Financial Product Ontology Population with Large Language Models

Chanatip Saetia[1], Jiratha Phruetthiset[2], Tawunrat Chalothorn[1], Monchai Lertsutthiwong[1], Supawat Taerungruang[2], Pakpoom Buabthong[3,*]

[1]Kasikorn Labs, Kasikorn Business-Technology Group, Nonthaburi, Thailand
[2]Department of Thai, Faculty of Humanities, Chiangmai University, Chiangmai, Thailand
[3]Faculty of Science and Technology, Nakhon Ratchasima Rajabhat University, Nakhon Ratchasima, Thailand

Abstract

Ontology population, which aims to extract structured data to enrich domain-specific ontologies from unstructured text, typically faces challenges in terms of data scarcity and linguistic complexity, particularly in specialized fields such as retail banking. In this study, we investigate the application of large language models (LLMs) to populate domain-specific ontologies of retail banking products from Thai corporate documents. We compare traditional span-based approaches to LLMs-based generative methods, with different prompting techniques. Our findings reveal that while span-based methods struggle with data scarcity and the complex linguistic structure, LLMs-based generative approaches substantially outperform, achieving a 61.05% F1 score, with the most improvement coming from providing examples in the prompts. This improvement highlights the potential of LLMs for ontology population tasks, offering a scalable and efficient solution for structured information extraction, especially in low-resource language settings.

1 Introduction

With an increasing volume of text document repositories, the need for efficient and accurate information management systems has become inevitable. Ontology is one of the tools that facilitate structured representations of knowledge within specific domains, promoting interoperability and reasoning from unstructured data sources (Gruber, 1993). In addition, a specialized subset of ontologies, such as Schema markup, can also empower organizations to publish machine-readable web pages, thereby enhancing their visibility on search engines (Schema.org, 2008).

*Corresponding author: pakpoom.b@nrru.ac.th

However, the task of extracting and populating domain-specific ontologies from unstructured texts presents significant challenges due to the diversity of the source materials (Chasseray et al., 2023). Particularly in the banking sector, source documents, often authored by various internal units, lack a standardized format, frequently comprising only phrases or fragmented information rather than complete sentences (Petrova et al., 2017). In this domain, especially under low-resource language settings, structured storage can also be leveraged to construct a knowledge graph to facilitate downstream tasks such as recommendation systems (Guo et al., 2020) or question answering systems (Khongcharoen et al., 2022).

Recent advances in natural language processing (NLP), especially the development of large language models (LLMs), have led to new approaches to semantic annotation and ontology population (Babaei Giglou et al., 2023). These models, with their capacity for language comprehension, allow for automating the extraction of structured information, even in languages with limited training resources (Huang et al., 2023; Saetia et al., 2024). However, the effectiveness of LLMs for ontology populations in specific domains, such as banking, where the accuracy of extracted information is paramount, remains underexplored.

Tuning the prompts is one of the techniques to optimize LLMs for a specific task. A range of prompting techniques, including few-shot learning (Brown et al., 2020), chain-of-thought (CoT) prompting (Wei et al., 2022), and others, have been proposed to enhance the performance of LLMs across various NLP tasks, from named-entity recognition (NER) to complex question answering. These techniques aim to guide the model's attention and reasoning process, facilitat-

Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, pages 53–60

ing a more accurate extraction and interpretation of the desired information. While the impact of these prompting strategies has been extensively studied in other NLP applications, their application in structured information extraction from unstructured text, particularly within the context of low-resource languages, has yet to be thoroughly investigated. This gap necessitates a study to understand the potential of various prompting methods in improving the efficiency and accuracy of ontology population tasks especially in low-resource language. Herein, we study the extraction of structured information from corporate banking documents, particularly credit card product descriptions, using both traditional span-based methods and innovative LLMs-based generative approaches.

The main contributions of our work are summarized as follows:

- We provide a comparative study between span-based and generative approaches for ontology population tasks within the banking sector.
- We present LLM-based generative approaches with different prompting techniques for extracting structured information from text in a low-resource language context.

Our proposed approach could offer benefits to organizations maintaining internal documents in low-resource languages, seeking to streamline their data warehousing and enhance data interoperability across departments, particularly when resources for comprehensive data annotation are limited.

2 Related Work

2.1 LLMs for Generative Information Extraction

The recent advancements in LLMs have attracted attention to investigate their role in generative structured information extraction (IE), such as Named Entity Recognition (NER) and Relation Extraction (RE). Early studies have pioneered the approaches to address the limitations of LLMs in NER, introducing methods that formulate NER into a generative task and employ self-verification strategies for accuracy enhancement (Xia et al., 2023; Wang et al., 2023). Particularly, (Xia et al., 2023) proposes a training-free framework that improves the LLM performance in zero-shot NER. Similarly, frameworks like QA4RE (Zhang et al., 2023) have been proposed to improve LLM accuracy for RE tasks by aligning the task with question-answering tasks. GPT-RE (Wang et al., 2023) develops further by incorporating task-aware representations and reasoning logic to mitigate the issues of low relevance between entity and relation. To address the issues on the large number of relation types, (Li et al., 2023a) integrates the LLM with a dedicated inference module to improve document-level relation extraction.

Another paradigm to tailor the models to specific tasks is through prompt tuning, which has been shown to improve the overall performance (Yin et al., 2023). Code4UIE utilizes prompts that align the input-output pair with the pre-training stage of LLM for code generation (Guo et al., 2023). Few-shot prompting has also been used to provide task-specific examples for the LLMs to learn from (Brown et al., 2020). Techniques like CoT also encourage logical inferences and reasoning from the models (Wei et al., 2022). Additionally, interactive prompt strategies, like multi-turn QA, facilitate iterative refinement and feedback on generated extractions (Zhang et al., 2023). In IE, explicitly stating the definition of the field in the prompt is also reported to have a substantial influence on the extraction accuracy (Sousa et al., 2023).

3 Methodology

3.1 Data Collection

In this study, we analyze internal corporate documents detailing 20 retail banking products, specifically credit cards. These documents were manually tagged by a Thai linguist and subsequently verified by two computational linguistics researchers, to follow banking product ontology. The ontology employed herein aligns closely with the Schema's PaymentCard concept. Namely, we consider four main properties from PaymentCard (floorLimit, monthlyMinimumRepaymentAmount), FinancialProduct (annualPercentageRate), and Service (availableChannel). Schema's structure is selected because of its relevancy to the original documents. Other ontologies are discussed in A.1.

The primary objective of the ontology population task is to extract these properties from the original, unstructured text. This objective is achieved through the use of either span-based approaches or variations of the LLMs-based generative approaches.

3.2 Span-based Approach

This approach adopts the widely utilized BIO extraction concept (B-named entity for the beginning; I-entity name for the inside; O; for non-entity tokens) (Ramshaw and Marcus, 1999), which effectively detects the beginning and intermediate tokens within spans or entities. As pre-trained language models have shown success in this task (Devlin et al., 2018; Brown et al., 2020), here, we fine-tune Wangchanberta, a Thai pre-trained language model (Lowphansirikul et al., 2021), on the dataset while incorporating the BIO concept. A few-shot setting is also applied using two-state prototypical networks similar to previous few-shot NER methods (Ding et al., 2021; Li et al., 2023b)

3.3 LLMs-based Generative Approach

In this approach, GPT-3.5 (Brown et al., 2020; OpenAI, 2023b) is employed to extract structured properties or entities from the provided text (The comparison between GPT-3.5 and GPT-4 is provided in A.2). The prompts are designed as a prefix-prompt (Liu et al., 2023) to generate JSON-formatted outputs, ensuring ease of parsing and storing.

The prompt initially consists solely of a task description to guide the model in a zero-shot setting. As illustrated in Figure A.1 in part A, the prompt comprises three sentences. The first sentence provides the instruction while specifying the output format. The second sentence includes the name of the primary field, and the final sentence lists the associated sub-properties.

To improve the conciseness in the context of extracting the structured properties, three prompt construction strategies are presented as follows.

3.3.1 Few-shot Prompting

We adhere to the original method outlined in the previous work (Brown et al., 2020), which involves the insertion of examples after the task description but before the expected input text. Specifically, we use “TEXT:" to mark the commencement of the input, and “ANSWER:" to indicate the beginning of the output for each example. The structure of this few-shot setting prompt is depicted in Figure A.1 part C.

In a positive example, the input text contains all sub-properties in the task description. Conversely, the negative example deliberately excludes all sub-properties.

3.3.2 CoT Prompting

To guide the reasoning capability of the model, we follow a similar CoT approach as presented in (Wei et al., 2022). We include a segment of the extracted text that mentions all sub-properties as the preceding reasoning before generating the sub-properties themselves. The initial step guides the model to extract this text as a preliminary step before extracting each sub-field. Consequently, the model can identify the relevant text within the primary field without necessarily comprehending the sub-field at the beginning. In the positive example, the extracted text is populated as the first field (labeled as “extracted_text”) within the JSON-formatted output. The structure of the prompt, where “extract_text” is integrated into a few-shot setting, is shown in Figure A.1 part C.

3.3.3 Definition from schema.org

The definition of each primary field and its respective sub-properties are sourced from “schema.org” (Schema.org, 2008) and are provided explicitly in the prompts. Notably, only the meaning or description of each field is selected, with other details such as examples being excluded. The segment containing these definition within the prompt is after the task description but before the inclusion of example, as shown in Figure A.1 part B.

3.4 Evaluation Metrics

To evaluate the performance of both span-based approaches and the LLM-based generate approach, we employ F1 scores to compare the extracted entities and the annotated ones.

For the span-based approaches, evaluation entails the computation of a macro-averaged F1 score based on token prediction. In this work, macro-averaging is used to ensure equal consideration of all classes.

Meanwhile, the evaluation of the generative uses an F1 score based on exact matches. The evaluation involves determining the intersection between the predicted entities and the labeled entities within each sub-field.

4 Results and Discussion

4.1 Experiment - Span-based Approach

The results of the span-based approaches are shown in Table 1. When the Wangchanberta model is fine-tuned using the BIO extraction concept, the model yields only a 4.92% F1 score across all four main

Model	Precision (%)	Recall (%)	F1 score (%)
BIO extraction model (with classes)	13.89	2.99	4.92
Span detection (no classes)	**96.43**	11.79	21.01
Zero-shot	14.00	35.90	20.14
Few-shot (pos)	27.27	69.23	39.13
Few-shot (pos + neg)	55.81	61.54	58.54
Few-shot + Definition	47.79	69.23	56.54
Few-shot + CoT	49.54	69.23	57.75
Few-shot + CoT + Definition	51.79	**74.36**	**61.05**

Table 1: The results of span-based and LLMs-based generative approaches.

classes. This result confirms the limited efficiency of the fundamental span-based approach when employed with a restricted volume of training data.

In the subsequent experiment, conducted in a few-shot setting, span detection without considering specific classes yields a 21.01% F1 score. This result indicates that using traditional span-based approaches is still relatively limited even for just identifying the span without classifying the class.

4.2 Experiment - LLMs-based Generative Approach

The results, shown in Table 1, indicate that when utilizing a prompt containing solely a task description, the model achieved 20.14% F1 score. This poor performance likely arises from the model making assumptions on the definition of given fields.

To address this limitation, the incorporation of a positive example as a context, results in a notable improvement, yielding a 39.13% F1 score. This improvement is particularly significant in terms of recall, as it mimics the provided example, enhancing the models' ability to recognize relevant information. Subsequently, after introducing the negative example, the precision further increases (an F1 score of 58.54%). This enhancement suggests a better model comprehension, being able to identify what should be disregarded. Importantly, this outcome underscores that adopting even with only two labeled examples can yield substantial improvements.

To further improve performance, the inclusion of the proposed CoT and the field definition from schema.org provides a more comprehensive context for the model to understand each respective field. While there is a slight decrease in precision and the F1 score when utilizing CoT and definition separately, their combined integration results in approximately 2.5% increase in the F1 score compared to the prompt without these components, yielding the overall F1 score of 61.05%.

The results herein show the ability of LLMs to adapt to complex, domain-specific tasks using relatively simple prompt adjustments, even in a low-resource setting. Particularly, providing both positive and negative examples has the most influence on improving the performance in this information extraction task.

As these prompting techniques are not language specific, the prompting sequences proposed here can also be applied to other languages. However, the accuracy of the output depends on the proficiency of the selected LLM in the target language (Nguyen et al., 2023; Le Scao et al., 2023).

5 Conclusion

In this study, we conduct an evaluation of structured information extraction from unstructured Thai corporate documents describing retail banking products. We show that while LLM-based approaches allow for the extraction of relevant concepts without prior task-specific, the models face limitations in accurately interpreting unstructured text in low-resource language. This work demonstrates the efficiency of LLMs-based generative approaches enhanced by advanced prompting techniques, achieving an F1 score up to 61.05%. Our findings reveal that providing both positive and negative examples leads to the most improvement in the F1 score.

6 Future Work

This work can be extended outside of financial domain, leveraging Schema.org's extensive entities and properties. Additionally, the generated knowledge graph can be employed for question-answering or other information retrieval applications (Yang et al., 2015; Yani and Krisnadhi, 2021).

Integrating this graph into Large Language Models (LLMs) can potentially enhance their capabilities through graph neural networks and other reasoning methods (Kang et al., 2022; Liu et al., 2020).

7 Limitations

In this study, we note several limitations. First, GPT-3.5 was trained for general-purpose tasks and can be replaced with more robust, task-specific models. Specifically, Thai, as a low-resource language, tends to exhibit lower performance when compared to high-resource languages such as English. Second, the writing style can vary among different authors and languages, potentially making consistent annotation challenging. Also, identifying the span of the target entities within the document can sometimes be subjective. The span of the target entities can be a word or a phrase depending on the entity type. For example, “serviceLocation" can include a long phrase of an address but “value" in “floorLimit” can be only one number.

References

Hamed Babaei Giglou, Jennifer D’Souza, and Sören Auer. 2023. LLMs4OL: Large language models for ontology learning. In *The Semantic Web – ISWC 2023*, pages 408–427, Cham. Springer Nature Switzerland.

Bank-Ontology-Project. 1999. Bank ontology project. https://www.bankontology.com. Accessed on June 15, 2023.

Mike Bennett. 2013. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3-4):255–268.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yohann Chasseray, Anne-Marie Barthe-Delanoë, Stéphane Négny, and Jean-Marc Le Lann. 2023. Knowledge extraction from textual data and performance evaluation in an unsupervised context. *Information Sciences*, 629:324–343.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *arXiv preprint arXiv:2003.00911*.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2023. Retrieval-augmented code generation for universal information extraction. *arXiv preprint arXiv:2311.02962*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Minki Kang, Jinheon Baek, and Sung Ju Hwang. 2022. KALA: Knowledge-augmented language model adaptation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5144–5167.

Wirit Khongcharoen, Chanatip Saetia, Tawunrat Chalothorn, and Pakpoom Buabthong. 2022. Question answering over knowledge graphs for thai retail banking products. In *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–5.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023a. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Yongqi Li, Yu Yu, and Tieyun Qian. 2023b. Type-aware decomposed framework for few-shot named entity recognition. *arXiv preprint arXiv:2302.06397*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

OpenAI. 2023a. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Introducing ChatGPT. `https://openai.com/blog/chatgpt`. Accessed on February 5, 2024.

GG Petrova, Anatoly Tuzovsky, and Nataliya Valerievna Aksenova. 2017. Application of the financial industry business ontology (FIBO) for development of a financial organization ontology. In *Journal of Physics: Conference Series*, volume 803. 012116.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Chanatip Saetia, Areeya Thonglong, Thanpitcha Amornchaiteera, Tawunrat Chalothorn, Supawat Taerungruang, and Pakpoom Buabthong. 2024. Streamlining event extraction with a simplified annotation framework. *Frontiers in Artificial Intelligence*, 7:1361483.

Schema.org. 2008. Schema.org. `https://schema.org/`. Accessed on June 15, 2023.

Hugo Sousa, Nuno Guimarães, Alípio Jorge, and Ricardo Campos. 2023. GPT Struct Me: Probing gpt models on narrative entity extraction. In *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 383–387. IEEE.

Ana Tănăsescu. 2016. A financial reporting ontology design according to IFRS standards. *Economic Insights-Trends & Challenges*, 68(4):37–44.

Eileen Z Taylor and Ann C Dzuranin. 2010. Interactive financial reporting: an introduction to extensible business reporting language (XBRL). *Issues in Accounting Education*, 25(1):71–83.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yu Xia, Yongwei Zhao, Wenhao Wu, and Sujian Li. 2023. Debiasing generative named entity recognition by calibrating sequence likelihood. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1137–1148, Toronto, Canada. Association for Computational Linguistics.

Min-Chul Yang, Do-Gil Lee, So-Young Park, and Hae-Chang Rim. 2015. Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, 42(23):9086–9104.

Mohammad Yani and Adila Alfa Krisnadhi. 2021. Challenges, techniques, and trends of simple knowledge graph question answering: a survey. *Information*, 12(7):271.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *arXiv preprint arXiv:2305.11150*.

A Supplementary Information

A.1 Financial Ontology

Current ontologies in the financial and banking sectors are designed to provide a formal representation of financial knowledge. This often encompasses the categorization of financial entities and the definition of properties for each entity. In the early stages of ontology and knowledge engineering, the primary objective was to create a machine-readable Semantic Web to enhance information retrieval and enable reasoning with structured knowledge (Berners-Lee et al., 2001).

While various financial ontologies have been developed to represent concepts from both the customer perspective (such as transactions) and the organization perspective (like banking products), there is often a degree of conceptual overlap among these ontologies, each capable of representing information relevant to the financial and banking industries. The Financial Industry Business Ontology (FIBO), developed by the Enterprise Data Management (EDM) Council, stands out as one of the most comprehensive, encompassing loans, securities, and financial processes (Bennett, 2013). The eXtensible Business Reporting Language (XBRL) provides a global framework for exchanging business information, primarily focused on processing financial statements and regulatory filings (Taylor and Dzuranin, 2010).

Similarly, the International Financial Reporting Standards (IFRS) ontology is tailored to financial reporting standards (Tănăsescu, 2016). In contrast, the Bank Ontology offers a wider array of concepts covering products offered by banking institutions (Bank-Ontology-Project, 1999). Conversely, although not strictly adhering to the W3C Web Ontology Language (OWL) standards initially intended for the Semantic Web, the Schema.org markup provides centralized, extensible schemas for representing structured data vocabularies across various industries, including financial and banking products (Guha et al., 2016).

A.2 The comparison of GPT-3.5 and GPT-4

In the following comparative analysis, both GPT-3.5 and GPT-4 were subjected to identical prompts, with the results shown in Table A.1. Contrary to GPT-3.5, prompting GPT-4 under a zero-shot setting yields a better result, likely from broader general reasoning capability (OpenAI, 2023a). When positive and negative examples are provided, GPT-4 exhibits improvements similar to GPT-3.5. Nevertheless, GPT-3.5 demonstrates a slightly better F1 score, likely because the outputs from GPT-4 are paraphrased into more readable text, while those from GPT-3.5 maintain the exact text extracted from the input text. Moreover, the GPT-4 output may include additional information, in some cases, the start and end dates, or the special condition of the property. A similar comparison between GPT-3.5 and GPT-4 also be observed when prompting with CoT and definition. In summary, although GPT-4 may be employed for this task and may generatively extract accurate information, additional post-processing or restrictive prompting will need to be used to prevent the model from generating additional information. Other metrics that measure semantic similarity or human evaluation can be used to further investigate the comparative performance of the two models.

Model	Precision (%)	Recall (%)	F1 score (%)
Zero-shot	22.82	43.59	29.96
Few-shot (pos + neg)	53.41	60.26	56.63
Few-shot + CoT + Definition	50.49	66.67	57.46

Table A.1: The results of LLMs-based generative approach using GPT-4.

Your task is to extract structured data (JSON format) from the TEXT given the DEFINITION. The structured data has a field as availableChannel. In the field "availableChannel" includes three sub fields: "serviceLocation", "servicePhone", and "serviceUrl". A

DEFINITION:
serviceLocation: **{Definition from schema.org}**
... B

TEXT: **{NEGATIVE_EXAMPLE_TEXT}**
ANSWER: {'availableChannel': []}

TEXT: **{POSITIVE_EXAMPLE_TEXT}**
ANSWER: {'availableChannel': [{
'extracted_text': '**{EXTRACTED_TEXT}**',
'serviceLocation': '...',
'serviceUrl': '...',
'servicePhone': '...'
}]} C

TEXT: **{INPUT_TEXT}**
ANSWER:

Figure A.1: The prompt for extracting “availableChannel” integrating all construction strategies