Struct2Unstruct: Creating Tender NER Datasets from Structured Procurement Records using Large Language Models

Asim Abbas[1]*, **Mark Lee**[1], **Niloofer Shanavas**[2], **Venelin Kovatchev**[1], **Mubashir Ali**[1]
[1]School of Computer Science, University of Birmingham, B15 2TT, UK
[2]School of Computer Science University of Birmingham, Dubai, UAE
axa2233@student.bham.ac.uk, {m.g.lee, n.shanavas, v.o.kovatchev, m.ali.16}@bham.ac.uk

Abstract

Named Entity Recognition (NER) in the tender and procurement domain is critical for tasks such as contract monitoring, supplier analysis, and compliance tracking. However, unlike general-purpose NER, no open-source datasets exist for Tender NER, largely due to data sensitivity and confidentiality restrictions. This scarcity limits the development of automated entity extraction models. To address this gap, we propose struct2unstruct, a data preparation pipeline that generates and annotates tender-specific datasets using large language models (LLMs). Starting from structured procurement data published by the Singapore government (2015–2021) available in English language, we employ Llama-3 to generate synthetic tender narratives in multiple writing styles, ensuring each contains at least one tender-related entity. Post-processing steps correct inconsistencies in dates, symbols, and entity formats. Entities are then annotated using a BIO tagging scheme through deterministic alignment with structured fields, followed by expert validation to ensure accuracy. This study focuses on data preparation and evaluation, not model training. The resulting dataset provides a scalable resource for future Tender NER research in low-resource environments. By releasing both the dataset and pipeline as open-source resources, we establish a foundation for advancing domain-adapted information extraction and automated tender entity recognition.

Keywords: Named Entities Recognition, Data Augmentation, Large Language Models, Data Preparation

1. Introduction

Named Entity Recognition (NER) is one of the cornerstone tasks in Natural Language Processing (NLP). While general-purpose corpora such as CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), OntoNotes (Hovy et al., 2006), and WNUT (Tabassum et al., 2020) have enabled significant progress in NER, specialized domains like public procurement and tenders remain under-resourced and less explored compared to well-studied areas such as clinical or legal NLP. This scarcity stems from the sensitivity of procurement data, inconsistencies in document structure, and the absence of publicly available annotated corpora (Taoufik and Azmani, 2024; Abbas et al., 2025b).
Tenders are formal requests for proposals or offers, typically issued by organizations or government agencies seeking goods, services, or works (Cao, 2025; Siciliani et al., 2023). Extracting structured information such as supplier names, buyer agencies, contract values, and deadlines from tender documents is critical for applications in market transparency, supplier analysis, and fraud detection (Toikka et al., 2021). However, tender documents vary widely in format across institutions and are rarely shared publicly. Due to commercial sensitivity, existing NER models trained on general corpora often fail to generalize to procurement text. Manual extraction, on the other hand, is costly, time-consuming, and prone to errors.

Data annotation, a crucial step in data preparation, is widely recognized as labor-intensive (Furche et al., 2016). Surveys shows that data scientists spend nearly 80% of their time on tasks such as cleaning, collating, and annotating data (Fernandes et al., 2023). While indispensable, this process remains a major bottleneck in building domain-specific AI systems. Recent advances in Large Language Models (LLMs) have opened new opportunities by generating synthetic corpora enriched with diverse and contextually appropriate entity mentions (Brown et al., 2020). This approach improves adaptability in low-resource domains, enabling the creation of training data at scale (Dao et al., 2025).
Nonetheless, reliance on LLMs introduces challenges. LLMs are prone to hallucinations, producing plausible but factually incorrect content, which complicates entity alignment with structured source data (Dao et al., 2025). Additionally, variations in phrasing and prompt adherence can further hinder deterministic span-level tagging (Hu et al., 2024). Moreover, domain-specific terminologies often lack grounding in pre-trained LLMs, leading to misclassifications or semantic drift (Ling et al., 2023). Likewise, automated pipelines thus require extensive post-processing and human oversight to ensure annotation quality and label consistency (Klie et al., 2024). Finally, models trained predominantly on synthetic data may overfit to artificial linguistic patterns and fail

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 96–106 May 11, 2026.

to generalize to real-world documents (Dao et al., 2025).

To address these limitations, we propose struct2unstruct, a data augmentation pipeline that transforms structured tender data into synthetic unstructured narratives for the Tender NER task. The proposed study makes following key contributions:

- **First open synthetic tender dataset:** We introduce the first publicly available dataset for Tender NER, generated from structured procurement records, addressing the lack of accessible corpora in this sensitive domain.
- **LLM-based pipeline for low-resource NER:** We design a structured-to-unstructured generation method using Llama-3 that produces diverse, entity-grounded narratives without requiring manual annotations, making it applicable to other low-resource domains.
- **Efficient annotation via entity alignment:** By aligning generated text with structured records using deterministic span-matching, we minimize manual effort while maintaining high annotation quality verified by domain experts.

The remainder of this paper is organized as follows: Section 2 reviews related work on NER in both general and tender domains. Section 3 presents our proposed data preparation pipeline for Tender NER. Section 4 reports experimental and evaluation results. In Section 5, we discuss the limitation and future plan of study and finally, Section 6 concludes the study.

2. Related Work

Reviewing prior literature reveals that research in NER has been advanced through a series of shared tasks and benchmark datasets such as CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Hovy et al., 2006), which have shaped progress in both general-purpose and specialized contexts. However, no shared task has explicitly addressed NER in the procurement or tender domain. Similarly, existing corpora are largely drawn from newswire, conversational, or web text, and while meticulously annotated by experts, they provide little coverage of procurement-specific entities such as tender title, IDs, contract values, or agency names etc.

On the other-hand, NER systems rely on high-quality annotated datasets, but such resources are scarce in sensitive domains like procurement, where manual annotation is costly and often infeasible. To alleviate this bottleneck, synthetic data generation using LLMs has emerged as an alternative to manual labeling (Xu et al., 2024). However generating synthetic data for domain specific NER task is also a challenge because its not just generating natural language text but also must ensure the entity correctness, contextual consistency, and domain relevance. Unlike general text generation, NER data must contain entities that are correctly labeled and naturally embedded within the context, reflecting real-world sentence structures (Dao et al., 2025).

Earlier data augmentation strategies for low-resource NER primarily involved transformations of existing text. For instance back-translation (Dai and Adel, 2020) introduced lexical and syntactic variety while preserving entities. Contextual word substitution (Torres et al., 2024) enriched datasets with semantic alternatives for non-entity tokens. Distant supervision (Xiaoqin et al., 2021) aligned unlabeled text with knowledge bases to automatically annotate entities, albeit with noisy results. Paraphrasing (Sharma et al., 2022) generated alternative sentence formulations to enhance linguistic diversity. While effective in expanding training data, these methods generally fail to capture the specialized terminology and structures found in procurement texts.

Recent advances in LLMs enable more sophisticated augmentation approaches tailored to scarce domains (Abbas et al., 2025a). Few-shot prompting (Liu et al., 2022) uses annotated examples to guide the creation of entity-rich sentences. Similarly, Schema-driven generation (Tsai et al., 2021) converts structured records into narrative text while enforcing stylistic constraints, offering strong alignment with our setting. Moreover, Domain-grounded generation (Liu et al., 2020) incorporates real-world records to improve entity accuracy and reduce hallucination. Additional techniques such as contextual similarity–based augmentation and transformer-based text generation (Yili and Haonan, 2023; Abbas et al., 2024) further demonstrate significant improvements in biomedical and other specialized domains. Despite these advances, LLM outputs often deviate from prompt specifications or introduce irrelevant content (Min et al., 2023), motivating the need for robust pipelines that ensure entity correctness and contextual fidelity.

Our work addresses this gap by leveraging structured tender data (e.g., contract types, buyer names, supplier names etc) as input prompts to guide LLM-based text generation. This ensures that synthetic examples include target entities relevant to the tender domain. We further apply deterministic span alignment and post-processing to produce high-quality BIO-formatted annotations, thereby addressing data scarcity in Tender NER while preserving domain-specific integrity.

3. Proposed Pipeline

In this study, we present a data preparation pipeline for Tender NER see Figure 1. The tender domain is highly sensitive, making access to real tender documents extremely limited. Even when data is available, such as from commercial sources, it is often too restricted to be shared openly. To address this challenge, we explore an alternative: leveraging a small set of public procurement records and augmenting them with an LLMs (Llama-3:8b) to generate synthetic yet realistic tender narratives. This approach not only contributes to advancing Tender NER but can also be generalized to other domains where annotated data is scarce or confidential. The pipeline consists of four stages: structured data acquisition, structured-to-unstructured transformation, content repair, and entity annotation and evaluation. This approach supports reproducible research in Tender NER and can be adapted to other low-resource domains. The code and dataset is available on Github [1]

3.1. Structured Data Acquisition

We use a publicly available procurement dataset released by the Singapore Government Procurement on Kaggle (Dataset, 2024) available in English language, covering tenders awarded between 2015 and 2021. The dataset contains 23,909 records in CSV format with seven structured fields like *Tender No, Tender Description, Agency (Buyer), Award Date, Tender Status, Supplier Name, and Awarded Amount*. During analysis, we found the *Tender Description* field is highly diverse and difficult to annotate, so we excluded it from the dataset for consistency. Finally, we have six entities included in the data preparation process.

3.2. Structured to Unstructured Data Transformation

An NER task requires unstructured text where entities are naturally embedded in context. However, our source dataset was only available in structured CSV format. To overcome this limitation, we developed the Struct2Unstruct pipeline, which converts structured tender records into realistic narratives using the open-source Llama-3:8b model. The pipeline is designed around four components: field mapping, writing pattern variation, generation constraints, and careful model selection.

- **Structured Data Field Mapping:** To ensure that generated text always contains relevant entities, we directly link structured fields to narrative outputs. In each generation step, one or more fields (e.g., Tender No, Supplier Name, Tender Amount) and their values are sampled from the dataset. These values are embedded into the prompt, guaranteeing that at least one tender-related entity is present in the final text. This approach improves consistency and ensures that the data remains useful for Tender NER training.
- **Writing Pattern Variation:** Tender documents differ widely in format and tone depending on the issuing organization. To reflect this diversity, we designed a set of writing patterns, including *formal, descriptive, regulatory, announcement, technical, press release, and project proposal styles*. During text generation, one style is selected at random and applied to the prompt. This variation produces narratives that not only differ in content but also in structure and style, improving the robustness and generalization of NER models.
- **Text Generation Constraints:** LLMs often generate long or inconsistent outputs, which are difficult to annotate and more likely to contain hallucinations. To avoid these issues, we restricted Llama-3 to produce short outputs between one to three (1-3) sentences. Following this setup, we generated about 8,000 tender narratives, in between one to three sentences. This length constraint makes the data easier to annotate, reduces noise, and ensures efficiency in later training stages.
- **Model Selection:** For generation, we used Llama-3:8b, that is widely adopted in the research community and performs competitively compared to closed-source systems such as GPT-4. Its accessibility through frameworks like Ollama makes it practical for both high- and low-resource environments. Using prompt-tuning strategies, we guided Llama-3 to incorporate structured tender fields directly into the generated narratives. We used a temperature of 0.7 and nucleus sampling (top-p = 0.9) to promote lexical diversity while maintaining factual consistency and domain relevance.

In a nutshell, the Struct2Unstruct pipeline systematically transforms structured procurement records into unstructured, entity-rich narratives. By combining field mapping, stylistic diversity, length constraints, and an open-source generation model, it creates synthetic corpora that are realistic, consistent, and suitable for high-quality Tender NER training.

3.3. Data Pre-processing and Repairing

Following the generation of unstructured tender narratives using Llama-3, we obtained a total of 8,000

[1] Synthetic Dataset and Generation Code

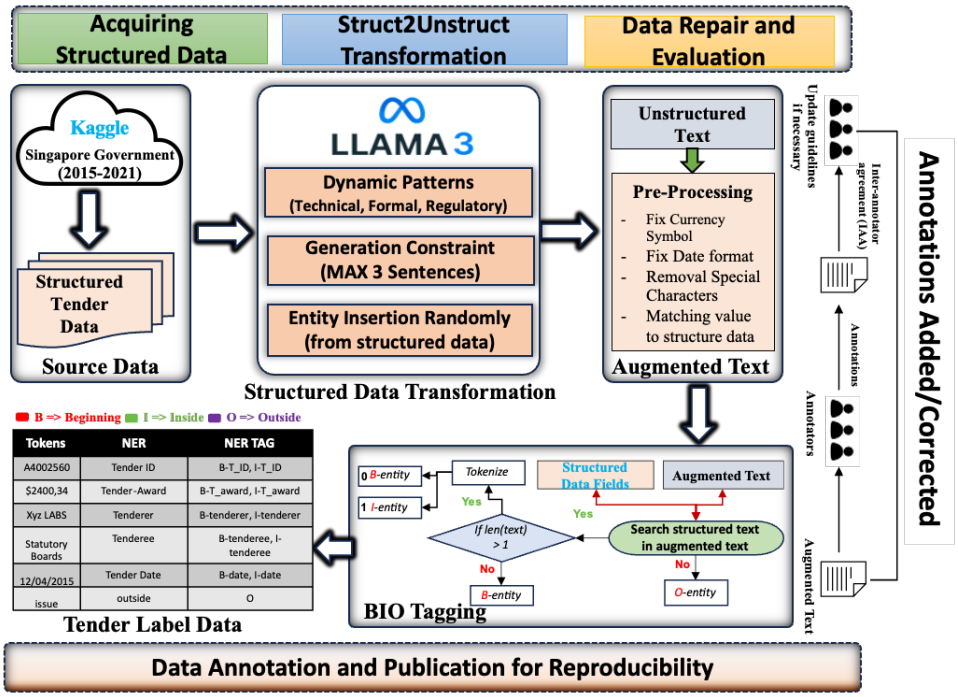

Figure 1: Detailed Workflow Diagram for Struct2Unstruct Data Transformation

augmented instances. Each record was designed to include at least one entity drawn from the structured tender dataset, ensuring relevance for downstream NER tasks. To maintain consistency in text length, half of the records were constrained to 1–2 sentences, while the remaining half comprised 2–3 sentences. Although the generated text generally adhered to these constraints, closer examination revealed inconsistencies and structural ambiguities that required manual or systematic pre-processing. A known limitation of LLMs is their occasional deviation from explicit instructions. In our case, Llama-3 sometimes altered entity formats or patterns: currency symbols were replaced or spelled out (e.g., "$" converted to "£" or "Euro"), numerical values were expressed in abbreviated or alternative forms (e.g., "2,800,000$" as "2.8 million dollars"), and dates were reformatted inconsistently (e.g., "23/07/2020" to "23 July, 2020" or "07/2020"). While these variations are contextually accurate, they introduce morphological inconsistencies that can hinder precise entity annotation.

To address these issues, we adopted a hybrid pre-processing strategy. Critical entities, including dates and currency amounts, were manually corrected to align with the structured source data, ensuring fidelity and consistency. Simultaneously, certain variations were retained to introduce linguistic diversity, allowing the NER model to generalize across multiple formats and representations. Additionally, extraneous special characters (e.g., "§", "<§") were removed to eliminate noise and maintain semantic clarity. As a result of this pre-processing and cleaning phase, the augmented tender texts became consistent, semantically meaningful, and sufficiently diverse for training robust NER models. This high-quality dataset provides a reliable foundation for downstream Tender NER tasks and ensures that models can accurately recognize entities across varying formats, styles, and representations typical of real-world tender documents.

3.4. Data Annotation and Publication

Annotating data for NER is a challenging task. Manual annotation is accurate but time-consuming and costly, particularly in domains such as procurement, where documents are long and specialized. Automatic annotation using LLMs is an alternative, but general-purpose models often make mistakes in sensitive or domain-specific contexts. To overcome these limitations, we designed a hybrid strategy that leverages structured data as a reliable source of entity information while automating alignment with unstructured, LLM-generated text. Our approach uses a heuristic matching algorithm to align structured fields such as *tender number, agency(Buyer), award date, tender status, supplier name, and awarded amount* with their corresponding mentions in the generated narratives. By aligning annotations to structured values, we reduced manual effort and ensured consistency across the dataset.

For labeling, we adopted the widely used BIO scheme (Begin, Inside, Outside), which offers a balance of simplicity and expressiveness. Each

generated tender text was tokenized using SpaCy, and entity values from the structured data were matched against token spans through a sliding window search. When a match was found, the first token was labeled with a B-tag, subsequent tokens with I-, and all other tokens with O. Records with unmatched or ambiguous spans were excluded to preserve annotation quality. This process produced high-precision, span-level annotations without requiring manual span marking.

The final dataset follows the standard IOB format proposed by Ramshaw and Marcus (1995). Each row contains two aligned columns: tokens and ner_tags. The tokens column stores the tokenized text, while the ner_tags column assigns the corresponding BIO labels. Entities such as *tender number, award date, and awarded amount* usually appear as single tokens and are therefore marked only with B-tags. In contrast, entities such as *supplier name, tender status, and agency* often span multiple tokens and thus include both B- and I-labels. In contrast, O-tagged text carries no specific entity information. In total, 11 entity types were annotated: B-TENDER_NO, B-AWARD_DATE, I-AWARD_DATE, B-TENDER_STATUS, I-TENDER_STATUS, B-SUPPLIER, I-SUPPLIER, B-AWARDED_AMT, I-AWARDED_AMT, B-AGENCY, and I-AGENCY. This dataset structure is fully compatible with modern NER frameworks, including Hugging Face datasets and CoNLL-style sequence labeling models.

3.5. Dataset Evaluation Strategy

Scientific research requires systematic validation of both methods and outcomes. To ensure the reliability of our proposed data generation and annotation pipeline, we designed a multi-level evaluation framework. This framework combines quantitative similarity measures, heuristic-based alignment, and expert validation to confirm the quality of the generated tender dataset.

At the first level of evaluation, we measured the diversity and clustering behavior of the generated tender documents. Our experimental design and visualization pipeline is shown below:

$$\text{Documents} \xrightarrow{\text{SBERT}} \mathbb{R}^{384} \xrightarrow{\text{UMAP}} \mathbb{R}^{2} \xrightarrow{\text{Analysis}} \text{Insights}$$

Where as , each document d_i in the dataset $D = \{d_1, d_2, \ldots, d_n\}$ was embedded into 384 dimensional vector space using Sentence-BERT (SBERT) as shown in Equ[1]:

$$\mathbf{e}_i = \text{SBERT}(d_i) \in \mathbb{R}^{384}, \tag{1}$$

To reduce dimenionality for visualization, we applied Uniform Manifold Approximation and Projection (UMAP):

$$\mathbf{e}_i^{2D} = \text{UMAP}(\mathbf{e}_i) \in \mathbb{R}^2, \tag{2}$$

where $\mathbf{e}_i^{2D}$ represents the two-dimensional projection of the original embedding $\mathbf{e}_i$. Clustering was then performed applying Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which assigns cluster labels c_i to each document:

$$c_i = \text{HDBSCAN}(\mathbf{e}_i) \in \{-1, 0, 1, 2, \ldots, k\}, \tag{3}$$

where $c_i = -1$ indicates noise points, and k is the number of discovered clusters. Finally, for any two document $\mathbf{e}_i$ and $\mathbf{e}_j$, the cosine similarity is computed as:

$$\text{sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \, \|\mathbf{e}_j\|} = \frac{\sum_{k=1}^{384} e_{i,k} \, e_{j,k}}{\sqrt{\sum_{k=1}^{384} e_{i,k}^2} \sqrt{\sum_{k=1}^{384} e_{j,k}^2}}. \tag{4}$$

This ensured that the generated dataset was both diverse and contextually coherent within the tender domain.

In the next step, we validated entity correctness through a heuristic matching approach. Entities generated by Llama-3 were compared against structured references and categorized into six match types: Exact, Reformatted, Normalized, Type-preserving but semantically altered, Hallucinated, and Not Found (see Table 1).

Let $N = 8{,}000$ be the total number of records and $E = 6$ the number of entity types, yielding

$$T = N \times E = 48{,}000 \tag{5}$$

total evaluations. For each entity type $i \in \{1, \ldots, E\}$, let F_i denote the number of correctly detected entities. The overall detection rate is then given by:

$$\text{Overall\_Detection\_Rate} = \left(\frac{\sum_{i=1}^{E} F_i}{T} \right) \times 100\%. \tag{6}$$

This allowed us to quantify performance while accounting for variations such as date reformatting or currency normalization.

Finally, we conducted human validation to assess annotation quality. Domain experts reviewed the BIO-tagged outputs and their agreement with the automatic annotations was measured using Inter-Annotator Agreement (IAA). An IAA score above 81% to 100% was considered a perfect indicator of reliability. Datasets surpassing this threshold were judged suitable for publication and for use in training AI models for Tender NER.

% smaller text for ACL style % row spacing

Match Type	Description	Performance Matrices
Exact	Entity is identical to gold	$P(\text{Exact}) = \left(\sum_{i=1}^{E} E_i\right)/T \times 100\%$
Reformatted	Formatting changed (e.g., date styles, punctuation)	$P(\text{Reformatted}) = \left(\sum_{i=1}^{E} R_i\right)/T \times 100\%$
Normalized	Value reformulated (e.g., "2800000" → "2.8 million")	$P(\text{Normalized}) = \left(\sum_{i=1}^{E} \text{Norm}_i\right)/T \times 100\%$
Type-preserving but semantically altered	Same entity type, but meaning changed (e.g., $ → £)	$P(\text{Type-preserved}) = \left(\sum_{i=1}^{E} \text{TP}_i\right)/T \times 100\%$
Hallucinated	Entity not in gold dataset and intended)	$P(\text{Hallucinated}) = \left(\sum_{i=1}^{E} \text{H}_i\right)/T \times 100\%$
Not Found	Entity not in augmented text	$P(\text{Not found}) = \left(\sum_{i=1}^{E} \text{NF}_i\right)/T \times 100\%$

Table 1: Match types, descriptions, and corresponding performance matrices.

4. Dataset Construction and Evaluation

In this study, we evaluated the dataset at multiple levels. First, we assessed the robustness of the LLM-generated data by measuring variation and domain-specific consistency using different metrics. Next, we examined the semantic and syntactic correctness of tender entities in the generated text through an advanced evaluation approach. Finally, we validated the entity annotations by calculating Inter-Annotator Agreement (IAA). These evaluation steps are visually summarized below.

4.1. Domain-specific data variation and consistency

The cluster visualization shown in Figure 2, the generated tender documents group together based on their writing style, with each point representing one document positioned according to its semantic similarity. We used SBERT embeddings (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019) to capture the semantic meaning of each document and then applied UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018), a method that reduces the high-dimensional data into two dimensions while preserving similarity patterns. Subsequently, the HDBSCAN (Hierarchical Density-Based Spatial Clustering) (Campello et al., 2013) algorithm is used to detect clusters of documents with similar writing styles and to identify outliers that do not fit well into any group. The resulting interactive scatter plot shows that styles such as formal and press release overlap heavily in the center due to shared linguistic features, while others like technical, government report, and project proposal form smaller, more distinct clusters. We can also see a few outliers that do not fit well with the main clusters, likely because those texts were written in a very specific or unusual way. Although all documents share a common tender-related vocabulary and professional tone, the distribution demonstrates contextual diversity across styles, with meaningful differences in tone and structure alongside areas of overlap. This suggests that the data successfully captures variation in writing style while maintaining consistency within the tender domain.

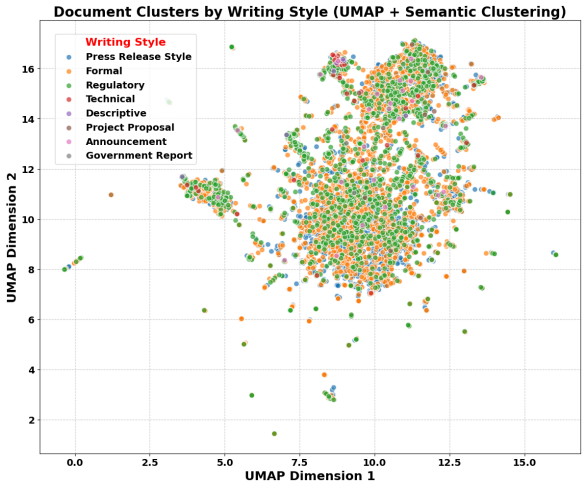

Figure 2: Visualizing Domain-Specific data variation across Writing Styles using SBERT and HDBSCAN

4.2. Intra-style Similarities by Writing Style

The intra-style mean similarity analysis highlights how consistent documents are within each writing style. Formal (3354 documents) and press release style (3218 documents) dominate the dataset, producing 5.6 million and 5.1 million document pairs, respectively, with mean similarity scores of 0.49 and 0.48 as shown in Table 2. These large numbers of pairs reveal moderate consistency but also considerable variation, as shown by wide ranges between minimum and maximum values. Regulatory texts (1351 documents, 0.9 million pairs) show

slightly higher internal similarity (mean 0.52), while technical (33 documents, 528 pairs) and announcement (26 documents, 325 pairs) achieve stronger consistency (means 0.57 and 0.53) despite their smaller sample sizes. Project proposal (8 documents, 28 pairs) shows lower similarity (0.45), reflecting more variation in this style, whereas descriptive (7 documents, 21 pairs, mean 0.65) and government report (3 documents, 3 pairs, mean 0.63) exhibit the highest similarity, which is expected due to fewer combinations and less stylistic diversity. The disproportionate number of formal and press release documents arises from the LLM's tendency to favor professional and announcement-like tones, which closely align with the tender domain. Even though styles were randomly assigned from the list, the model more frequently generated text in these styles because they overlap with its training distribution and the common linguistic features of tender-related documents.

4.3. Entity-Level Semantic and Syntactic Evaluation

The entity-level semantic and syntactic evaluation highlights both the overall distribution of match types and the detection performance of individual entities. At the aggregate level, exact matches account for 22.2% of the evaluated cases, with reformatted and normalized matches contributing 8.6% and 8.8,% respectively. Type-preserved but semantically altered cases make up 13.4% of the total, while instances classified as hallucinated are 3%. A substantial portion, representing 44.2% of all cases, falls into the not found category. This outcome does not indicate system failure but rather reflects that these entities were not available in the augmented text, since entities were randomly added during the large language model augmentation process see Figure 3.

At the entity-specific level, the detection rates vary considerably across different entity types. The highest detection performance is observed for tender_status with a rate of 86.3%, followed by supplier_name at 69.5%. Awarded amount achieves a moderate detection rate of 50.7%, while tender_no and award date yield lower rates of 38.5% and 37.2% respectively. Agency (Buyer) exhibits the lowest detection rate at only 11.2%. These results indicate not only the system's variable ability to detect different entity types but also the fact that the large language model augmented text contained these entities in such proportions, as they were included randomly during augmentation. This explains both the distribution across categories and the variation in detection rates across entity types see Figure 3.

4.4. Annotator Agreement and Reliability

To assess the reliability of the tender entity annotations, we conducted an inter-annotator agreement analysis using two independent annotators. The confusion matrix shows that the two annotators agreed on the majority of cases, with an overall observed agreement of 98.53%. However, a portion of this agreement (83.14%) could be expected to occur by chance. To account for this, we calculated Cohen's κ, which provides a more conservative measure of inter-rater agreement by adjusting for chance effects. The resulting value of $\kappa = 0.913$ indicates an almost perfect level of agreement between the annotators. The standard error of the kappa estimate was very small ($SE = 0.001$), and the corresponding 95% confidence interval ranged from 0.911 to 0.915. This narrow interval suggests that the reliability estimate is highly stable and statistically meaningful.

Although the raw agreement rate is very high, the kappa statistic confirms that this agreement remains strong even after correcting for chance effects. In other words, while some agreement can be attributed to both annotators assigning labels within common or dominant categories, the high kappa value demonstrates that the consistency between annotators is substantial and not merely due to baseline agreement. The strong kappa score therefore, reflects that the annotators applied the entity labels in a highly consistent and reliable manner, with only minimal variation. From a practical perspective, this result is highly encouraging for downstream use of the annotated dataset. The very high observed agreement demonstrates that the annotations are largely consistent, and the statistically significant and strong kappa value confirms that the agreement is not driven by chance. These findings indicate that the annotation guidelines are well defined and effectively followed. Nevertheless, continuous refinement of entity definitions may further improve clarity, particularly for rare or ambiguous entity cases, to ensure sustained annotation quality in future expansions of the dataset.

5. Discussion, Limitation and Future Work

NER in sensitive, low-resource domains like public procurement faces significant bottlenecks due to data scarcity, confidentiality constraints, and inconsistent document structures. Consequently, there are currently no open shared tasks or benchmark datasets available specifically for the tender domain. While LLMs offer strong zero-shot and few-shot capabilities (Abbas et al., 2025a), they frequently struggle with domain-specific terminology, hallucinate irrelevant details, and fail to maintain

Writing Style	Docs	Pairs	Mean Similarity	Range
Formal	3354	5,622,981	0.49	[0.03–0.98]
Press Release Style	3218	5,176,153	0.48	[0.01–0.98]
Regulatory	1351	911,925	0.52	[0.04–0.97]
Technical	33	528	0.57	[0.21–0.92]
Announcement	26	325	0.53	[0.22–0.80]
Project Proposal	8	28	0.45	[0.29–0.62]
Descriptive	7	21	0.65	[0.49–0.79]
Government Report	3	3	0.63	[0.57–0.68]

Table 2: Summary statistics by writing style. Reported values include the number of documents, generated pairs, mean, and range(min/max values).

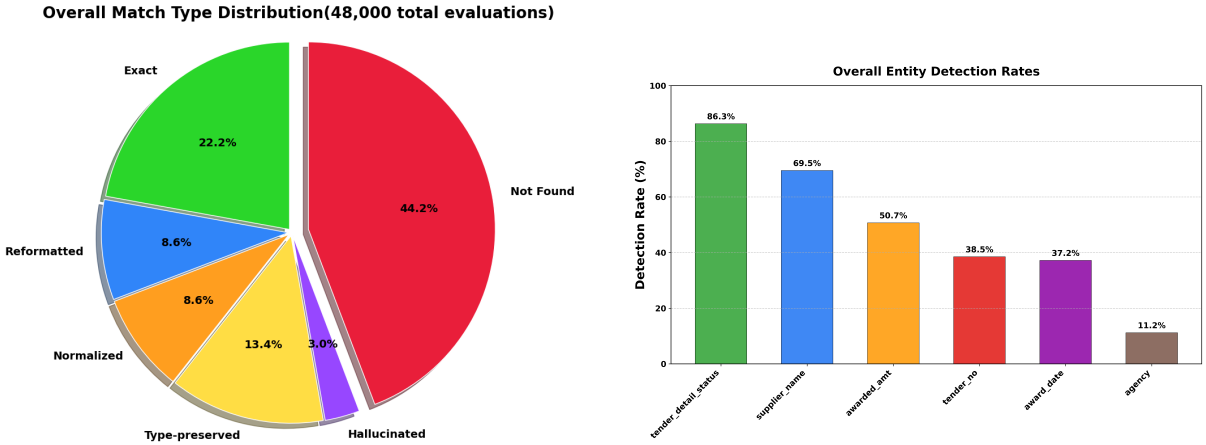

Figure 3: Overall match type distribution and entity level detection rates beyond match type

Annotator A \ Annotator B	Entity	Not Entity	Total	Observed Agreement (%)	98.53%
Entity	33,227	3660	36,887	Agreement Expected by chance (%)	83.14%
Not Entity	2031	349,338	351,369	Kappa Score	0.913
Total	35,258	352,998	388,256	Standard Error of Kappa	0.001
95% confidence interval: From 0.911 to 0.915					

Table 3: Tender Entities annotation agreement confusion matrix between two annotators with observed and expected agreement, Cohen's κ, and its standard error.

consistent formatting without strict guidance.

To bridge this gap, this study introduced the struct2unstruct pipeline, utilizing a locally deployed Llama-3 (8B) model to ensure data privacy while producing a cost-effective, synthetic corpus. By dynamically mapping structured fields from the Singapore Government Procurement dataset (2015–2021) into constrained 1-3 sentence narratives, the pipeline guarantees the inclusion of valid tender entities while mitigating LLM hallucinations. A key strength of this approach is the demonstrated stylistic and semantic diversity of the generated dataset. As evidenced by our semantic clustering (UMAP and HDBSCAN) and intra-style similarity analyses, the pipeline successfully generated distinct document clusters across multiple styles—such as formal, technical, regulatory, and press releases—while maintaining high semantic coherence within those styles. Furthermore, our multi-level evaluation framework validated the robustness of the data. The entity-level evaluation revealed realistic variations in entity detection rates—such as an 86.3% detection rate for tender status compared to 11.2% for agency names—which accurately reflects the random distribution of entities introduced during the augmentation process. Most notably, the high inter-annotator agreement (Cohen's $= 0.913$) proves that combining heuristic span-matching with structured references produces highly reliable, scalable BIO-tagged annotations, drastically reducing the labor-intensive manual effort typically required for domain-specific NER.

Despite these contributions, our study has some

limitations. The primary focus of this study was the creation and evaluation of the data preparation pipeline. We did not fine-tune or train a base transformer model to establish baseline NER performance on this new dataset. Although the dataset encompasses diverse writing styles, the underlying entities and contextual seeds are derived exclusively from Singapore Government procurement records. Consequently, the dataset may not fully capture the diverse structural and terminological variations present across all global organizations and procurement contexts. While our hybrid preprocessing resolved many LLM-induced morphological inconsistencies (e.g., varied date formats or currency symbols), the automated BIO-tagging still relies on a sliding-window heuristic matching algorithm, which inherently required excluding records with ambiguous or unmatched spans to preserve quality.

The current dataset establishes a foundational step toward standardized tender datasets, but it should be viewed as an initial benchmark rather than a complete solution. We plan to train and fine-tune state-of-the-art transformer-based NER models on this newly prepared dataset to evaluate baseline performance for automated tender entity recognition. Similarly, we aim to broaden the scope and generalizability of the benchmark by integrating our synthetic tender dataset with existing open-source NER corpora, making it suitable for both specialized and broader NER tasks. Ultimately, we intend to extend this pipeline by training models on restricted, highly sensitive procurement data provided by a commercial partner, thereby advancing domain-specific NER for real-world industry applications.

6. Conclusion

This study addressed the persistent challenge of NER in the tender domain, where data scarcity and confidentiality limit model performance and dataset availability. To mitigate this, we developed a pipeline for generating a Tender NER dataset by combining structured tender data from Singapore with synthetic data produced by the open-source Llama-3 model. The approach introduced textual diversity while maintaining entity accuracy through explicit generation constraints. To ensure data reliability, we implemented a multi-level evaluation framework integrating similarity analysis, heuristic alignment, and expert review. Additionally, data were formatted in BIO structure using SpaCy and validated by domain experts to enhance quality. Although the dataset and models remain in early stages, this work establishes a foundational step toward creating standardized, domain-specific tender datasets. Future research will focus on fine-tuning transformer-based models and expanding dataset generalization for broader applicability across procurement and related domains.

7. Ethical Consideration

This study used an open-source dataset publicly available on Kaggle that do not include personally identifiable information. The data were used intended research purposes.

Human participants were involved only as domain experts for annotation validation. All experts participated voluntarily and provided informed consent before their involvement. Their feedback was used solely for research validation, and no personal or sensitive data were collected or disclosed.

ChatGPT was used exclusively to refine the English language and presentation of the manuscript. The conceptualization, design, experimental analysis and execution of the study were entirely performed by the author.

8. References

Asim Abbas, Venelin Kovatchev, Mark Lee, Niloofer Shanavas, and Mubashir Ali. 2025a. Harnessing open-source LLMs for tender named entity recognition. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Asim Abbas, Mark Lee, Niloofer Shanavas, and Venelin Kovatchev. 2024. Clinical concept annotation with contextual word embedding in active transfer learning environment. *Digital Health*, 10:20552076241308987.

Asim Abbas, Mark Lee, Niloofer Shanavas, Venelin Kovatchev, and Mubashir Ali. 2025b. Structured tender entities extraction from complex tables with few-short learning. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 59–67, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based

on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Danrun Cao. 2025. *Information extraction from heterogeneous multilingual documents for the exploitation of a global tender database*. Ph.D. thesis, Université de Bretagne Sud.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Dao, Hiroki Teranishi, Yuji Matsumoto, Florian Boudin, and Akiko Aizawa. 2025. Overcoming data scarcity in named entity recognition: Synthetic data generation with large language models. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 328–340, Viena, Austria. Association for Computational Linguistics.

Singpore Government Procurement Dataset. 2024. Kaggle link.

Alvaro AA Fernandes, Martin Koehler, Nikolaos Konstantinou, Pavel Pankin, Norman W Paton, and Rizos Sakellariou. 2023. Data preparation: A technological perspective and review. *SN Computer Science*, 4(4):425.

Tim Furche, George Gottlob, Leonid Libkin, Giorgio Orsi, and Norman Paton. 2016. Data wrangling for big data: Challenges and opportunities. In *Advances in Database Technology—EDBT 2016: Proceedings of the 19th International Conference on Extending Database Technology*, pages 473–478.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Low-resource ner by data augmentation with prompting. In *IJCAI*, pages 4252–4258.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. 2020. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195. IEEE.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, page 3982–3992. Association for Computational Linguistics.

Saket Sharma, Aviral Joshi, Namrata Mukhija, Yiyun Zhao, Hanoz Bhathena, Prateek Singh, Sashank Santhanam, and Pritam Biswas. 2022. Systematic review of effect of data augmentation using paraphrasing on named entity recognition. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

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.

Jeniya Tabassum, Wei Xu, and Alan Ritter. 2020. WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 260–267, Online. Association for Computational Linguistics.

Amina Oussaleh Taoufik and Abdellah Azmani. 2024. Ai-enhanced techniques for extracting structured data from unstructured public procurement documents. In *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pages 1–8. IEEE.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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

Arthur Elwing Torres, Edleno Silva de Moura, Altigran Soares da Silva, Mario A Nascimento, and Filipe Mesquita. 2024. An experimental study on data augmentation techniques for named entity recognition on low-resource domains. *arXiv preprint arXiv:2411.14551*.

Alicia Tsai, Shereen Oraby, Vittorio Perera, Jiun-Yu Kao, Yuheng Du, Anjali Narayan-Chen, Tagyoung Chung, and Dilek Hakkani-Tur. 2021. Style control for schema-guided natural language generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 228–242, Online. Association for Computational Linguistics.

MA Xiaoqin, GUO Xiaohe, XUE Yufeng, YANG Lin, and CHEN Yuanzhe. 2021. Data augmentation technology for named entity recognition. *Journal of East China Normal University (Natural Science)*, 2021(5):14.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Qian Yili and Xu Haonan. 2023. Datg: data augmentation with transformer-based generation for low-resource named entity recognition. In *2023 China Automation Congress (CAC)*, pages 6188–6193. IEEE.