Harnessing Open-Source LLMs for Tender Named Entity Recognition

Asim Abbas[1*], Venelin Kovatchev[1], Mark Lee[1], Niloofer Shanavas[2], Mubashir Ali[1]
[1]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK
[2]School of Computer Science, University of Birmingham, Dubai Campus, UAE
axa2233@student.bham.ac.uk, {v.o.kovatchev, m.g.lee, n.shanavas, m.ali.16}@bham.ac.uk

Abstract

In the public procurement domain, extracting accurate tender entities from unstructured text remains a critical, less explored challenge, because tender data is highly sensitive and confidential, and not available openly. Previously, state-of-the-art NLP models were developed for this task; however developing an NER model from scratch required huge amounts of data and resources. Similarly, performing fine-tuning of a transformer-based model like BERT requires training data, as a result posing challenges in training data cost, model generalization, and data privacy. To address these challenges, an emerging LLM such as GPT-4 in a Few-shot learning environment achieves SOTA performance comparable to fine-tuned models. However, being dependent on the closed-source commercial LLMs involves high cost and privacy concerns. In this study, we have investigated open-source LLMs like Mistral and LLAMA-3, focusing on the tender domain for the NER tasks on local consumer-grade CPUs in three different environments: Zero-shot, One-shot, and Few-shot learning. The motivation is to efficiently lessen costs compared to a cloud solution while preserving accuracy and data privacy. Similarly, we have utilized two datasets open-source from Singapore and closed-source commercially sensitive data provided by Siemens. As a result, all the open-source LLMs achieve above 85% F1-score on an open-source dataset and above 90% F1-score on a closed-source dataset.

1 Introduction

One of the basic tasks in the Natural Language Processing (NLP) domain is Name Entity Recognition (NER), which plays a cornerstone role by identifying and classifying entities such as names, persons, organizations, addresses, dates, locations, etc. available in unstructured format (Ji et al., 2019). These entities could be used for information retrieval available online, text summarization, chatbots, etc., facilitating downstream tasks and improving decision power (Toikka et al., 2021).

Similarly, the research community has given a lot of attention to NER in other domains, but less attention is given to the tender domain. Tenders are formal requests for proposals or offers, typically issued by a company, organization, or government agency that seeks to provide goods, services, or work to be provided (Siciliani et al., 2023). Moreover, tender documents are large, often consisting of more than 100 pages each. Manually extracting relevant information from such huge documents requires a lot of energy and time and is a labor-intensive task that is often prone to errors and inefficiencies.

State-of-the-art (SOTA) NER models are generally based on supervised fine-tuning of transformer base models possess, several challenges: a) Typically, it requires a domain expert to manually prepare high-quality annotated data, time consuming and tedious; b) The NER task in the tender domain is subjective and error-prone, as different annotators may classify the same entity differently. For example, given the word "Siemens," one person may classify it as a vendor (company), while another may label it as a product brand if it refers to Siemens-manufactured equipment or machinery in a tender document; c) A fine-tuned NER model that performs well on one dataset but does not perform well on another dataset in the same domain due to different terminologies used by each organization due to regional preferences, industry standards, or organizational terminology. For instance, "Tender Value" can be label variantly in the documents as "Contract Amount", "Estimated Budget", "Project Cost" etc; While adding new terminologies in the existing dataset is a challenging task that requires careful review; d) Using annotated datasets for low-resource languages adheres to additional challenges

Proceedings of Recent Advances in Natural Language Processing,pages 1–10
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-001

such as data scarcity (Zhu et al.).

Recent advancements in LLMs have significantly transformed the NLP domain, exhibiting successful outcomes in tasks such as Named Entity Recognition (NER) by effectively identifying and classifying entities across multiple languages (Brown et al., 2020). Leveraging LLMs for NER can efficiently mitigate the challenges as discussed above, adopting the in-context Learning (ICL) approach, where LLMs only require a few examples to learn from and perform a specific task. To the best of our knowledge, there is no such research available for Tender Named Entities Recognition (TNER) using open-source LLMs. This is the motivation of our study, ”Can open-source LLMs efficiently perform for Tender NER”, because the tender document is mostly confidential within the organization. Similarly, each organization uses its structure and standard terminology in the tender documents.

Answering this research question entails three basic concerns: data privacy, cost, and annotation. Compared to commercial and closed-source LLMs like GPT-3 (OpenAI) that oblige users to upload data by paid APIs, possibly compromising sensitive information (Das et al., 2025). In contrast, open-source LLMs can be freely available, downloaded, and run locally, which requires low-resource hardware like DeepSeek-R1 (Guo et al., 2025). Using LLMs in our applications, two popular approaches are available: Fine-tuning on domain-specific annotated data and in-context learning (ICL), where the user is required to provide only a few examples, also known as Few-shot learning, to perform a specific task.

In this study, we have explored the efficiency and ability of open-source LLMs towards TNER in three different scenarios: Zero-shot, One-shot, and Few-shot learning by leveraging five LLMS in our experiments, such as Mistral, Phi-3, LLAMA-3, Falcon-3, and Deepseek-R1. These models are selected on the following criteria: having a model size not over 8 billion parameters, popularity, best for general-purpose chat and reasoning, released after July 2023, and ranked position on the Huggingface and Ollama leaderboards. We have detailed information about these models presented in section 3.1. Finally, our contribution to this study is:

- Unlike previous studies that rely on expensive and proprietary LLMs for general Named Entity Recognition (NER), as a result, raising serious concerns about cost and data privacy. To the best of our knowledge, we are the first to challenge this norm by rigorously evaluating open-source LLMs for their NER capabilities in the tender domain, paving the way for a more accessible, transparent, and cost-effective future in the field.

- While LLMs are primarily trained on open-source data, our key concern is assessing how effectively open-source LLMs can identify entities within tender documents. Unlike general text, tender documents are not freely available, and each organization employs its own unique terminologies and structural formats. This variability presents a significant challenge, making it crucial to evaluate the adaptability and efficiency of open-source LLMs in this specialized domain.

- We performed comparative analysis of various LLMs in three different environments: Zero-shot, One-shot, and Few-shot learning, which enables us to assess their adaptability and learning efficiency in identifying entities within tender documents, which often feature domain-specific terminologies and diverse structures. By analyzing their strengths and limitations under varying levels of data availability, we provide insights into the feasibility of using LLMs for automated tender information extraction. This research empowers organizations to perform entity recognition on tender documents without relying heavily on costly domain experts.

- Our experimental design incorporates a confidential tender dataset from a commercial company, alongside a publicly available dataset. This combination ensures a comprehensive evaluation while enhancing the generalizability of our findings to a broader audience.

2 Literature Review

There are three different steps for developing, customizing, and utilizing the model to perform specific tasks.

2.1 Model Training from scratch to Fine-Tuning

When training LLMs from scratch, they learn from raw data while using substantial computational capabilities. Similarly, natural language model training requires extensive corpus data collection and neural architecture design, followed by multiple epochs of training for achieving model generalization across tasks. Among the most impactful Transformer-based models, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) began their training from scratch using large datasets to build effective generalization abilities. Recently, the training pipelines of LLaMA-3 (Grattafiori et al., 2024) and DeepSeek R1 (Guo et al., 2025) utilize optimized training pipelines together with bigger datasets. The benefit of training from scratch enables full control over architecture, along with vocabulary selection and data choice, which results in highly specialized models that perform better than fine-tuned models in domain-specific applications (Bommasani et al., 2021). Nevertheless, this methodology requires substantial computational strength, together with extended training time and extensive labeled data, which limits its practical use in various scenarios.

Moreover, the practical choice for model development involves fine-tuning a pretrained model by training it with additional task-specific data from a smaller dataset. This technique has proven effective for legal and procurement-related NLP tasks as demonstrated by LEGAL-BERT (Chalkidis et al., 2020), which goes through fine-tuning from legal and government procurement text sources. Likewise, the process of fine-tuning provides multiple benefits, which include reduced computational expenses, accelerated training time, and superior results on specialized tasks. In addition, the training process has become more efficient through LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) fine-tuning techniques, which reduce the quantity of updated parameters.

Regardless of its advantages, the fine-tuning technique comes with certain constraints. The principal disadvantage of using this method is catastrophic forgetting, which depicts how models discard information learned in their initial training data when they are adapted for domain-specific use (McCloskey and Cohen, 1989). Moreover, fine-tuning needs domain-specific data of high quality, but sometimes this data remains inaccessible. When fine-tuning occurs too many times, it results in overfitting, which means the model achieves high performance on its fine-tuned dataset but fails to generalize to new examples (Li et al., 2021).To address these challenges, instruction-tuning (Longpre et al., 2023) has been developed as an alternative solution that enables models to achieve better generalization performance between various tasks through minimal adaptation.

2.2 Beyond Fine-Tuning: Learning from Context with Few-shot, Zero-shot, and Prompting

The way LLMs grow in capacity while expanding their ability to generalize has changed from basic fine-tuning to ICL, which enables models to perform downstream tasks through prompts only. OpenAI introduced GPT-3 (Brown et al., 2020), which showed that LLMs could solve new tasks based solely on examples provided within context (Few-shot) or without any examples (Zero-shot) through prompts. The ability to learn from a wide range of pretraining tasks and instructions allowed the model to ingest universal problem-solving patterns. Likewise, the high cost of fine-tuning became impractical because training a GPT-3-sized model costs more than USD 4.6 million and requires hundreds of GPU years (Li, 2020), so ICL became a practical, scalable solution for low-resource and domain-specific tasks such as NER in tender documents.

Similarly, the annotation burden for extracting structured data from unstructured documents decreases significantly through ICL because it allows researchers to use prompts and minimal labeled examples without requiring domain-specific annotated corpora. The extraction of tender entities requires domain knowledge for labeling contracts, workflows, organizations, dates, and procurement items. Current models such as LLaMA3 (Grattafiori et al., 2024) and Phi-3 (Abdin et al., 2024), together with DeepSeek-R1 (Guo et al., 2025), show effective Zero-shot and Few-shot generalization abilities that perform better than traditional fine-tuned models on NER benchmarks. Additionally, these models produce JSON-structured outputs from natural language instructions, which makes them suitable for information extraction pipelines.

Recently, studies like (Sanh et al., 2021) and (Wei et al., 2022b) establish that proper prompt engineering enables LLMs to generalize tasks effectively without requiring supervised fine-tuning and extensive data. The paradigm has evolved through chain-of-thought prompting (Wei et al., 2022b) and retrieval-augmented ICL (Ram et al., 2023), which improve both factual grounding and reasoning capabilities essential for processing complex tender documents with embedded images, tabular data, and hierarchical workflows. Research demonstrates that entity extraction works efficiently with partial PDF content extraction from OCR processes when using a small number of curated examples that match procurement language structure and meaning.

3 Methodology

In this study, we perform Zero to Few-shot learning employing the ICL strategy towards tender entities extraction from unstructured tender documents and evaluate the performance of each learning approach. We apply Few-shot learning by extending our prompt to 3 examples to extract error-prone entities. Examples of Few-shot learning setups have been provided in Table 1.

In Zero-shot learning, the prompt template consists of three components: instructions, a list of required entities, and the test input data. Whereas in One/Few-shot learning, we add one more component, such as a complete example showing both input and output, which helps the model understand the task. As a result, the model doesn't require as much deep reasoning or explanation as in the Zero-shot case. Moreover, in this setup, the model is given slightly more context, which takes more time for the LLMs to process compared to Zero-shot, but the response tends to be more accurate and reliable. Similarly, in the Few-shot setup, the prompt template will be the same, but the size example would be more than one.

Before the ICL approach, a classification model required annotated data to fine-tune LLMS like BERT (Devlin et al., 2019), which is very expensive, time-consuming, and limits the efficient usage of LLMs. Additionally, LLMs are trained on generic data and can be used for broader applications, but fine-tuned LLMs tend to narrow down their capability and cause issues like catastrophic forgetting (McCloskey and Cohen, 1989).

To run these LLMs, we leveraged the Ollama[1] framework, which makes it very easy to deploy these models on a standard desktop system. The desktop setup we used had 16 GB of RAM and 2.8 GHz AMD Ryzen 5 CPUs. This demonstrates that there is no longer a need for high-end GPUs or powerful systems; LLMs can now be efficiently run even on regular desktop machines.

3.1 LLMs Model Selection

The considerations of computational cost, hardware constraints, and environmental impact, as outlined in (Kaplan et al., 2020), guide us in the selection of LLMs for tender entity extraction experiments. We limit model size to 3 to 8 billion parameters while focusing on compact, high-performing architectures. The models used in our study: Phi-3, LLaMA-3, Mistral, Falcon-3, were chosen because they are well-known and high-performing models on the Ollama leaderboard, which represents both community consensus and empirical effectiveness. Our study differs from previous work that only evaluates models between 7B–34B parameters and excludes models above 70B because of their high computational requirements and carbon footprint. The research shows that models with under 7 billion parameters are becoming more competitive in NLP tasks despite the scaling laws that benefit larger LLMs (Wei et al., 2022a). The models become more appropriate for real-world applications because they offer better efficiency and deployment capabilities. Our method of selecting lightweight models supports sustainable AI practices and keeps a solid empirical base. The table below shows the models we chose along with their parameter counts (Table 2).

Model	# Parameters	Release Year
LLaMA-3	8 billion	2024
Mistral-0.3	7 billion	2024
Deepseek-R1	7 billion	2025
Phi-3 mini	3.8 billion	2024
Falcon-3	7 billion	2024

Table 2: Selected Models for TNER Experiments

4 Experiments and Evaluation

4.1 Datasets

In this study, we have utilized two datasets: Open-source datasets by Singapore Government Procurement Dataset (2015–2021), available on Kaggle (on Kaggle, 2024) and a closed-source commercial dataset by Siemens, which is confidential to share. Open-source datasets comprise data on government tenders awarded by the Singapore Government during that period. This dataset comprises 23909 instances, available in a structured format, see Table 3. We employ a data augmentation technique to make it unstructured and prepare

[1] https://ollama.com/search

You are an expert tender information (Entities) extractor. Follow the examples below.

Example 1:

Input Text: ”’Status update: Awarded to Suppliers. In a call for competitive bids, Ministry of Education Schools (Government Ministries & Departments) released Tender No. MOESCHETT18300005 regarding provision of 16-day BICULTURAL STUDIES EDUCATIONAL IMMERSION PROGRAMME TO UNITED STATES OF AMERICA FOR DUNMAN HIGH SCHOOL. Following a thorough evaluation, the award was granted on 17/7/2018 to STA TRAVEL (PTE) LTD for $121879.7.”’

Output JSON:

{{{{
”Tender NO”: ”MOESCHETT18300005”,
”Tender Description”: ”PROVISION OF 16-DAY BICULTURAL STUDIES EDUCATIONAL IMMERSION PROGRAMME TO UNITED STATES OF AMERICA FOR DUNMAN HIGH SCHOOL”,
”Agency”: ”Ministry of Education - Schools (Government Ministries & Departments)”,
”Award Date”: ”17/7/2018”,
”Tender Detail Status”: ”Awarded”,
”Supplier Name”: ”STA TRAVEL (PTE) LTD”,
”Awarded Amount”: ”$121879.7”,
”Main Category”: ”Educational Immersion Programme”
}}}}

Test Input:

`{unstructured_tender_data}`

JSON output (do not add any extra keys or modify the keys or information)

Table 1: Prompt Template: One-shot/Few-shot Learning for Entities Extraction

Datasets	Size (Train/Test)	TNER Type	No of TNER Types
Siemens	30	Tenderee/Tenderer Name, Date, Tenderee/Tenderer Address, Tender Number, Tender Name, Tendree Personal, Telephone	9
Singapore	23909	Tender No, Description Award Date,Tender Status, Supplier Name Awarded Amount, Main Category	7

Table 3: Datasets Utilized During Experiments

for the ICL reasoning towards entity extraction. The dataset augmentation techniques for data transformation is available on Github [2]. On the other hand closed-source dataset comprised 30 tender documents, see Table 3. The required tender entities are available in the form of a complex table, which is difficult to extract accurately, and the text structure is lost after extraction See (Abbas et al., 2025) for table structure.

4.2 Evaluation Metrics

In this study, we have computed three matrices such precision, recall, and F1 score. These metrics are computed on the document-level entities identification. Since LLMs might not be able to exactly identify the entities, so we apply soft match criteria. For example, we have tenderee address ”SCREAT FAKES PLUMBING & HEATING COMPANY, 4017 W DIVERSEY AVE CHICAGO IL 69659-1225, US” that includes the name of tenderee, which is ”SCREAT FAKES PLUMBING & HEATING COMPANY”, address(street no, city, postal code, and country) which is ”4017 W DIVERSEY AVE CHICAGO IL 69659-1225, US”. If the LLMs predict it together and label as tenderee, it is still correct.

4.3 Open-source Data Evaluation

We process open-source datasets by Singapore Government Procurement in Zero-shot, One-shot, and Few-shot environments and calculate the results at the document level entities extraction for each LLM as discussed in Table 4.

Zero-shot: As depicted in Table 4, in Zero-shot evaluation LLAMA-3 as the strongest performer with precision at 92.74%, recall at 90.19%, and F1 score at 91.18% which demonstrates its robustness even without prior examples. Interestingly, Falcon-3 achieves similar performance, almost 89% across three matrices, but Deepseek-R1, Mistral, and PHI-3 demonstrate slightly reduced yet strong results. Ultimately, LLama-3 achieves better pretraining exposure and information extraction alignment despite not receiving any fine-tuning or example con-

[2] https://github.com/TuriAsim/Tender-NER.git

ditioning.

One-shot: Moving from Zero to One-shot evaluation, performance results show considerable variations. The majority of models either fail to improve or decline their performance levels compared to the Zero-shot results. The F1-score of LLama-3 decreases to 78.19%, which shows that the model becomes more sensitive to both prompt writing and the quality of provided examples. The improvement in Falcon-3 F1-score up to 86.93%, mainly results from a substantial increase in recall performance (96.62%) because it adapts well to minimal supervision. The inferior performance of LLama-3 and other models in this scenario may result from prompt misalignment because One-shot examples sometimes present noise or bias when the single instance does not effectively represent the full data distribution.

Few-shot: Curiously, all models achieve their highest performance levels during Few-shot prompting. The F1-score of Falcon-3 reaches 98.80% while Mistral reaches 97.72% and PHI-3 reaches 95.96% and LLama-3 achieves 94.89%, and Deepseek-R1 reaches 94.13%. These results depict that demonstration-based learning plays a vital role in LLMs. The ability of models to learn data structures and task-specific details improves their accuracy and generalization when they receive more examples. Our observation at entities level indicate that the models consistently extract entities such as ”Tender No” and ”Supplier Name”, ”Award Date, ”Amount” and ”Status”. However, they struggle with extracting ”Main Category” and ”Tender Description” because these entities embedded in unstructured text and often consist of long, complex patterns, making them more challenging to identify accurately.

4.4 Close Source Data Evaluation

Likewise, we evaluated commercially sensitive Siemens tender documents, revealing high overall performance, with particularly strong results in Few-shot learning settings as shown in Table 5. These models are evaluated for tender entity extraction from a restricted dataset of 30 proprietary documents through Zero, One, and Few-shot testing.

Zero-shot: The Zero-shot evaluation revealed PHI-3 and LLama-3 as top performers, while achieved F1 scores of 96.97% and 96.95% without needing any in-context examples. Similarly, the Falcon-3 and Deepseek-R1 models demonstrated outstanding performance through their F1 scores of 96.55% and 91.27%, which indicated their ability to detect regular patterns without prior exposure. In contrast, Mistral achieved a lower F1 score of 91.57% because it might have experienced underfitting during domain adaptation when no prior examples were available.

One-shot: Similarly, with the demonstration of an example (One-shot learning) led to performance enhancements across all models, particularly Deepseek-R1, achieving an F1 score of 94.07% and Mistral reaching 97.11%, LLama-3 and Falcon-3 scoring an almost perfect F1 of 98.88% and 98.11%. These models demonstrated their quick learning ability through contextual cues when only one in-domain sample was provided. The F1 score of PHI-3 reached 95.35% in the One-shot setting, but its performance increase was less than other models because it relied on thoughtful generalization from minimal context.

Few-shot: Interestingly, all models reached their peak performance in the Few-shot environment. The Few-shot learning produced the best results for Mistral and Falcon-3 while achieving F1 scores of 99.25%, which demonstrated their strong contextual learning abilities when multiple examples were provided. The performance of LLama-3 remained high at an F1 score of 98.88%, which depicted its consistent results. Moreover, Deepseek-R1 and PHI-3 experienced F1 score reductions from One-shot to Few-shot (92.22% and 93.14% respectively) because of experiencing pattern recognition disruption from the multiple slightly different examples in a restricted domain. Similarly, we utilized a PDF text extraction tool such as PDFMiner. However, during the extraction process, much of the original structure and contextual integrity of the text were lost, making it difficult to interpret for both humans and machines. Without restructuring or restoring the logical flow of the extracted text, we fed it directly into the LLMs. This lack of preprocessing may have contributed to the challenges the models faced in accurately understanding and extracting entities, even when provided with distinct examples.

4.5 Open-Source Vs Closed-Source LLMS Comparative Analysis

Our research aimed to investigate how open-source Large Language Models (LLMs) handle commer-

Methods	Deepseek-R1			LLAMA-3			PHI-3			Mistral			Falcon-3		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Zero-shot	85.13	87.57	85.28	**92.74**	**90.19**	**91.18**	83.17	81.35	81.99	86.46	83.71	84.91	89.06	89.92	89.05
One-shot	84.30	81.23	82.60	85.02	72.47	78.19	81.12	78.33	79.42	84.95	82.99	82.49	**86.50**	**96.62**	**86.93**
Few-shot	92.62	97.37	94.13	93.95	97.17	94.89	94.95	97.23	95.96	96.77	98.91	97.72	**99.35**	**98.28**	**98.80**

Table 4: Model Performance Comparison Across Various LLMS on Singapore Government Procurement Datasets

Methods	Deepseek-R1			LLAMA-3			PHI-3			Mistral			Falcon-3		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Zero-shot	96.35	86.82	91.27	96.21	**97.70**	96.95	**98.46**	95.52	**96.97**	91.94	91.20	91.57	96.92	96.18	96.55
One-shot	98.35	90.15	94.07	**98.50**	**99.25**	**98.88**	96.85	93.90	95.35	98.48	97.74	97.11	98.48	97.74	98.11
Few-shot	98.41	87.30	92.22	**99.25**	98.50	98.88	97.67	89.51	93.14	**100**	**99.25**	**99.63**	**100**	**99.25**	**99.63**

Table 5: Model Performance Comparison Across Various LLMS on Siemens Datasets

cially confidential data. The privacy restrictions of our data prevented us from using closed-source LLM applications such as ChatGPT for processing sensitive datasets. We began our analysis by manually anonymizing the confidential Siemens data before testing ChatGPT on this data for tender entity extraction. The analysis of the Singapore Procurement dataset using ChatGPT served as a broader assessment in addition to the processing of this dataset. The evaluation method enabled us to assess how ChatGPT (closed-source) performed relative to different open-source LLMs. We chose open-source models that achieved the highest F1-scores in all three settings (Zero-shot, One-shot, and Few-shot) across both datasets (see Sections 4.3 and 4.4).

As depicted in Figure 1, the open-source dataset showed ChatGPT (GPT-4) outperforming overall open-source LLMs in each of the learning environments. Consequently, GPT-4 achieved an outstanding F1-score of 96.64% in the Zero-shot setting that exceeded LLAMA-3's 91.18%, thus demonstrating superior generalization capabilities without any examples. Moreover, GPT-4 achieved the highest F1-score of 91.61% in One-shot tasks, but Falcon-3 achieved 86.93%, which indicates that GPT-4 adapts better with minimal contextual information. The F1 score of GPT-4 reached 99.15% in the Few-shot setting while Falcon-3 scored 98.8% in this same scenario. The performance results indicate that GPT-4 maintains high consistency across different supervision levels when operating with large, diverse datasets. The open-source LLMs demonstrated acceptable performance while Falcon-3 displayed solid Few-shot learning performance but struggled in Zero and One-shot tasks.

In contrast, GPT-4 (ChatGPT) dominated all testing conditions with the closed-source Siemens dataset by achieving 100% F1 in Few-shot tasks and 99.63% F1 in One-shot tasks. The Zero-shot performance of GPT-4 reached 97.34% F1, although PHI-3 outperformed it at 96.97%, indicating PHI-3 has better capability to process unknown data, but GPT-4 reaches 100% Few-shot accuracy, proving its better ICL abilities. LLAMA-3 showed exceptional One-shot performance with 98.88% F1, but Mistral obtained 99.63% F1 in Few-shot settings because it learned better from multiple examples.

Overall, GPT-4 demonstrated superior performance to all open-source models on both datasets by showing exceptional results, especially on the small sensitive Siemens data through its strong Few-shot and One-shot performance. The open-source LLMs LLAMA-3, Falcon-3, and Mistral delivered strong performance on both datasets yet failed to match GPT-4's performance, particularly when Zero-shot and One-shot settings were applied. The Few-shot capabilities of Falcon-3 and Mistral proved exceptional because they effectively used contextual examples, while LLAMA-3 demonstrated strong performance across datasets with notable One-shot learning abilities. GPT-4 demonstrates superior performance in smaller domain-specific datasets because its closed-source architecture benefits from extensive fine-tuning and diverse instructions and proprietary alignment approaches that lead to better results than open-source models.

5 Discussion, Conclusion and Future work

The discussion section presents various important findings regarding capabilities, limitations, and comparative analysis of open-source and closed-source LLMS on open-source and commercially sensitive data for the TNER task. First, the study

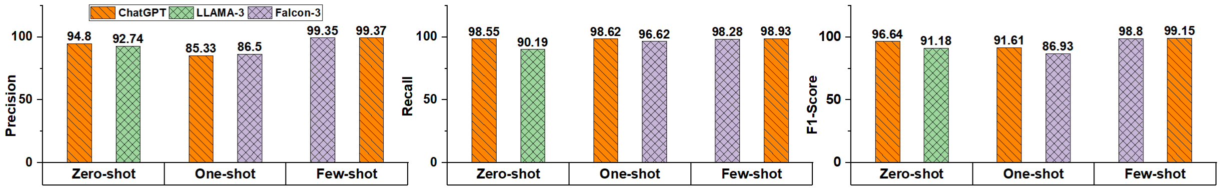

Figure 1: Performance comparison of open-source and closed-source LLMs on Singapore Procurement(Open-source) data for TNER Tasks

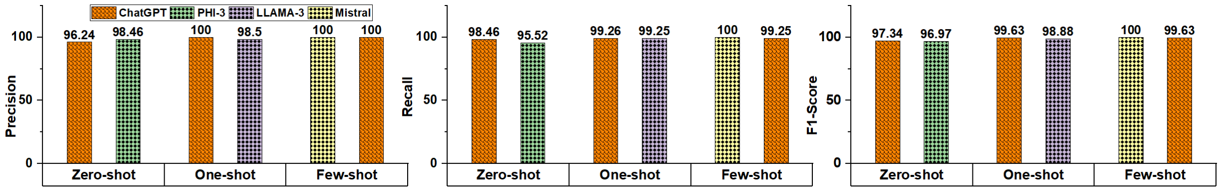

Figure 2: Performance comparison of open-source and closed-source LLMs on Siemens (closed-source) data for TNER (structured) tasks.

emphasizes that LLMs are trained on open-source data, encompassing broad general information, which allows them to respond across multiple domains, but this general approach leads to inaccurate or unhelpful answers in specific tasks. The adaptation of LLMs for specialized tasks like tender named entity recognition (TNER) requires fine-tuning with domain-specific data, although this process needs labeled data as well as domain expertise and time, which creates significant challenges. In this study, we have evaluated and categorized LLMs into two broad categories: Open-source LLMs, which can be run and deployed locally with limited hardware resources, and closed-sourced models, which require an API for accessibility, creating privacy risks and cost challenges. Data privacy is primarily concerned for sensitive domains such as tenders because tender documents contain confidential information, which must be handled carefully. Additionally, LLMs face challenges when processing tender data because each organization exhibits diverse complex structures where LLMs trained on open-source data do not prepare them to identify unique entities like tender names and tenderer personal details .

This study evaluated two different datasets, including an open-source and a closed-source sensitive data from a commercial entity. The open-source data was available in structured form, we developed rule-based data augmentation algorithm and transformed only 200 instrances into an unstructured format. Similarly, closed-source data was available in complex table format, after extraction the structure of the text is lost which hampers model performance. Further, we manually anonymized the sensitive data and processed it through GPT-4. The evaluation process included Zero, One, and Few-shot model testing. The open-source models LLAMA-3 and Falcon-3 achieved good results on the open-source dataset, but these models experienced difficulties with specific entities because of their limited exposure. In contrast, GPT-4 achieved superior performance than all open-source models across both datasets because of its extensive parameter size, reaching 1.7 trillion compared to millions. However, open-source models achieved remarkable results in TNER tasks despite their limited fine-tuning capabilities.

In the future, we will fine-tune small language models like BERT on open-source data related to the tender domain and then, by following a transfer learning approach, we will evaluate them on closed-source data for entity extraction. Similarly, we aim to extend this research to other domains where data is not easily accessible, such as the clinical domain. Our goal is to explore how LLMs can be used without tuning or training from scratch through effective prompt engineering to extract valuable insights and information.

Acknowledgments

The authors gratefully acknowledge **Siemens** for supplying the tender dataset used in this study. The availability of this data was vital for conducting meaningful experiments and achieving the research objectives.

References

Asim Abbas, Mark Lee, Niloofer Shanavas, Venelin Kovatchev, and Mubashir Ali. 2025. Structured tender entities extraction from complex tables with few-short learning. *COLING 2025*, page 59.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yunfei Ji, Chao Tong, Jun Liang, Xi Yang, Zheng Zhao, and Xu Wang. 2019. A deep learning method for named entity recognition in bidding document. In *Journal of Physics: Conference Series*, volume 1168, page 032076. IOP Publishing.

Singpore Government Procurement Dataset on Kaggle. 2024. Kaggle link: https://www.kaggle.com/datasets/shivamb/government-procurement-dataset, 2015-2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Chuan Li. 2020. Openai's gpt-3 language model: A technical overview. *Blog Post*.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. 2021. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Lucia Siciliani, Vincenzo Taccardi, Pierpaolo Basile, Marco Di Ciano, and Pasquale Lops. 2023. Ai-based decision support system for public procurement. *Information Systems*, 119:102284.

Esa Toikka et al. 2021. Information extraction from procurement contracts. Master's thesis.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Dengya Zhu, Sirui Li, Nik Thompson, and Kok Wai Wong. Open-source large language models excel in named entity recognition.