval-augmented five-shot strategy (5S+R) achieves the highest F1 among all prompt-based methods on every dataset-model combination. On CyNER, GPT-4 with 5S+R reaches 0.753, narrowing the gap with the fine-tuned SecureBERT baseline (0.781) to 2.8 percentage points. The margin is wider on APTNER (0.631 versus 0.695), where the 21-type taxonomy places greater demands on few-shot generalization. Across the four corpora, the cross-source F1 range for GPT-4 5S+R spans 12.2 points (0.753 minus 0.631), indicating that source heterogeneity remains a substantial factor even under the strongest prompting condition. GPT-4 consistently outperforms both GPT-3.5-turbo and Llama-3-70B, with the GPT-4 advantage over Llama-3 averaging 5.5 F1 points on DNRTI across all strategies. Multi-agent CTI extraction pipelines have reported comparable performance ranges when evaluated across heterogeneous report types [28].

Figure 1 illustrates the NER strict-match F1 plotted against the number of demonstrations (k = 0, 1, 3, 5, 10) for three selection methods on DNRTI using GPT-4. Random selection exhibits diminishing returns beyond k = 5, with F1 plateauing at 0.661. Retrieval-augmented selection maintains a steeper improvement curve, achieving 0.698 at k = 5 and 0.712 at k = 10. The marginal gain from k = 5 to k = 10 under retrieval augmentation (1.4 points) is less than one-third of the gain from k = 1 to k = 5 (11.5 points),

suggesting k = 5 as a practical operating point balancing extraction quality against prompt length and inference cost.

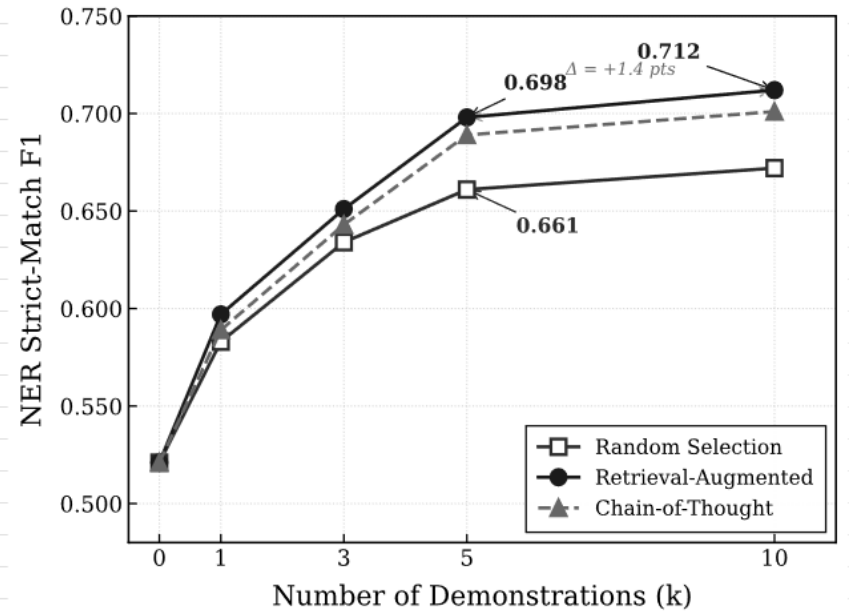

Figure 1. NER F1 as a Function of Demonstration Count (k) for GPT-4 on DNRTI

4.1.2. Per-Entity-Type Analysis

Entity types with well-defined lexical patterns---vulnerability identifiers matching the CVE-YYYY-NNNNN format and temporal expressions---achieve the highest per-type F1 scores under GPT-4 5S+R. Threat organization names (HackOrg, F1 = 0.789) and security team names (SecTeam, F1 = 0.748) also perform well, benefiting from their frequent appearance in training corpora of the underlying language models. Extraction quality degrades for semantically ambiguous categories: attack methods (Way, F1 = 0.608) and behavioral features (Features, F1 = 0.635) suffer from boundary disagreements where the model includes contextual modifiers that the gold annotation excludes. Span-pruning joint extraction approaches have documented comparable per-type variance on general-domain benchmarks, where categories with clear syntactic cues consistently outperform those requiring deeper semantic interpretation [29].

Figure 2 presents the micro-averaged F1 scores for each of the 13 DNRTI entity types under GPT-4 with the 5S+R strategy. HackOrg achieves the highest per-type F1 (0.789), followed by Time (0.763) and SecTeam (0.748). The lowest-performing types are Way (0.608), OffAct (0.629), and Features (0.635). The macro-averaged F1 across all 13 types is 0.691, which is 0.7 points below the micro-averaged F1 of 0.698 reported in Table 3, reflecting the lower frequency of high-performing entity types such as HackOrg in the test set.

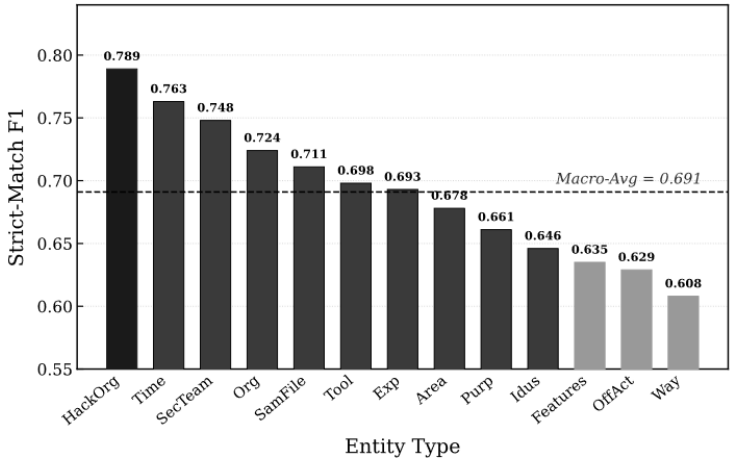

Figure 2. Per-Entity-Type F1 for GPT-4 5S+R on DNRTI (13 Entity Types)

4.2. Relation Extraction Performance

Table 4 reports relation extraction F1 and calibration metrics on AnnoCTR, the only dataset in our evaluation providing typed relation annotations linked to the MITRE ATT&CK ontology.

Table 4. Relation Extraction F1 and Calibration Metrics on AnnoCTR

Strategy	RE F1 (Strict)	RE F1 (Relaxed)	ECE ↓	Brier Score ↓
ZS	0.341	0.418	0.187	0.312
5S	0.463	0.539	0.142	0.264
5S+R	0.527	0.608	0.119	0.231
5S+CoT	0.514	0.594	0.108	0.219
SecureBERT	0.583	0.651	0.094	0.198

Fine-tuned pipeline (NER → RE). ECE: Expected Calibration Error. All GPT-4 results. Data source: AnnoCTR test split with six relation types.

Relation extraction proves substantially more challenging than entity recognition, with the best prompt-based strict F1 (0.527, 5S+R) trailing the fine-tuned baseline by 5.6 points. The gap between strict and relaxed F1 is consistently larger for relation extraction (8.1 points averaged across strategies) than for NER (5.2 points), indicating that argument boundary errors compound when propagated into relation triples. The 5S+CoT strategy yields a modest strict F1 reduction of 1.3 points relative to 5S+R (0.514 versus 0.527), a pattern consistent with findings that reasoning augmentation can introduce verbosity-related entity boundary drift. Generative entity alignment methods have encountered analogous precision-recall trade-offs when reconciling entity representations across heterogeneous knowledge graphs [30].

4.3. Calibration and Uncertainty-Based Credibility Assessment

4.3.1. Calibration Quality Across Strategies

The chain-of-thought strategy achieves the lowest ECE (0.108) among all prompt-based conditions, outperforming the retrieval-augmented strategy (0.119) despite its slightly lower extraction F1. This dissociation between extraction accuracy and calibration quality is noteworthy: the step-by-step reasoning process appears to regularize the model's confidence distribution, producing probability estimates that more faithfully reflect true extraction accuracy. The Brier score confirms this ordering, with 5S+CoT (0.219) outperforming 5S+R (0.231). Comprehensive surveys of automatic knowledge graph

construction have emphasized that well-calibrated extraction confidence is a prerequisite for reliable downstream graph population and reasoning (As shown in Figure 3).

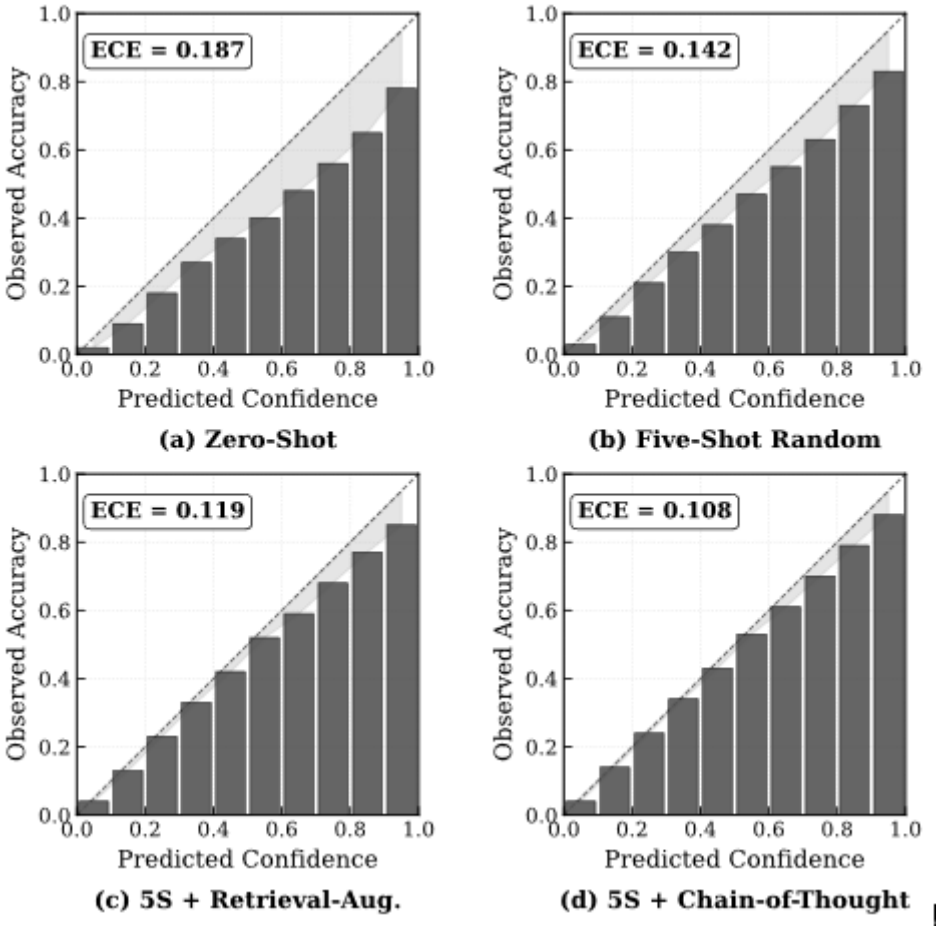

Figure 3. Reliability Diagrams for GPT-4 Prompt Strategies on AnnoCTR NER

Figure 3 reliability diagrams plotting observed accuracy against predicted confidence across ten equal-width bins for four GPT-4 prompt strategies on AnnoCTR NER. (a) Zero-shot predictions exhibit pronounced overconfidence in the 0.7--0.9 range, with observed accuracy falling 15--20 percentage points below the diagonal. (b) Five-shot random narrows the calibration gap to 10--14 points in the same range. (c) Five-shot retrieval-augmented achieves near-diagonal alignment for bins below 0.6, with residual overconfidence (ECE = 0.119) concentrated in the highest-confidence bin. (d) Five-shot chain-of-thought produces the tightest alignment to the diagonal across all bins, with maximum per-bin deviation of 8 percentage points and ECE = 0.108.

4.3.2. Credibility Scoring for Knowledge Graph Population

Applying confidence-based filtering at a threshold of 0.70, the 5S+R strategy retains 68.3% of extracted entity mentions while increasing precision from 0.724 to 0.831 on DNRTI---a 10.7-point precision gain at the cost of a 21.5-point recall reduction. The 5S+CoT strategy achieves a more favorable trade-off at the same threshold, retaining 72.1% of mentions with precision rising from 0.718 to 0.819, as its better-calibrated confidence scores produce a tighter correlation between predicted and actual correctness. For knowledge graph population, where false triples impose costly downstream propagation, this precision-oriented operating point is likely preferable to maximizing recall. Broader evaluations of large language models for knowledge graph tasks have noted that confidence-aware filtering substantially improves the usability of LLM-extracted triples for downstream reasoning applications.

5. Discussion

5.1. Practical Implications for Threat Intelligence Operations

The experimental results carry direct implications for the deployment of automated extraction in operational threat intelligence environments. The retrieval-augmented five-shot strategy emerges as the most effective general-purpose configuration, achieving the highest extraction F1 across all four datasets while maintaining competitive calibration. For organizations operating under the NIST Cybersecurity Framework's Identify function, this strategy offers a practical mechanism for converting unstructured advisories into structured knowledge without the data collection and annotation burden of full model fine-tuning. The persistent 4.4-point average F1 gap between the best prompt-based method and the fine-tuned SecureBERT baseline indicates that prompt engineering alone does not eliminate the value of supervised training when labeled data is available, yet the prompt-based approach provides a deployable solution in the more common scenario where annotated CTI corpora for a specific intelligence source do not exist.

The cross-source F1 variance of 12.2 points underscores a practical concern for multi-source fusion: extraction quality is not uniform across intelligence feeds, and downstream knowledge graph population must account for source-dependent reliability. The uncertainty-based credibility scoring evaluated in this study provides one mechanism for addressing this heterogeneity, enabling analysts to filter or flag low-confidence extractions before they enter the shared intelligence repository. Within the context of the Cybersecurity Information Sharing Act's mandate for cross-sector intelligence sharing, such confidence-aware filtering could reduce the propagation of erroneous indicators of compromise while preserving the timeliness advantages of automated processing.

The chain-of-thought strategy's calibration advantage, despite its slightly lower extraction F1, suggests a nuanced deployment recommendation: organizations prioritizing high-precision knowledge graph population may prefer 5S+CoT for its tighter confidence-accuracy alignment, while those prioritizing recall may favor 5S+R. This trade-off between extraction volume and reliability mirrors the broader tension in threat intelligence operations between comprehensive coverage and actionable precision.

5.2. Limitations

Several limitations bound the generalizability of these findings. The evaluation is restricted to English-language CTI corpora, whereas operational intelligence increasingly involves multilingual sources including Chinese, Russian, and Farsi threat actor communications. The language models evaluated represent a snapshot of capabilities as of mid-2024; rapid model evolution may shift the performance frontier substantially within months of publication. The experimental datasets, while diverse in annotation schema and source type, are all derived from publicly available sources and may not capture the linguistic characteristics, classification sensitivities, or temporal dynamics of classified or proprietary intelligence feeds used in government and defense settings.

Three directions warrant future investigation. Lightweight domain-adaptive fine-tuning methods, such as low-rank adaptation applied to open-weight models, could combine the data efficiency of prompt engineering with the performance advantages of parameter updates. Real-time incremental knowledge graph population under streaming intelligence feeds presents engineering and consistency challenges that static batch evaluation does not address. Cross-lingual transfer of CTI extraction capabilities, leveraging multilingual large language models to process non-English threat reports with English-trained prompt templates, could substantially expand the operational scope of automated intelligence processing. Standardized benchmarking protocols that span the full extraction-to-knowledge-graph pipeline, rather than evaluating individual subtasks in isolation, would accelerate progress toward deployment-ready threat intelligence automation.

References

1. M. Büchel, et al., "SoK: Automated TTP extraction from CTI reports --- Are we there yet?," in *Proceedings of the 34th USENIX Security Symposium*, USENIX Association, 2025.

2. T. Satyapanich, F. Ferraro, and T. Finin, "CASIE: Extracting cybersecurity event information from text," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 8749--8757, AAAI Press, 2020.
3. Y. Cheng, O. Bajaber, S. A. Tsegai, D. Song, and P. Gao, "CTINexus: Automatic cyber threat intelligence knowledge graph construction using large language models," in *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2025.
4. Z. Jie and W. Lu, "LinkNER: Linking local named entity recognition models to large language models using uncertainty," in *Proceedings of the ACM Web Conference 2024 (WWW '24)*, ACM, 2024. Available: https://doi.org/10.1145/3589334.3645414
5. Z. Li, J. Zeng, Y. Chen, and Z. Liang, "AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports," in *European Symposium on Research in Computer Security (ESORICS 2022)*, pp. 589--609, Springer, 2022.
6. M. T. Alam, D. Bhusal, Y. Park, and N. Rastogi, "LADDER: Looking beyond IoCs --- Automatically extracting attack patterns from external CTI," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2023)*, ACM, 2023. Available: https://doi.org/10.1145/3607199.3607208
7. E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "SecureBERT: A domain-specific language model for cybersecurity," *arXiv preprint arXiv:2204.02685*, 2022.
8. P. T. Chung, "Enhancing Dental Polymer Formulation through Interpretable Machine Learning: A Comparative Analysis of Feature Selection and Algorithm Performance," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 234-241, Dec. 2025.
9. W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "UniversalNER: Targeted distillation from large language models for open named entity recognition," in *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.
10. T. Xie, Q. Li, J. Zhang, Y. Zhang, Z. Liu, and H. Wang, "Empirical study of zero-shot NER with ChatGPT," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 7935--7956, ACL, 2023.
11. M. Han, "Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 679-684, Dec. 2025.
12. D. Liang and C. Cai, "Optimizing Large-Scale Contract Review through Data Analytics: Practical Evidence from IPO Audits," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 242-249, Dec. 2025.
13. Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, and S. Kurohashi, "GPT-RE: In-context learning for relation extraction using large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 3534--3547, ACL, 2023.
14. S. Wadhwa, S. Amir, and B. Wallace, "Revisiting relation extraction in the era of large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pp. 15566--15589, ACL, 2023.
15. P. T. Chung, "Multi-Objective Optimization of Process Parameters for Dental Resin 3D Printing Using Improved NSGA-II Algorithm," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 276-287, 2026.
16. Y. Liu, "AI-Enhanced Healthcare Data Quality Governance: An Integrated Approach for Anomaly Detection and Integrity Verification," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 1, pp. 215-229, 2026.
17. X. Long, "Performance Evaluation of Anomaly-Based Detection Approaches for Zero-Day Attack Early Warning in Cloud Infrastructure," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 352-363, 2026.
18. R. Zhang, Y. Su, B. D. Trisedya, X. Zhao, M. Yang, H. Cheng, and J. Qi, "AutoAlign: Fully automatic and effective knowledge graph alignment enabled by large language models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3168--3182, 2024. Available: https://doi.org/10.1109/TKDE.2023.3325484
19. X. Zhao, Y. Jia, A. Li, R. Jiang, and Y. Song, "Multi-source knowledge fusion: A survey," *World Wide Web*, vol. 24, pp. 1947--1987, 2021. Available: https://doi.org/10.1007/s11280-020-00811-0
20. Y. Zhang, T. Du, Y. Ma, J. Yan, S. Li, Z. Li, et al., "AttacKG+: Boosting attack knowledge graph construction with large language models," *Computers & Security*, vol. 150, p. 104220, 2025. Available: https://doi.org/10.1016/j.cose.2024.104220
21. Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10572--10601, ACL, 2023.
22. A. Jagannatha and H. Yu, "Calibrating structured output predictors for natural language processing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 2078--2092, ACL, 2020.
23. J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A survey of confidence estimation and calibration in large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*, pp. 6577--6595, ACL, 2024.
24. L. Huang and X. Xiao, "CTIKG: LLM-powered knowledge graph construction from cyber threat intelligence," in *Proceedings of the First Conference on Language Modeling (COLM 2024)*, 2024.
25. Z. Yan, S. Yang, W. Liu, and K. Tu, "Joint entity and relation extraction with span pruning and hypergraph neural networks," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 7512--7526, ACL, 2023.
26. L. Guo, Z. Chen, J. Chen, Y. Fang, W. Zhang, and H. Chen, "Revisit and outstrip entity alignment: A perspective of generative models," in *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.

27. M. Zhong, "Multi-Dimensional Feature Analysis and Evaluation Methods for Anomalous Fund Flow Identification in Cross-Border Financial Transactions," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 2, pp. 1-13, 2026.
28. Y. Zhang, "A Comparative Study of Machine Learning Methods for Automated Customer Service Dialogue Quality Assessment," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 328-338, 2026.
29. L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1--62, 2024. Available: https://doi.org/10.1145/3618295
30. Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang, "LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities," *arXiv preprint arXiv:2305.13168*, 2023.