Evaluating Large Language Models for Strategic Knowledge Extraction in Capability-Based Planning

Hein Kolk, Julia García-Fernández, Julia Bronkhorst, Roos Bakker
TNO
Anna van Buerenplein 1, 2595 DA The Hague, The Netherlands
{hein.kolk, julia.garciafernandez, julia.bronkhorst, roos.bakker}@tno.nl

Abstract

In a security environment that is growing more complex, large national organizations like the police rely on strategic frameworks to guide their decision-making. Frameworks like the Capability Based Planning (CBP) system are used to address this, but require a vast amount of information to function properly. A significant but underused store of information lies within an organization's own internal flow of documents, like vision statements or annual reports. We tap into this flow by proposing a method to automatically extract relevant strategic entities and structuring them within a knowledge graph. We evaluate the performance of various Large Language Models (LLMs) on a corpus of policy excerpts from the Dutch National Police in extracting relevant strategic entities and linking them to core police capabilities. We employ the novel alternative annotator test (Alt-Test) to determine if an LLM can serve as a reliable substitute for a human domain expert on this highly subjective task. Our evaluation shows that while LLMs cannot fully replace human experts, they prove to be valuable support tools by frequently identifying the same strategic information as the annotators, successfully extracting core entities and linking them to predefined capabilities.

Keywords: Capability-Based Planning, Knowledge Graphs, Large Language Models, Information Extraction

1. Introduction

The recent proliferation of Large Language Models (LLMs) presents an opportunity for public security organizations to unlock stored intelligence from their own archives of documents. This ability to extract intelligence can be used to inform strategic frameworks that guide decision-making in an increasingly complex security environment. One such framework is Capability Based Planning (CBP), a framework originally developed for military planning to address "volatile and uncertain" post-Cold War environments (Hales and Chouinard, 2011, p. 1). Described as the "gold standard" for strategic planning, its use has since been extended from defense into the public security sector to help organizations prepare for a wide range of future challenges (Hales and Chouinard, 2011). In summary, CBP focuses on developing and maintaining organizational capabilities to meet future challenges.

Implementing CBP effectively depends on good intelligence gathering, as the process must "start with a holistic appreciation of the problem space" (Hales and Chouinard, 2011). To aid in understanding the problem space, organizations can use a vast and underutilized source of strategic information already present within the organization: vision statements, annual reports, and other internal documents. These documents contain expert-curated critical knowledge about an organization's strategic situation. However, they are typically fragmented across countless unstructured texts and different organizational units. This fragmentation means that strategic insights are difficult to discover and use. Consequently, the holistic view required for CBP is hard to obtain, forcing planners to rely on incomplete information and making it difficult to effectively align strategy with capabilities (Hales and Chouinard, 2011). This difficulty is not only a data management issue; it reflects a core challenge within CBP of creating consistent, traceable models that connect high-level requirements to the systems that deliver them (Koivisto et al., 2022). The critical step in this process is explicitly linking abstract strategic insights, such as goals and trends, to the concrete, internal capabilities an organization possesses.

The recent emergence of LLMs offers the potential to support the process of structuring this knowledge. Unlike traditional information extraction methods, their demonstrated proficiency in processing long-form text (Chang et al., 2024) allows them to comprehend the nuanced and context-dependent information typical of strategic documents. However, for supporting the CBP process across multiple documents, simply extracting separate insights would lead to a collection of disconnected phrases. A Knowledge Graph (KG) is a suitable format to structure this information, defining entities—such as capabilities, threats, and organizational goals—and their relations in a schematic way. This creates a model that is both human- and machine-readable, providing the structure that the CBP field requires (Koivisto et al., 2022), built entirely from knowledge sourced from within.

Proceedings of the Knowledge Graphs and Large Language Models Workshop 2026 (KG-LLM@LREC26), pages 210–220
May 16, 2026. ©2026 European Language Resource Association

This paper, therefore, proposes and evaluates a method for using LLMs to automatically construct a CBP-oriented KG from internal police documents, aiming to provide a structured knowledge base that can directly inform the CBP process.

We introduce the following research questions:

1. **RQ1:** Can an LLM serve as a reliable substitute for a human domain expert in extracting a pre-defined knowledge graph from internal police documents?
2. **RQ2:** Is an LLM capable of coupling this knowledge to the organization's internal capabilities?

2. Related Work

Over the past decades, Natural Language Processing (NLP) techniques for information extraction have evolved significantly. Early approaches relied on syntax-driven methods such as dependency and constituency parsing (De Marneffe et al., 2006; Nivre et al., 2016). These were followed by representation-based methods such as word embeddings (e.g., Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)), which capture words in a vector space. Contextualized embeddings from transformer-based models (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) further advanced the field by representing words in context. Most recently, generative large language models (LLMs) have become popular since they were first introduced (Brown et al., 2020). Due to them being trained on large amounts of textual data, they have achieved state-of-the-art performance across many tasks (Chang et al., 2024). In this paper, we will compare LLM performance against human performance on a custom information extraction task for the security environment. In the next section, we will discuss related work on similar tasks and on similar work in the security environment.

Information Extraction Our task is related to several well-studied Natural Language Understanding (NLU) tasks that aim to extract structured information from text. One of such tasks is Semantic Role Labeling (SRL), where predicates and the relations between their semantic arguments are identified in a sentence. Often used roles are agent, object, and recipient (Li and Gao, 2025). Techniques for this task vary from early syntactic-driven approaches (Punyakanok et al., 2008), to models specifically trained for this task (He et al., 2017), to early attempts with LLMs (Cheng et al., 2024). Closely related, frame-semantic parsing allows for more specialized roles than SRL and is therefore often used in domain-specific use cases that require specific role types (Das et al., 2014; Marzinotto et al., 2018; Ferreira and Pinheiro, 2021). It identifies a central frame and its roles (frame elements) in text (Johnson et al., 2016). In our case, the frame corresponds to a Strategy, with linked elements Goal and Trend, but unlike standard frame parsing, we work at the document level and rely on a domain-specific ontology. We will elaborate on this in Section 3.

Police domain Research on information extraction (IE) in the security environment, such as the police and legal domain, has focused mainly on identifying entities, relations, and factual details from texts. Early work demonstrated the feasibility of extracting crime-related information from police and witness narratives using rule-based techniques such as Part-of-speech tagging (POS) with high precision and recall (Ku et al., 2008). Later studies introduced statistical methods to extract entities specific to police reports (Chau et al., 2002), and reviewed relation extraction techniques for detecting semantic relations in police filings in multiple languages (Carnaz et al., 2019). Other contributions proposed automated systems that preprocess, transform, and connect police reports from disparate sources to support forensic information analysis (Carnaz et al., 2018). More recently, transformer-based methods have been applied, such as leveraging BERT for extractive summarization of federal police documents (Barros et al., 2023), fine-tuning BERT for SRL for Dutch legal texts (Bakker et al., 2022; van Drie et al., 2023), and adapting lightweight LLMs for entity extraction in Chinese police reports through fine-tuning and prompt engineering (Xing and Chen, 2024).

These developments illustrate the gradual shift from rule-based and early machine learning approaches toward neural and transformer-based methods. However, existing work remains focused on extracting factual or relational information at the sentence level. To our knowledge, no prior research has targeted the extraction of strategies, goals, and trends in police documents. Our approach addresses this gap by (1) grounding extraction in an ontology-based representation, and (2) operating at the document level, thereby capturing higher-level reasoning and cross-document patterns.

3. Method

Our methodology is designed to compare the performance of LLMs against human domain experts on a strategic information extraction task composed of two main components: schema extraction and capability linking. The experiment consists of four main phases. First, we developed a domain-specific ontology to define the strategic entities (Strategies, Goals, Trends) and their relations

to core police capabilities (Capabilities). Second, we prepared a corpus by selecting and segmenting a public strategic document from the Dutch National Police. Third, we conducted a parallel annotation process where a group of human experts and several LLMs annotated the same document segments according to the ontology. Finally, to evaluate the LLMs' performance relative to the human experts (RQ1 and RQ2), we employed the alternative annotator test (Calderon et al., 2025), which compares the inter-annotator agreement between humans to the agreement between a human and an LLM. The subsequent sections will detail each of these phases.

3.1. Ontology

To construct the ontology depicted in Figure 1, we first conducted a domain analysis by reviewing a broad set of relevant police documents to identify key concepts and relations. Next, we refined the initial ontology with two domain experts. During this refinement, some borderline cases between Strategies, Goals, and Trends were discussed explicitly, reflecting genuine conceptual overlap in strategic texts rather than simple annotation error. The discussions led to the formulation of four core classes and three relations or links:

Strategy A plan of action or internal program developed in response of a **trend** and intended to achieve a **goal**. It requires the use of **capabilities**, decision-making, and resource allocation. Strategies can range from long-term plans (e.g., invest in digital infrastructure) to short-term actions (e.g., update the website of the organization).

Goal The desired result of a **strategy**.

Trend A general development or change in the world and the way people behave. Trends are external to the organization, which does not create them but responds to them through the design, development, and implementation of **strategies**.

Capability The set of abilities or resources that the Police possesses or seeks to develop to accomplish their defined **strategies**.

The relations between these classes are as follows:

isResponseTo This link defines the relation between the classes **Strategy** and **Trend**. A Strategy can be a response to one or more Trends.

hasGoal This link defines the relation between the classes **Strategy** and **Goal**. A Strategy can have one or more Goals.

requires This link defines the relation between the classes **Strategy** and **Capability**. A Strategy can require one or more Capabilities.

Capabilities are defined by the Police and are defined as being necessary to execute the strategies. This implies that in our annotation experiment, Capabilities are not extracted from the text but are drawn from a pre-defined, official list provided by the organization. This will be further elaborated in Section 3.3.

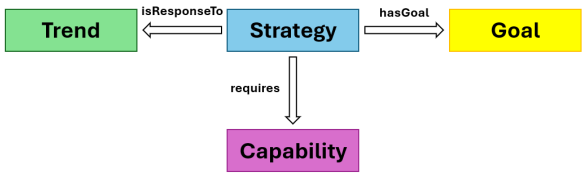

Figure 1: Ontology for the information extraction experiment.

We expect the high-level Strategy–Goal–Trend–Capability schema to transfer more easily than the specific capability inventory, which is organization-dependent and would require local adaptation.

3.2. Corpus

The corpus for this study consists of strategic police documents annotated by humans (domain experts) and three different LLMs.

Annotation Documents While an initial set of 50 internal strategic documents was provided by the Dutch Police, their confidential nature restricted the choice of LLM models and would have hindered the reproducibility of our results. Therefore, this study utilizes a comparable public document: the "Begroting en beheerplan politie 2025-2029"[1]. The entire 112-page document was considered too large for human annotators and so the document was segmented. Ultimately, 8 sections with an average word count of 242 were selected for the experiment. The primary selection criterion was that each section should contain instances of all concepts from our ontology.

Police Capability selection A confidential list of 70 police capabilities was provided for the experiment. To simplify the matching task for the human annotators during the experiment, this list was shortened. To ensure that the capability list still represented the different parts of the organization, each capability was grouped into a category based on their description. One capability was

[1] https://www.politie.nl/informatie/begroting-en-beheerplannen-politie.html/

then selected from each category for the final list used in the annotation task. A slightly modified and anonymized version of the categorized capability list is provided in our repository for reproducibility purposes, as the original operational capability list used in the experiment is confidential and cannot be publicly released.

3.3. Annotation Procedure

3.3.1. Human Annotation Process

To create annotation instructions, we defined a common understanding of the concepts of the ontology, drafted instructions, and tested those internally before the final instructions. The annotation task was performed in Label Studio[2], an open source annotation tool that allows labeling texts, grouping concepts and creating relations. Four police experts, all with management experience and familiarity with strategic documents, annotated the corpus. The tool was customized to allow linking **strategies** to the predefined **capabilities**.

3.3.2. LLM Annotation Process

To ensure a fair comparison with the human annotators (RQ1), we kept the LLM instructions as close as possible to those used by the human annotators. Since human annotators had continuous access to the full annotation guidelines, we included these in every prompt. Prompts were in English, but the document excerpts remained in Dutch; models were instructed to produce structured outputs in Dutch. All prompt templates and the full annotation guidelines used for both humans and LLMs are included in the repository to support reproducibility. Furthermore, we designed each task to be independent, reducing cognitive load and enabling separate evaluation of each stage. This independence also facilitates ablation studies, allowing us to assess the contribution of each annotation task individually.

To obtain the LLM annotated dataset, we have used the structured model outputs API by OpenAI[3] for GPT models running Azure (servers located in Europe), and the structured outputs API by Ollama[4] for open-source models running locally on our own servers. These APIs were used to constrain the LLMs' output to a schema that we defined using Pydantic models, directly mirroring the ontology shown in Figure 1. Our prompting strategy combined several techniques to ensure structured and accurate outputs. We used task decomposition, prompting the LLM for each extraction step separately, and output chaining to feed the results of one step into the next. Specifically, for the link extraction tasks, the LLM was provided with the list of previously extracted strategies and prompted to identify their corresponding links. Each prompt was ontology-constrained using Pydantic models and included the full annotation instructions with examples (few-shot prompting). Additionally, the temperature was set to 0.2 across all tasks to prioritize precision and consistency over creative recall.

We tried the annotation task with different OpenAI models: GPT-4.1 (version 2025-04-14), GPT-4.1-mini (version 2025-04-14), and GPT-5-mini (version 2025-08-07). The GPT-4.1[5] series offers strong instruction-following capabilities and supports long-context inputs, which is essential for our task requiring adherence to the detailed annotation instructions and handling prompts that include the full instructions, the output from the previous task and the specific instruction and the document text. GPT-4.1-mini was also included to evaluate whether its lower cost and faster inference could deliver comparable performance to the full GPT-4.1 model. Moreover, GPT-5-mini[6] was tested, a more efficient version of OpenAI's latest model in terms of cost and speed, which excels in instruction following for well-defined, complex tasks.

We also executed the task with several open-source models locally: Mistral-small-3.1-24B[7], Gemma3-27B[8], and Qwen3-4B[9], all in instruction-tuned variants to ensure strong adherence to annotation guidelines. To accommodate hardware constraints and improve inference speed, these models were used in quantized GGUF formats (Q8_0: 8 bits per parameter and standard symmetric quantization), which reduces memory footprint and energy consumption. Although quantization introduces a trade-off by potentially lowering model accuracy (Shi and Ding, 2025), it allowed us to assess whether these models could still deliver satisfactory performance under resource-limited conditions. Although smaller open-source models often cannot compete with their larger commercial counterparts, we still wanted to measure their performance on the task due to their use for privacy-sensitive data. We carried out the task with all aforementioned models and manually reviewed

[2] https://labelstud.io/

[3] https://platform.openai.com/docs/guides/structured-outputs

[4] https://ollama.com/blog/structured-outputs

[5] https://platform.openai.com/docs/models/gpt-4.1

[6] https://platform.openai.com/docs/models/gpt-5-mini

[7] https://huggingface.co/bartowski/mistralai_Mistral-Small-3.1-24B-Instruct-2503-GGUF

[8] https://huggingface.co/bartowski/google_gemma-3-27b-it-GGUF

[9] https://huggingface.co/unsloth/Qwen3-4B-Instruct-2507-GGUF

the output quality. From the open-source models, Gemma3-27B performed best. From the commercial models, GPT-4.1 was selected for the primary in-depth evaluation against the human baseline due to its overall superior performance. In the remainder of the text, we refer to GPT-4.1 as the LLM.

3.3.3. Matching Algorithm

To compare free-text annotations from multiple annotators, we designed a semantic matching algorithm to aggregate extracted text spans into a set of unique concepts. These concepts then served as the items on which we performed the evaluation. The algorithm, performed separately for each entity type (Strategy, Goal, Trend), operates as follows:

1. Concept Identification: For each annotator, their individual text spans are grouped into distinct concepts, based on a predefined semantic similarity threshold. Semantic similarity between spans was computed as cosine similarity between multilingual sentence embeddings (SentenceTransformers `paraphrase-multilingual-MiniLM-L12-v2`); spans were merged when similarity $\geq \tau$ (we use $\tau = 0.8$).

2. Concept Aggregation: These individual concepts are aggregated across all annotators into a final, universal set of unique concepts. This cross-annotator matching merges concepts if their similarity exceeds the threshold. Special rules were created for ambiguity: the highest-scoring match is chosen for one-to-many alignments, and all related concepts are merged in many-to-many cases.

The process yields a binary result matrix where rows represent the unique concepts and columns represent the annotators. A cell is marked *1* if an annotator identified a given concept. This matrix serves as the direct input for our evaluation metrics.

Worked example. (Excerpt in Dutch.) Consider the sentence: *"Om de risico's zo goed mogelijk te beheersen zet de politie met de arbeidsmarktstrategie een breed pakket aan maatregelen in."* Annotator H1 marks the Strategy span *"de arbeidsmarktstrategie"*, H2 marks *"met de arbeidsmarktstrategie een breed pakket aan maatregelen in"*, and GPT-4.1 marks *"arbeidsmarktstrategie"*. We compute pairwise semantic similarity between all extracted spans; spans with cosine similarity $\geq \tau$ (we use $\tau = 0.8$) are merged into a single Strategy concept. This yields a concept-by-annotator incidence row where H1=1, H2=1, and GPT-4.1=1 for this concept, which then serves as input to Krippendorff's α and the Alt-Test.

Link evaluation was handled separately. An annotator was credited with a **Strategy** $\rightarrow$ **Capability** link only if they found the source Strategy and the capability string was an exact match from the predefined list. For the **Strategy** $\rightarrow$ **Goal/Trend** links, a match required that the linked entities belonged to the same unique concept identified in the aggregation step.

3.4. Evaluation Metrics

3.4.1. Inter Annotator Agreement

Using the semantic matching algorithm defined in Section 3.3.3, we calculate the IAA to establish the human baseline performance. This step is crucial for contextualizing the subsequent LLM evaluation. The reliability of the Alt-Test's conclusions is related to the consistency of the human annotators (Calderon et al., 2025). A stable human IAA demonstrates that the annotation task is sufficiently well-defined and interpretable, providing a meaningful benchmark against which the LLMs can be compared.

We employ two different metrics to calculate the IAA:

- **Schema extraction:** We used **Krippendorff's alpha**, a chance-corrected metric, to measure agreement among annotators on whether a unique *strategy*, *goal* or *trend* concept was "found" or "not found" within the text (Krippendorff, 2011).

- **Capability linking:** For the task of linking found strategies to the pre-defined list of capabilities, we measure agreement by calculating the pairwise Jaccard index. This was calculated as the intersection over the union of the sets chosen by two annotators. For Strategy$\rightarrow$Capability, we compute Jaccard only on cross-annotator Strategy pairs that match at similarity $\geq \tau$ within a document; unmatched/missing strategies are excluded rather than encoded as empty sets. We then average Jaccard scores across matched Strategy pairs and annotator pairs.

Similarity Threshold Selection The choice of the semantic similarity threshold significantly impacts the matching process. To select an appropriate value, we performed a sensitivity analysis by calculating the Inter-Annotator Agreement (IAA) at various thresholds ranging from 0.45 to 0.95 (see Figure: 2). The analysis revealed that a single threshold was not optimal for all tasks; entity agreement (Krippendoff's Alpha (Krippendorff, 2011)) peaked at more lenient thresholds, while capability link agreement (Jaccard index) was highest at stricter ones.

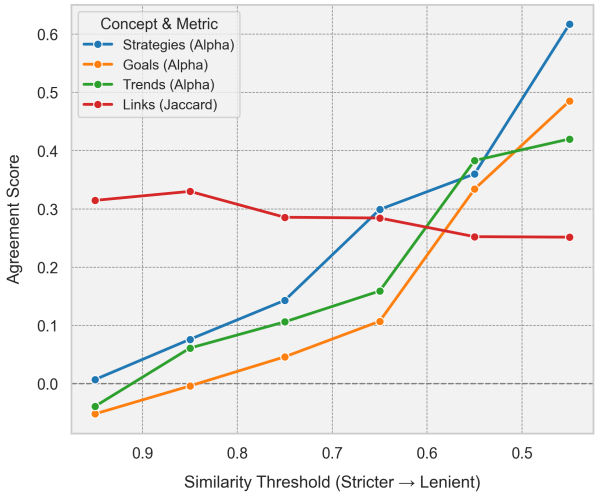

Figure 2: IAA at different threshold levels

We therefore set $\tau = 0.8$ as a practical compromise: in this region, concept-level agreement (Krippendorff's α) has moved into a positive range, while capability-link agreement (pairwise Jaccard) remains close to its maximum before declining at more lenient thresholds.

3.4.2. Alt-Test

Recent research has highlighted the difficulty of directly comparing LLM performance against a curated ground truth dataset. This is due to several key challenges: traditional metrics like F1 score lack clear decision thresholds, a true 'gold standard' is often unavailable for subjective tasks, and LLMs may even outperform the human annotators who create such datasets (Calderon et al., 2025; Nahum et al., 2024). Furthermore, a subjective task like strategic knowledge extraction allows for multiple valid interpretations, meaning a range of different annotations can be considered correct. For this reason, comparing the generated annotations against a ground truth is undesirable. Instead we used the alternative annotator test (Alt-Test), first proposed in Calderon et al. (2025). The procedure works by comparing two values: the LLMs' agreement with the human consensus, and an individual human's agreement with that same consensus. This test thus does not measure whether an LLM performs as well as a gold standard, but rather whether it can serve as a reliable replacement for a human on this task. A key feature of this test is the cost-benefit hyperparameter, epsilon (ϵ), which accounts for the practical advantages of using an LLM. This allows us to assess whether an LLM is a justifiable replacement, even if its performance is slightly below that of a human, by considering the savings in time and resources. Our results will show the LLMs performance across different epsilon levels.

4. Results

We first report human inter-annotator agreement (IAA) as a baseline, followed by Alt-Test results for schema extraction (RQ1) and capability linking (RQ2).

4.1. Human Inter-Annotator Agreement

A descriptive analysis of the human annotations revealed significant variability in the quantity of entities each expert identified. For example, an average of 29 strategies were annotated per person, but with a high standard deviation of 12.19. In contrast, the number of links created per strategy was more consistent across the annotators.

At the selected semantic similarity threshold (detailed in section: 3.4.1) of 0.8, the human annotators showed very low agreement on the free-text entity extraction task. Krippendorff's alpha was between 0 and 0.1 for Goals, Strategies and Trends. In contrast, agreement on the more constrained capability linking task was substantially higher, with a pairwise Jaccard index of 0.30. These scores, particularly the low alpha values, highlight the high degree of subjectivity inherent in the task and establish a challenging baseline for the LLM evaluation.

Aware of the low value achieved for the human inter-annotator agreement, we acknowledge the subjectivity of the task. A major challenge of the task is the fact that the annotators can freely choose parts of the text. For this reason we had to perform a semantic similarity alignment of the outputs. Moreover, while all annotators were experienced in the topic of strategic management and familiar with the documents, we observed how different annotation outputs could be different but valid. As explained, the Alt-Test does not evaluate similarity to a gold standard, but whether an LLM can act as an alternative annotator when multiple valid annotations exist (Calderon et al., 2025). Given the very low human IAA, our Alt-Test results should be interpreted narrowly as measuring annotator-like behavior under ambiguity, not as evidence that the extracted knowledge is sufficiently reliable for fully automated downstream use. Under these conditions, the main value of the evaluation is to identify which sub-tasks are suitable for expert-in-the-loop support rather than to establish absolute extraction quality.

4.2. LLM Performance (Alt-Test Results)

When examining the overall statistics for both human annotators and LLMs, several observations were made. First, LLMs produce substantially more entities and links than human annotators. A key difference in annotation behavior was also noted: for human annotators, link annotation exhibited

higher consistency than entity extraction, whereas for LLMs, outputs were more consistent for entity extraction than for linking. Finally, upon manual examination of the results produced by each model, we observed the highest output quality for GPT-4.1 in a manual qualitative check (e.g., fewer malformed outputs and more plausible entity spans), and therefore selected it for the primary Alt-Test evaluation. The results of this test are presented in Table 1.

4.2.1. Knowledge Extraction (RQ1)

The results for the entity extraction task, are presented in the first section of Table 1. At the recommended epsilon level for skilled human annotators (ϵ = 0.2 (Calderon et al., 2025)), the LLM failed the Alt-Test for all concept types, resulting in a winning rate (ω) of 0.0. Despite this, the average advantage probability (ρ) was consistently above 0.5, with values of 0.65 for Strategies, 0.68 for Trends and 0.62 for Goals. At a more lenient epsilon (ϵ = 0.3), does the LLM begin to pass the test for Strategies (ω = 0.75) and Trends (ω = 1.00). For Goals, however, the LLM failed to pass the test across all evaluated settings.

4.2.2. Capability Linking (RQ2)

As shown in the bottom section of Table 1, the analysis highlights a significant difference between the LLMs ability to handle closed-set and open-set linking tasks. For the closed-set linking task (Strategy $\rightarrow$ Capability (L S/C)), the LLM failed the Alt-Test at the expert-level epsilon level (ϵ = 0.2) with a winning rate of 0.0, despite a high average advantage probability (ρ = 0.65). When the epsilon was relaxed (ϵ = 0.3), the test passed with a winning rate of 1.0.

In contrast, for the open-set linking tasks (Strategy $\rightarrow$ Trend (L S/T) and Strategy $\rightarrow$ Goal (L S/G)), the model achieved a winning rate of 0.0 across all tested epsilon values. This corresponded to low average advantage probabilities (ρ) of 0.41 for S/G links and 0.19 for S/T links.

5. Discussion

Our results show that the LLM cannot be considered a direct substitute for a human expert for this strategic extraction task (**RQ1**). As detailed in Table 1, the model failed the Alt-Test at the strict epsilon level of 0.2, which is recommended for skilled annotators. However, the model's high raw advantage probability (of more than 0.6) for entity extraction indicates that its performance is not random; it frequently aligns with the majority opinion of the human annotators. Furthermore, when the epsilon is relaxed to a level of 0.3, the LLM did pass the test for the Strategy and Trend concepts. The failure on the Alt-Test shows that this alignment is just not strong or consistent enough to match the standard of an expert. Regarding **RQ2**, the LLM demonstrated a partial capability in coupling knowledge to the organization's internal capabilities. It showed promise on the task of connecting a Strategy to a predefined list of Capabilities, passing the Alt-Test at a relaxed threshold (epsilon=0.3). However, this relative success stood in sharp contrast to its complete failure on more complex linking tasks, such as connecting a Strategy to a Trend. This suggests the model can handle classification-like linking but struggles with establishing novel semantic relationships within the text, a finding consistent with the broader NLP literature that identifies document-level argument linking as a "decidedly harder" task than its sentence-level counterparts (Gantt, 2021).

The LLM's performance is best understood in the context of the task's inherent subjectivity, something clearly evident in our human annotation baseline. For the human annotations, the average standard deviation of the number of extracted concepts was very high compared to the mean. This confirms that human experts do not agree on one single interpretation, making multiple interpretations (and thus KGs) possible. Thus, the LLMs inability to pass the Alt-Test at the desired epsilon level is not only a sign of bad model performance, but also reflects that it is being compared to a varied set of human judgments. However, this challenge is offset by several significant practical advantages. First, the models demonstrated far greater output consistency than the human annotators. The reversal in consistency between humans and LLMs is also informative. Human annotators were relatively more consistent on capability linking because it is a constrained classification task over a predefined list, which reduces interpretive freedom. By contrast, entity extraction requires unbounded span selection from free text, giving annotators many more valid ways to identify and phrase the same underlying concept. For LLMs, the opposite pattern is plausible: extraction benefits from prompt regularity and structured output constraints, whereas linking requires more difficult semantic judgments over cross-entity relationships. Furthermore, the automated process is dramatically cheaper and more efficient: annotating our corpus took approximately 5 minutes in total for an LLM at a cost of approximately €0.64 [10], compared to several hours and an estimated cost of approximately €136 for one domain expert[11]. However, any operational

[10] Calculated using the total input/output tokens for the model GPT-4.1 and the Azure OpenAI service pricing overview https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/

[11] Calculated based on 4 hours of experiment duration and the average hourly rate of a police officer in

Table 1: Alt-Test Results for Entity and Relation Extraction at a Fixed Similarity Threshold of 0.8. The winning rate (ω) is shown across a range of cost-benefit hyperparameters (ϵ). A test is considered passed if $\omega \geq 0.5$. Passing results are highlighted in bold.

			Winning Rate (ω) at Epsilon (ϵ)			
Task Type	**Sample Size (n)**	**Avg. Advantage Prob. (ρ)**	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
Entity Extraction						
Strategies	78	0.65	0.00	0.00	**0.75**	**1.00**
Trends	59	0.68	0.00	0.00	**1.00**	**1.00**
Goals	66	0.62	0.00	0.00	0.25	0.25
Relation Extraction (Linking Only)						
Strategy → Capability (L S/C)	182	0.65	0.00	0.00	**1.00**	**1.00**
Strategy → Goal (L S/G)	101	0.41	0.00	0.00	0.00	0.00
Strategy → Trend (L S/T)	115	0.19	0.00	0.00	0.00	0.00

use would still require expert validation of LLM outputs for accountability; thus these figures reflect *pre-annotation* cost/time rather than end-to-end production effort. We also observed that the LLMs produced substantially more entities and links than the human experts. The speed and low cost achieved by the automatic evaluation using LLMs make it possible to analyze an entire corpus of documents, a scale unattainable with manual annotation. Crucially, this scalability enables the knowledge graph to be continuously updated with new information. As new documents are produced, an LLM can process them at scale to populate the graph with new strategic instances, ensuring the repository reflects the latest information, a task difficult for humans.

These findings have direct implications for the initial goal of supporting CBP. Our results show that a fully automated system for creating an expert-level strategic KG is not yet reliable. However, the LLMs consistent positive advantage probability demonstrates that it does not fail randomly. Instead, we argue that it successfully generates a plausible first draft of the strategic landscape by identifying entities that frequently align with human consensus. Therefore, the key contribution for this technology is its ability to produce this initial draft at a scale and speed that is impossible for humans. This transforms the starting point for strategic analysis: instead of beginning with hundreds of disconnected documents, planners can start from a single pre-populated knowledge base. At the same time, such a starting point may also anchor analysts too strongly in the framings already present in the processed documents. For that reason, we see the KG as a support tool for exploration and critique, not as a substitute for strategic judgment or for considering missing and alternative perspectives. While this model still requires refinement from an expert, it shifts the focus from simple data extraction to higher level strategic analysis tasks. This then enables a more comprehensive and data driven approach to CBP.

the Netherlands https://www.salaryexpert.com/salary/job/police-officer/netherlands

6. Conclusion

In this paper, we explored the potential of large language models (LLMs) to support or replace human domain experts in extracting structured information from police documents using a domain ontology. We also examined the feasibility of linking the extracted knowledge to predefined organizational capabilities. Our main contributions include: (i) the development of an ontology for strategic planning in the police domain, validated by domain experts; (ii) the design of a semantic similarity-based matching algorithm capable of aligning semantically equivalent annotations despite differences in phrasing. This is essential for our complex document-level information extraction task, which resembles frame semantic parsing in that it requires identifying a central concept (Strategy) and linking related elements based on an ontology; (iii) the application of the novel Alt-Test, a recent approach for assessing whether LLMs can perform annotation tasks at a level comparable to humans. This is an important consideration in the challenging landscape of LLM evaluation, where gold standards are often absent and multiple correct answers may exist. Finally, we demonstrate an innovative approach for partially automating Capability-Based Planning by aligning organizational capabilities with strategic documents, providing a foundation for identifying alignment gaps between an organization's strategic vision and its capabilities.

Our results highlight clear strengths and weaknesses for the LLM. Regarding entity extraction (**RQ1**), the model fails as a direct substitute for an expert but serves as an effective assistive tool. Its performance on relation extraction (**RQ2**) is divided: it succeeds at the closed-set task of linking strategies to predefined capabilities, but fails to connect open-set concepts like goals and trends.

Code and Data Availability

The source code for our methodology is available at https://github.com/capopaper/capo. The repository contains all scripts for data processing and analysis. As part of this work, we also introduce **StratID**, our newly created corpus of annotated strategic police documents, which is included in the repository. To support reproducibility, the repository also contains an anonymized, slightly modified version of the reduced capability list used for the annotation setup, as the original capability list is confidential and cannot be shared publicly. The repository further includes the prompt templates, the full annotation guidelines, and the hardware specifications for local model runs.

7. Ethical considerations

Generative LLMs can produce plausible but incorrect extractions (hallucinations). In a policing context, such errors may cause harmful downstream decisions if outputs are treated as facts; we therefore position this method as decision support and require expert review and traceability of each extracted item to source text. Although we use a public document for reproducibility, operational deployment on internal police documents raises confidentiality constraints; depending on policy, this may require locally hosted models or strict contractual/data-handling controls when using external providers. Finally, model outputs and prompts should be logged and versioned to enable auditing and accountability. A further organizational risk is cognitive anchoring: a pre-populated KG may overemphasize the strategic framing already present in the processed documents, so it should be treated as a starting point for deliberation rather than as a complete representation of the problem space.

8. Acknowledgements

We gratefully acknowledge the collaboration and support of the Dutch police.

During the preparation of this work, the author(s) used Microsoft Copilot in order to: Grammar, spelling check, and improving the readability of some sentences. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

9. Limitations

This study has several limitations that suggest directions for future work. First, the study involved one public police document segmented into 8 sections and annotated by 4 human experts. This limited scale constrains the generalizability of the findings across document types, organizations, and governance contexts. Future work should evaluate the method on multiple documents and, ideally, across different public-safety organizations. Second, we evaluated only a small set of LLMs. Expanding the analysis to include a broader range of models, especially open-source models, would provide a more comprehensive view. In fact, we keep experimenting with smaller, open-source models that we run locally on our premises. Third, we did not isolate the effect of prompt language. Our setup used English prompts on Dutch source texts, with models instructed to return structured outputs in Dutch. Although this worked in practice, we did not test whether Dutch-only, English-only, or cross-lingual prompting leads to different extraction quality. Future work should compare these settings explicitly. Finally, we did not explore variations in hyper-parameters or alternative prompting strategies. Future experiments should consider prompt tuning and LLM-specific instructions, enabling us to assess the difference in performance compared to using the same annotation instructions as used for the human annotators.

10. Bibliographical References

Roos M Bakker, Romy AN van Drie, Maaike de Boer, Robert van Doesburg, and Tom van Engers. 2022. Semantic role labelling for dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 448–457.

Thierry S. Barros, Carlos Eduardo S. Pires, and Dimas Cassimiro Nascimento. 2023. Leveraging BERT for extractive text summarization on federal police documents. *Knowledge and Information Systems*, 65(11):4873–4903.

Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1):5.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever,

and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*.

Gonçalo Carnaz, Vitor Beires Nogueira, Mário Antunes, and Nuno Ferreira. 2018. An automated system for criminal police reports analysis. In *International Conference on Soft Computing and Pattern Recognition*, pages 360–369. Springer.

Gonçalo Carnaz, Paulo Quaresma, Vitor Beires Nogueira, Mário Antunes, and Nuno N. M. Fonseca Ferreira. 2019. A Review on Relations Extraction in Police Reports. In Álvaro Rocha, Hojjat Adeli, Luís Paulo Reis, and Sandra Costanzo, editors, *New Knowledge in Information Systems and Technologies*, volume 930, pages 494–503. Springer International Publishing, Cham. Series Title: Advances in Intelligent Systems and Computing.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Michael Chau, Jennifer J Xu, and Hsinchun Chen. 2002. Extracting Meaningful Entities from Police Narrative Reports.

Ning Cheng, Zhaohui Yan, Ziming Wang, Zhijie Li, Jiaming Yu, Zilong Zheng, Kewei Tu, Jinan Xu, and Wenjuan Han. 2024. Potential and Limitations of LLMs in Capturing Structured Semantics: A Case Study on SRL. In *Advanced Intelligent Computing Technology and Applications*, pages 50–61, Singapore. Springer Nature Singapore.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56. Place: Cambridge, MA Publisher: MIT Press.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2017. Improving Implicit Semantic Role Labeling by Predicting Semantic Frame Arguments. ArXiv:1704.02709 [cs].

Vanessa C. Ferreira and Vladia Pinheiro. 2021. SpiNet - A FrameNet-like Schema for Automatic Information Extraction about Spine from Scientific Papers. *AMIA Annual Symposium Proceedings*, 2020:452–461.

William Gantt. 2021. Argument Linking: A Survey and Forecast. ArXiv:2107.08523 [cs].

Ting Gao, Xue Zhai, Chuan Yang, Linlin Lv, and Han Wang. 2024. Joint extraction of entity and relation based on fine-tuning BERT for long biomedical literatures. *Bioinformatics Advances*, 4(1):vbae194.

Doug Hales and Paul Chouinard. 2011. Implementing capability based planning within the public safety and security sector: Lessons from the defence experience. *Defence R&D Canada – Centre for Security Science*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Christopher R Johnson, Myriam Schwarzer-Petruck, Collin F Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles J Fillmore. 2016. Framenet: Theory and practice. Technical report, International Computer Science Institute.

Jouni Koivisto, Risto Ritala, and Matti Vilkko. 2022. Conceptual model for capability planning in a military context–a systems thinking approach. *Systems Engineering*, 25(5):457–474.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Chih Hao Ku, Alicia Iriberri, and Gondy Leroy. 2008. Crime Information Extraction from Police and Witness Narrative Reports. In *2008 IEEE Conference on Technologies for Homeland Security*, pages 193–198.

Junjiao Li and Zhengjie Gao. 2025. A comprehensive survey of semantic role labeling. *International Journal of Advanced AI Applications*, 1(2):79–105.

Gabriel Marzinotto, Jeremy Auguste, Frederic Bechet, Géraldine Damnati, and Alexis Nasr. 2018. Semantic Frame Parsing for Information Extraction : the CALOR corpus. ArXiv:1812.08039 [cs].

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Frank Martin Mtumbuka and Steven Schockaert. 2024. Entity or Relation Embeddings? An Analysis of Encoding Strategies for Relation Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6003–6022, Miami, Florida, USA. Association for Computational Linguistics.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2024. Are llms better than reported? detecting label errors and mitigating their effect on model performance. *arXiv preprint arXiv:2410.18889*.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *Ieee Access*, 12:71876–71900.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, 34(2):257–287. Https://direct.mit.edu/coli/article-pdf/34/2/257/1798602/coli.2008.34.2.257.pdf.

Tianyao Shi and Yi Ding. 2025. Systematic Characterization of LLM Quantization: A Performance, Energy, and Quality Perspective. ArXiv:2508.16712 [cs].

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Romy AN van Drie, Maaike HT de Boer, Roos M Bakker, Ioannis Tolios, and Daan Vos. 2023. The dutch law as a semantic role labeling dataset. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 316–322.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wei Xiang and Bang Wang. 2019. A Survey of Event Extraction From Text. *IEEE Access*, 7:173111–173137.

Xintao Xing and Peng Chen. 2024. Entity Extraction of Key Elements in 110 Police Reports Based on Large Language Models. *Applied Sciences*, 14(17):7819. Publisher: Multidisciplinary Digital Publishing Institute.