Domain-Specific Summarization to Enhance LLM Effectiveness

Hamza Emin Hacıoğlu[1,*], İsmail Karakaya[1], Ensar Erdoğan[1], Serdar Kalaycı[1], Müge Ak[1], Özgür Umut Vurgun[1], Arif Furkan Mendi[1], Mehmet Akif Nacar[1]

Abstract

Large Language Models (LLMs) have become increasingly prevalent in the field of natural language processing, particularly for text summarization tasks. Their ability to capture semantic relationships and generate coherent summaries has made them a powerful alternative to traditional extractive and abstractive summarization methods. However, the effective summarization of long and domain-specific texts, such as petitions, remains a challenging problem. This study explores the problem of petition text summarization by applying state-of-the-art LLMs and evaluating various training strategies on a custom-built dataset. The paper investigates different model training approaches, including Low-Rank Adaptation (LoRA) and prompt engineering techniques, to optimize summarization quality. Human-in-the-loop summarization methods are also incorporated to refine the models, ensuring high-quality outputs in real-world scenarios. Furthermore, the model's performance is assessed through expert evaluation, which highlights areas for potential improvement in terms of both accuracy and coherence. The research further delves into methods for effectively summarizing long-form petition texts, exploring ways to preserve critical information while maintaining brevity and readability. The results demonstrate the feasibility of leveraging LLM-based summarization methods for complex, lengthy texts while providing insights into the challenges and opportunities within this evolving field.

Keywords: LoRA, fine-tuning, SFT, context length, petition text, prompt engineering.

Received: 12.11.2025 Accepted: 18.05.2026 Research Article

[1] HAVELSAN Inc., Ankara, Türkiye
*Corresponding author.
✉ hamzaeminh@havelsan.com.tr

Cite this article as:
https://doi.org/10.65834/jdsi.12.32
Hacioglu, H. E., Karakaya, I., Erdogan, E., Kalayci, S., Ak, M., Vurgun, O. U., Mendi, A. F., & Nacar, M. A. (2026). Domain-Specific Summarization to Enhance LLM Effectiveness. *Journal of Defence and Security Industries: Strategy and Technology 2*, 129-156.

1. Introduction

Text summarization refers to the task of condensing long texts while preserving key information, and significant progress has been made with the emergence of deep learning and transformer-based models. Traditionally, summarization techniques have been divided into two main approaches: extractive and abstractive summarization. Extractive methods select and extract important sentences directly from the source text, while abstractive methods aim to generate new sentences that capture the essence of the input. Although both approaches have proven effective in various domains, challenges remain when dealing with long, complex, and domain-specific texts that require specialized understanding. Recent advances in artificial neural network architectures, particularly transformer-based models, have led to substantial improvements in summarization tasks. Models such as BERT, GPT, and their derivatives excel at capturing long-range dependencies and contextual nuances in language. However, despite these advancements, challenges persist in summarizing long texts, especially legal documents such as petitions. These challenges stem from the need to effectively integrate both local (sentence-level) and global (document-level) context. To address these issues, researchers have developed hybrid approaches that combine extractive and abstractive methods, providing more robust solutions for handling longer and more complex documents. Despite progress in the field of text summarization, domain-specific characteristics of the documents to be summarized still present significant challenges. This is particularly true for petition texts, which are often written in diverse styles and legal language by different individuals. Traditional summarization approaches often fail to effectively summarize such texts. Petition documents tend to be lengthy, lack a standardized structure, and may contain unclear or vague claims from the author. Consequently, automatically generating summaries through artificial intelligence remains a difficult task. Traditional summarization methods, whether extractive or abstractive, struggle to capture the subtle nuances of such documents. Challenges are especially prominent in maintaining the integrity of the text, balancing realism with detail, and ensuring brevity without losing essential information. This study explores innovative techniques for summarizing petition texts that combine local and global contexts, incorporate discourse-aware structures, and implement fine-tuning strategies for domain-specific performance. The goal of this research is to bridge the gap between current summarization techniques and the specific requirements for summarizing petition texts, ensuring that the generated summaries are not only concise but also coherent, relevant, and contextually rich. By focusing on the efficient processing of long documents, this work demonstrates how the combination of extractive and abstractive methods, along with domain adaptation, can significantly enhance the quality and usability of petition text summarization.

Research on text summarization has undergone a significant evolution, progressing from traditional extractive and statistical methods to advanced neural and transformer-based approaches. Early models focused on sentence salience and topic distribution to identify the most representative information within a document. With the rise of deep learning, sequence-to-sequence architectures and attention mechanisms enabled more coherent and contextually aware abstractive summaries. Subsequent developments in transformer-based models

introduced improved contextual understanding, enabling hybrid frameworks that combine extractive precision with abstractive fluency. More recent efforts have concentrated on the challenges of long document summarization, exploring efficient attention mechanisms, hierarchical encoding strategies, and multi-stage frameworks that balance local and global contextual information. These advances have substantially improved the ability of models to handle extended inputs such as reports, dialogues, and multi-section documents. Parallel to advancements in summarization architecture, fine-tuning strategies for LLMs have emerged as a key enabler of task specialization and efficiency. Parameter-efficient adaptation methods, such as low-rank approximation and quantization, have been shown to significantly reduce computational overhead while maintaining or improving summarization performance. The integration of discourse-aware structures and task-specific adaptation techniques has enhanced coherence and factual consistency, particularly in long-form and domain-specific summarization tasks. Furthermore, the combination of supervised, unsupervised, and in-context learning approaches has allowed models to better generalize across diverse datasets while maintaining adaptability to specialized domains. Together, these developments reflect a broader shift toward efficient, scalable, and contextually rich summarization systems grounded in large language model fine-tuning.

Text summarization methods for long documents aim to preserve both local sentence-level information and global document-level coherence. A study shows that this challenge can be addressed by integrating local context, derived from sentence proximity, with global context representing document structure and themes (Xiao & Carenini, 2019). By jointly modelling these two levels of information, their extractive approach improves sentence selection and produces summaries that better reflect the central ideas of long documents while maintaining coherence across sections. Another line of work focuses on improving the efficiency of attention mechanisms in Transformer-based models for long inputs. Another study shows that standard full attention becomes impractical for long sequences due to quadratic computational and memory costs (Huang et al., 2021). To address this limitation, they propose HEPOS, a head-wise stride-based attention mechanism that distributes token coverage across attention heads. This design reduces attention complexity while preserving both local and global dependencies, enabling efficient processing of long documents with competitive summarization performance. Multi-stage summarization frameworks further mitigate context-length limitations by combining extractive and abstractive strategies. For example, SummN, which segments long inputs into manageable chunks, applies extractive filtering to identify salient content, and progressively refines summaries through staged compression and abstractive rewriting (Zhang et al., 2022). This approach allows coherent summarization of long, information-dense documents without exceeding model input limits. Earlier extractive approaches rely on structural importance rather than generation. For instance, LexRank represents sentences as nodes in a similarity graph and selects salient sentences based on graph centrality (Erkan & Radev, 2004). While effective for identifying important content, such graph-based methods may produce disjointed summaries and can be computationally expensive for very long documents.

Topic modelling methods, such as latent Dirichlet allocation, support summarization by identifying dominant themes within documents (Blei et al., 2003). Although these methods provide strong thematic coverage, they often overlook fine-grained details and local discourse structure, leading to overly coarse summaries. Neural sequence-to-sequence models with attention have significantly advanced abstractive summarization. Also, there are pointer-generator networks that combine text generation with copying mechanisms, improving coherence and factual consistency (See et al., 2017). However, these models are computationally intensive and require substantial training data. Sentence embedding–based extractive methods select salient sentences by comparing contextual sentence representations in a continuous semantic space. This approach fine-tunes BERT for extractive summarization, enabling improved semantic matching beyond surface-level similarity, albeit at increased computational cost (Liu, 2019). Finally, hybrid approaches seek to balance factual accuracy and linguistic fluency by combining extractive and abstractive techniques. Furthermore, a study proposes a multi-task framework that jointly optimizes both objectives, improving readability while preserving content coverage, at the cost of increased system complexity (Chen et al., 2019).

Recent work on fine-tuning LLMs has increasingly emphasized achieving strong task performance with reduced computational cost, especially for summarization. A dual-stage framework for news summarization combines in-context learning and parameter-efficient fine-tuning (Guan et al., 2024). In this approach, the model is first adapted to the document-specific context using in-context learning (ELearn), which can be particularly useful when annotated examples are limited (Guan et al., 2024). A second stage then applies an efficient fine-tuning step (EFit) to update only a small subset of parameters, aiming to retain general capabilities while improving task-specific summarization quality under constrained resources (Guan et al., 2024). Overall, this design targets a practical balance: fast adaptation to the current input distribution through prompting, followed by lightweight learning that stabilizes and improves output quality. Large-scale evidence for the effectiveness of parameter-efficient adaptation is provided by an extensive evaluation of 310 LoRA–fine-tuned LLMs across diverse benchmarks (Zhao et al., 2024). The results highlight how LoRA can achieve strong performance without updating full model weights, significantly reducing training compute and time compared with full fine-tuning (Zhao et al., 2024). This work suggests that LoRA supports scalable specialization, allowing models to be adapted to tasks or domains while preserving broadly useful language capabilities, which is attractive for organizations that need multiple specialized models but cannot afford repeated full fine-tuning. Beyond efficiency, recent research has explored incorporating higher-level structure into fine-tuning for long-document summarization. A discourse-aware approach integrates Rhetorical Structure Theory (RST) into LoRA-based adaptation to better capture hierarchical structure and rhetorical dependencies in long texts, improving coherence and reducing omissions or inconsistencies in abstractive summaries (Liu & Demberg, 2024). This line of work reflects a broader trend toward adapting LLMs not only to domain vocabulary but also to document organization and discourse-level constraints that matter for summary quality. Another practical constraint is deploying fine-tuned models under quantization. A method called IR-QLoRA mitigates the accuracy

degradation commonly observed when LoRA-based fine-tuning is performed under ultra-low-bit quantization (e.g., 2–3 bits), by combining Information Calibration Quantization (ICQ) and Information Elastic Connection (IEC) to preserve and recover information more effectively (Qin et al., 2024). Reported results on LLaMA-family models indicate measurable gains compared with prior approaches, while adding minimal overhead, which supports the feasibility of deploying quantified, fine-tuned LLMs in resource-constrained settings (Qin et al., 2024). Fine-tuning strategies have also been examined more systematically in the context of report summarization. A distinction is drawn between Knowledge Fine-Tuning (KFT), which adapts models to domain-specific language using specialized corpora (e.g., government or legal text), and Format Fine-Tuning (FFT), which optimizes task behaviour using paired inputs and reference summaries to reinforce output structure and content selection patterns (Rallapalli et al., 2025). Further discussion covers unsupervised and semi-supervised methods, such as contrastive learning, self-training, and hybrid training using labelled and unlabelled data, as practical alternatives when high-quality labelled summaries are limited (Rallapalli et al., 2025). Together, these studies motivate fine-tuning pipelines that combine efficient adaptation, structural awareness, and training-data pragmatism to make LLM summarization both effective and deployable in real-world scenarios.

2. Methodology

The methodology outlined in this study establishes a structured approach for training a summarization model tailored to Turkish public procurement petitions. The dataset used in this work is confidential and accessible only within institutional authorization limits. The workflow integrates dataset preparation, supervised fine-tuning with controlled prompt variation, efficient parameter adaptation using LoRA, and systematic strategies for handling documents that exceed the model's effective context window. To ensure that the model produces summaries aligned with official Public Procurement Authority writing standards, a two-stage inference procedure is used to separate argumentative content extraction from linguistic normalization. Throughout the process, human expert evaluation was incorporated to assess factual completeness, structural fidelity, and stylistic conformity, enabling iterative refinement of both the model and the prompting strategy.

2.1. Data Source

The dataset comprises 5215 petition text-summary pairs written in Turkish. The average length of the petition texts is 8449 characters, while the average length of the summaries is 3251 characters. The distribution of text lengths is illustrated in Figure 1. In addition, the most frequent words, excluding conjunctions, are presented in Figure 2.

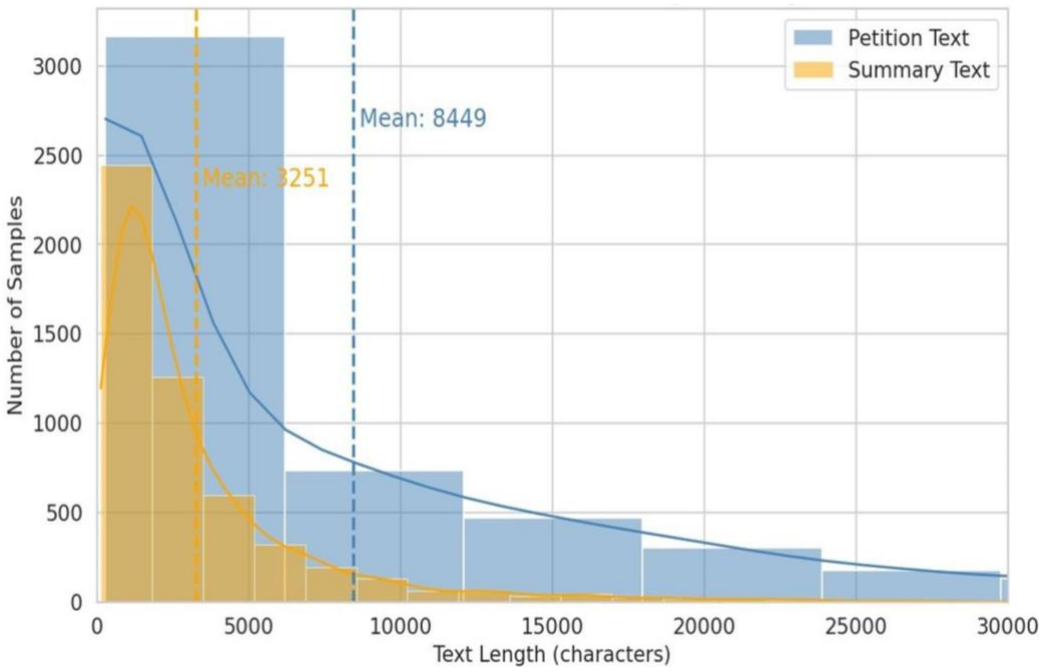

Figure 1. Distribution of petition vs. summary text lengths.

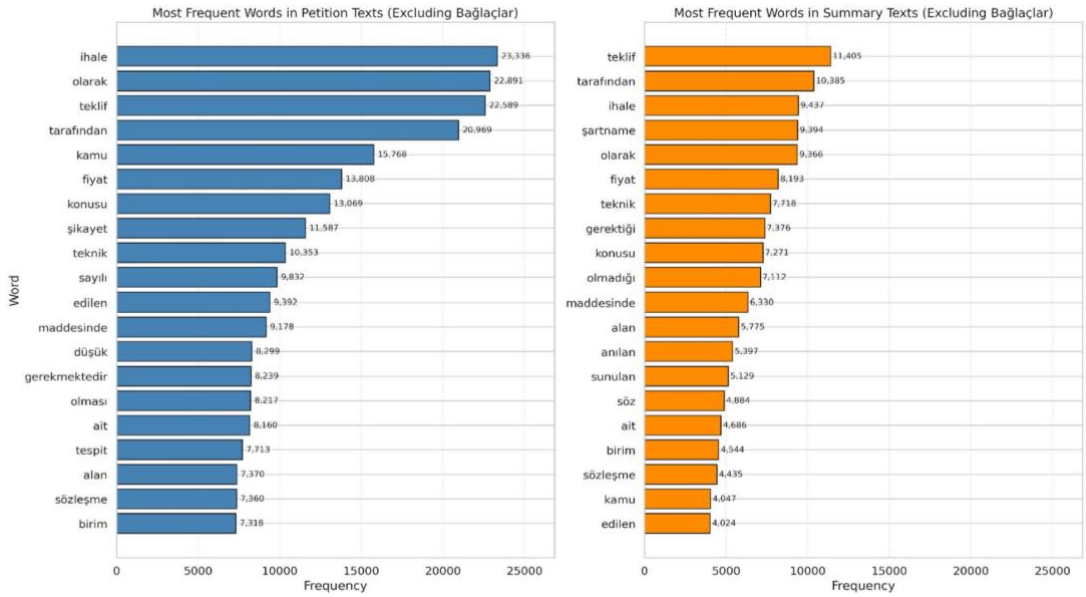

Figure 2. Most frequent words in the dataset, excluding conjunctions ("bağlaçlar" in Turkish language).

2.1.1. Characteristics of Petition Texts and Summaries

Petitions are primarily submitted by companies whose tender applications have been rejected, often written from a first-person perspective. These texts are marked by linguistic and structural inconsistencies, including misspellings, syntactic errors, inverted sentence constructions, and unrelated narrative digressions. Many petitions lack clear claim numbering, with some repeating the same argument or combining multiple issues into one claim. These

inconsistencies introduce variability in text quality, making automatic summarization challenging. In contrast, the summaries are formally consistent, well-structured, and written in clear, standardized Turkish language. Claims are numbered and separated into distinct statements, even when the original petition merges them. Repeated claims in the petition are consolidated in the summary. Written in the third person, the summaries use passive constructions and often end with phrases such as "… yapıldığı," "… edildiği," or "… iddialarına yer verilmiştir". Company names are mentioned once and referred to generically (e.g., "istekli," "birinci istekli," "ikinci istekli"), and the Public Procurement Authority (Kamu İhale Kurumu) is consistently referred to as "idare".

2.1.2. Dataset Selection and Filtering

To prepare the dataset for fine-tuning, several filtering and selection steps were applied. An initial examination of petition lengths showed that approximately 4000 petitions were shorter than 10 pages, 500 petitions ranged between 10 and 20 pages, and around 300 petitions exceeded 20 pages.

- **Claim Numbering:** Some petitions lacked explicit numbering of claims and instead presented arguments in long, unstructured paragraphs. Since claim numbering provides structural cues that facilitate alignment with summaries, petitions without numbering were deprioritized.
- **Length Harmony:** To ensure proportionality between source and target texts, the summary-to-petition length ratio was examined. Pairs with ratios between 33% and 47% were retained, reflecting realistic summarization behaviour and avoiding extreme length imbalances.

After filtering, the final fine-tuning dataset consisted of 2314 numbered petitions and 500 unnumbered petitions. A limited set of unnumbered samples was intentionally included to expose the model to both structured and unstructured examples, improving generalization. The remaining petitions were held out for evaluation on unseen data.

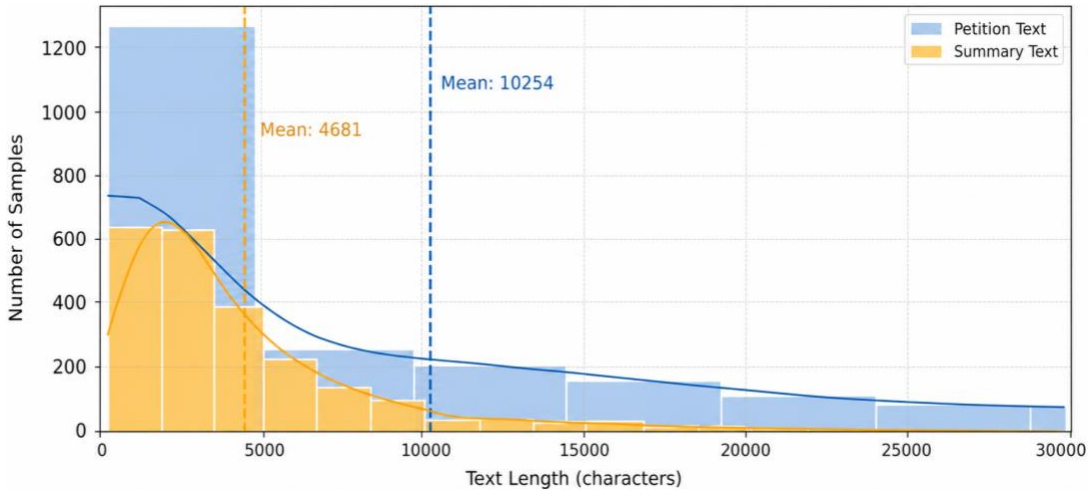

Figure 3. Distribution of petition vs. summary text lengths in the dataset after filtering.

2.2. Fine-Tuning Process

2.2.1. Supervised Fine-Tuning (SFT) Setup

The model was fine-tuned using a supervised, instruction-based approach inspired by instruction following training practices in prior work (Ouyang et al., 2022). Each training instance paired a petition text with its corresponding human-written summary, and the objective was to train the model to produce structured, numbered summaries in passive voice, using formal Turkish legal language while preserving factual content and numerical accuracy.

Early experiments used a single fixed instruction prompt for all samples, which increased prompt overfitting: the model tended to reproduce prompt phrasing rather than internalize the intended summarization behaviour. To reduce this effect, multiple semantically equivalent instruction variants were introduced and randomly assigned across the training set, encouraging the model to generalize the task specification instead of relying on surface-level prompt patterns.

Several prompt-design configurations were evaluated before finalizing the training setup. Early trials using long, inference-style prompts resulted in unstable behaviour, including verbatim text reproduction and overly short or structurally weak summaries. The final configuration employed ten syntactically distinct, but semantically aligned instruction prompts, each conveying the same objective: generating numbered, passive-voice Turkish summaries while preserving dates, percentages, and other factual details. These prompts were rotated across samples to reduce prompt dependency and improve robustness.

This prompt diversification strategy improved training stability and stylistic consistency, allowing the fine-tuned model to learn task-level summarization behaviour, while fine-grained stylistic control was deferred to inference-time system prompts.

2.2.2. Parameter Settings and LoRA Configuration

Fine-tuning was performed using LoRA applied selectively to the feed-forward layers of the LLaMA-3.3-70B-Instruct model (Hu et al., 2021). LoRA enables parameter-efficient adaptation by inserting trainable low-rank matrices into the model while keeping many pretrained parameters frozen, which substantially reduces memory usage and training compute compared with full fine-tuning (Hu et al., 2021).

The experiments were conducted using the LLaMA-3.3-70B-Instruct model as the base architecture, trained with mixed bfloat16 precision on two NVIDIA H200 GPUs providing a total of 288 GB memory. The optimizer was 8-bit Paged AdamW with hyperparameters β_1 = 0.9, β_2 = 0.999 and $\epsilon = 1 \times 10^{-8}$, and the learning rate was set to 1×10^{-6} with a cosine scheduler and a warm-up ratio of 0.3%. The batch size per device was 1 with a gradient accumulation step of 4, while gradient checkpointing was enabled to reduce memory usage. Weight decay was fixed at 0.0001, and LoRA was applied using a LoRA rank $r = 16$, LoRA α of 32, and a dropout rate of 0.05. Training was performed for a single epoch, and no evaluation strategy

was integrated into the training loop; instead, evaluation was conducted separately after fine-tuning.

Regarding numerical precision, the model was trained using bfloat16 (bf16) rather than standard float16 (fp16) precision. Although both formats halve memory consumption compared to full precision, bf16 offers a substantially wider dynamic range while retaining the same 16-bit storage footprint, which helps prevent gradient underflow and instability in large-model training, particularly under very low learning rates and long sequence lengths. Since NVIDIA H200 GPUs natively support bf16 operations, this choice enabled numerically stable optimization and removed the need for loss-scaling techniques typically required by fp16. Overall, this configuration achieved a favourable balance between stability and efficiency: mixed precision, gradient accumulation, and gradient checkpointing together allowed reliable fine-tuning of long petition texts without exceeding the available GPU memory.

2.2.3. Layer Selection for LoRA Adaptation

Initial experiments applied LoRA to all transformer layers, including attention and feed-forward components. While this increased adaptation capacity, it led to overfitting: the model began echoing long input segments instead of producing compact summaries. This indicated that modifying attention projections caused it to over-focus on surface lexical details rather than abstract structure. LoRA was therefore restricted to the feed-forward (FFN) layers (gate, up, and down projections). The attention mechanisms already modelled contextual relations effectively; the summarization task instead required refinement in semantic abstraction and reformulation. Adapting the FFN layers strengthened the model's ability to compress, paraphrase, and structurally organize content while preserving factual accuracy. This selective tuning reduced overfitting and improved generalization: summaries became more concise, consistently numbered, and aligned with legal-administrative style. In short, FFN-only LoRA enhanced semantic reasoning and controlled linguistic realization, achieving a balanced trade-off between precision and fluent summarization.

2.2.4. Model Capacity and Token Limit Testing

Comprehensive testing of model input capacity indicated a pronounced decline in performance once petition texts exceeded approximately 16,000 tokens. While longer sequences were technically accepted by the tokenizer and inference pipeline, the model's effective attention span deteriorated, leading to partial omission of later content or a breakdown in logical coherence across sections. On the output side, generation stability was maintained up to about 4.000 tokens. Beyond this point, completions exhibited structural drift, repetition, or premature truncation, suggesting a saturation of the model's active context window. To mitigate these limitations, the inference process was adapted into a hierarchical structure. Long petitions were first segmented into semantically self-contained sections, each treated as an independent summarization unit. The resulting partial summaries were then recursively merged through a secondary synthesis stage, ensuring that both local detail and global continuity were preserved.

This multi-pass summarization strategy enabled complete document coverage without exceeding the model's operational token and attention constraints.

2.2.5. Base Model Comparison

A comparative evaluation was conducted prior to fine-tuning to identify the most suitable base architecture. The evaluated candidates included GPT-OSS-20B (OpenAI, 2025), LLaMA-3.1-70B-Instruct and LLaMA-3.3-70B-Instruct (Meta AI, 2024), Aya-32B-Expanse (Dang et al., 2024), and Qwen3 variants (Qwen3-235B-A22B, Qwen3-30B-A3B, and Qwen3-32B) (Yang et al., 2025). All models were tested under identical conditions using the same petition sample and the same summarization prompt. Native Turkish-speaking legal experts evaluated the outputs in terms of factual completeness, clarity of argumentation, and structural coherence. During the evaluation, it was observed that model size strongly influenced generation length. Smaller models such as GPT-OSS-20B, Aya-32B-Expanse, Qwen3-30B, and Qwen3-32B consistently produced summaries that were disproportionately short relative to the source texts. In contrast, LLaMA-3.1-70B-Instruct, LLaMA-3.3-70B-Instruct, and Qwen3-235B-A22B generated summaries with sufficiently proportional length and preserved the overall argumentative structure of the petitions. Among these candidates, Qwen3-235B-A22B demonstrated the highest factual accuracy and generation stability; however, its computational requirements exceeded the available infrastructure. The LLaMA models provided comparably strong outputs, though both LLaMA-3.1-70B-Instruct and LLaMA-3.3-70B-Instruct exhibited a recurring issue: the insertion of foreign-language loanwords—particularly toward the end of the summaries—despite explicit Turkish-only prompting. Considering overall output quality, linguistic controllability, and resource feasibility, LLaMA-3.3-70B-Instruct was selected as the base model for fine-tuning. A quantitative comparison between the selected base model and its domain-adapted fine-tuned variant is presented in Table 1 below.

Table 1. Quantitative comparison between the base pretrained model (LLaMA-3.3-70B-Instruct) and the domain-adapted fine-tuned variant under identical inference and decoding settings on a held-out evaluation set of administrative petitions.

Metric	Base Model Mean	Base Std	Fine-Tuned Model Mean	Fine-Tuned Std	Δ (Fine-Tuned Minus Base)	Relative Improvement (%)
BERT Precision	0.6672	0.3758	0.7010	0.0903	+0.0338	+5.07%
BERT Recall	0.6719	0.0899	0.6974	0.0955	+0.0255	+3.79%
BERT F1	0.6688	0.0834	0.6976	0.0898	+0.0288	+4.31%
ROUGE-1 F1	0.5394	0.1380	0.5779	0.1386	+0.0385	+7.14%
ROUGE-2 F1	0.3629	0.1390	0.3991	0.1404	+0.0362	+9.97%
ROUGE-L F1	0.3758	0.1422	0.4090	0.1363	+0.0332	+8.84%

Note: M = mean; SD = standard deviation. Relative improvement is computed as (Mfine-tuned−Mbase)/Mbase×100. Positive Δ values indicate improvement.

2.3. Handling Long Petition Texts

One of the primary challenges encountered during this study was the summarization of long petitions, particularly those exceeding eight to ten pages. Beyond this length, the models were unable to generate outputs longer than approximately 4000 tokens, regardless of the input size or prompt configuration. This inherent limitation prevented the models from capturing all claims and contextual details within a single generation, often leading to incomplete or overly compressed summaries. To overcome this constraint, a degree of human intervention was initially required to manually divide the petitions into smaller, processable segments. Building upon this necessity, several automated strategies were subsequently developed and tested to minimize manual effort while maintaining summary completeness and coherence. The following subsections present these methods in detail, outlining their underlying logic, implementation steps, and empirical outcomes.

2.3.1. Fixed Length Segmentation and Merging Attempts

The initial approach for handling long petitions used fixed-length segmentation, where each document was divided into consecutive parts based on a character or word threshold. Each segment was summarized independently to ensure full coverage within the model's context window. In a subsequent step, the model was prompted to merge the partial summaries into a single coherent output, as shown in Algorithm 1. However, instead of aggregating the content, the model consistently treated the merge instruction as a new summarization task, producing an output comparable in length to a single partial summary and discarding information from earlier segments. For instance, a multi-page petition split into three segments yielded three detailed partial summaries, but the merged output retained only a fraction of the total claim content. While the segmentation stage itself was reliable, the merging process failed to preserve proportional detail and claim coverage. Consequently, this strategy was deemed unsuitable for long-document summarization.

Algorithm 1: Fixed-Length Segmentation and Merging Procedure

Require : petition_text, segment_size

Step-1 : segments ← SPLITBYLENGTH(petition_text, segment_size)

Step-2 : partial_summaries ← []

Step-3 : for seg ∈ segments do

Step-4 : s ← LLM_SUMMARIZE(seg)

Step-5 : APPEND(partial_summaries, s)

Step 6 : end for

Step 7 : merged_summary ← LLM_MERGESUMMARIES(partial_summaries)

Step 8 : return merged_summary

2.3.2. Context-Preserving Sequential Summarization

To improve coherence across long documents, a sequential summarization strategy was applied at Algorithm 2 in which each segment was summarized together with contextual information from the previous segment. Specifically, the model received both the final portion of the preceding text and its generated summary, allowing it to maintain continuity of arguments and narrative flow. This process was repeated iteratively for all segments, resulting in summaries that were more cohesive and less fragmented than those produced by isolated segmentation. However, as contextual information accumulated across iterations, GPU memory usage increased substantially, frequently leading to Out-of-Memory (OOM) errors for long petitions. As a result, despite improved coherence, this approach was not operationally stable or scalable for large-scale summarization.

Algorithm 2: Context-Preserving Sequential Summarization

Require : petition_text, segment_size, context_overlap = 500

Step-1 : segments ← SPLITBYLENGTH(petition_text, segment_size)

Step-2 : summaries ← []

Step-3 : prev_summary ← “”

Step-4 : prev_tail ← “”

Step-5 : for i ← 1 to |segments| do

Step 6 : current ← segments[i]

Step 7 : if i > 1 then

Step 8 : context ← CONCAT(prev_tail, prev_summary, current)

Step 9 : else

Step 10 : context ← current

Step 11 : end if

Step 12 : summary ← LLM_SUMMARIZE(context)

Step 13 : APPEND(summaries, summary)

Step 14 : prev_summary ← summary

Step 15 : prev_tail ← GETTAIL(current, context_overlap)

Step 16 : end for

Step 17 : final_summary ← LLM_MERGE(summaries)

Step 18 : return final_summary

2.3.3. Character-Position Claim Segmentation

This experiment attempted to segment petitions based on the approximate character positions of individual claims rather than using fixed-length boundaries. The primary objective was to identify where one claim ended and the next began within the raw text, thereby allowing the model to generate precise and semantically meaningful split indices. To achieve this, a specialized prompt was designed instructing the model to locate the approximate start and end positions of each distinct claim within the petition. The model was expected to return outputs in the following structured form:

- **Claim 1** → characters 136–1121;
- **Claim 2** → characters 1121–2530;
- **Claim 3** → characters 2530–6390.

Using these character ranges, each claim section was extracted from the original petition text and summarized independently. The individual summaries were then combined sequentially to form a complete, claim-by-claim summary of the petition. For instance, in one petition concerning a contract award, the model produced the following segmentation:

- **Claim 1:** Objection to the rejection of the bidder's offer (characters 220–1410).
- **Claim 2:** Argument that the awarded bidder's proposal was abnormally low (characters 1410–2890).
- **Claim 3:** Assertion that evaluation criteria were not applied consistently (characters 2890–4870).

Although this approach was conceptually sound, the practical results were highly inconsistent. Different models, and even multiple runs of the same model, returned divergent character boundaries for identical petitions. In many cases, the identified indices did not align with actual claim boundaries or fell within the middle of a sentence, leading to fragmented and incoherent sections. This instability made it impossible to reproduce the segmentation reliably. Consequently, despite its structural elegance, as illustrated in Algorithm 3, the character-

position method was deemed unsuitable for systematic summarization, as it failed to provide consistent or interpretable split locations across models and runs.

Algorithm 3: Character-Position Based Claim Segmentation and Summarization

Require : petition_text

Step-1 : boundaries ← LLM_DETECTCLAIMBOUNDARIES(petition_text)

Step-2 : summaries ← []

Step-3 : for all (start, end) in boundaries do

Step-4 : claim_text ← petition_text[start : end]

Step-5 : claim_summary ← LLM_SUMMARIZECLAIM(claim_text)

Step 6 : APPEND(summaries, claim_summary)

Step 7 : end for

Step 8 : final_summary ← LLM_MERGE(summaries)

Step 9 : return final_summary

2.3.4. Main Subclaim Decomposition

This experiment investigated a structured approach in which the model was instructed to identify the internal argumentative hierarchy of each petition by extracting main claims and their associated subclaims in a single pass. Each main claim, together with its supporting subclaims, was treated as a coherent semantic unit for subsequent summarization, and unit-level summaries were later merged in the original document order. For instance, in one petition the model produced the following hierarchical outline:

- **Main Claim 1:** Objection to the tender decision due to the rejection of the bidder's offer.
    - **Subclaim 1.1:** The bidder's proposal should have been considered the most economically advantageous offer.
    - **Subclaim 1.2:** The awarded company submitted an abnormally low bid that should have been disqualified.
    - **Subclaim 1.3:** The tender commission applied the evaluation criteria inconsistently with the relevant procurement regulation.
- **Main Claim 2:** The tender specifications were not applied uniformly to all participants.
    - **Subclaim 2.1:** Certain bidders were permitted to correct documentation errors, while others were not.

While this approach produced summaries with a clear logical structure, it exhibited significant variability in the number and granularity of extracted claims across petitions and runs. In some cases, redundant subclaims inflated summary length, while in others overly coarse claims omitted procedural details. This instability limited controllability, and the approach was therefore deemed unsuitable for large-scale or production use despite its conceptual advantages, exhibited in Algorithm 4.

Algorithm 4: Main- Subclaim Decomposition and Per-Unit Summarization

Require : petition_text

Step-1 : structure ← LLM_EXTRACTHIERARCHY(petition_text)

Step-2 : summaries ← []

Step-3 : for all (mc, sc_list) in structure do

Step-4 : claim_unit ← BUILDUNIT(petition_text, mc, sc_list)

Step-5 : unit_summary ← LLM_SUMMARIZEUNIT(claim_unit)

Step 6 : APPEND(summaries, unit_summary)

Step 7 : end for

Step 8 : final_summary ← LLM_MERGE(summaries)

Step 9 : return final_summary

2.3.5. Semantic Chunking – Final and Stable Approach

The final and most successful method was semantic chunking, which divided petitions into meaning-based sections rather than fixed lengths or indexed positions. The goal was to preserve the logical flow and internal consistency of each claim group while keeping the model's input size within its processing limit. In this method, the text was first analysed as a whole, and then divided into semantically coherent sections (e.g., Claim Group-1, Claim Group-2, . . .). Each section was summarized independently, and the resulting partial summaries were merged sequentially to form the complete output. This approach yielded the most consistent and stable results for several reasons:

- Claim integrity and logical order were preserved,
- Information loss was minimized across parts,
- The process remained within the model's effective limits (approximately 16,000 tokens input, and 4000 tokens output),
- No OOM issues were observed.

Extensive internal testing confirmed that semantic chunking produced the most balanced and reliable long summaries. Consequently, this approach was adopted in Algorithm 5 as the standard procedure for handling petitions beyond the model's natural context window.

Algorithm 5: Semantic Chunking and LLM-Based Summarization

Require: petition_text, max_words = 5000, flex_window = 1000

Step-1 : words ← TOKENIZETOWORDS(petition_text)

Step-2 : i ← 0

Step-3 : chunks ← []

Step-4 : while i < |words| do

Step-5 : base_end ← min(i + max_words, |words|)

Step 6 : win_start ← max(0, base_end - flex_window)

Step 7 : win_end ← min(|words|, base_end + flex_window)

Step 8 : candidate ← JOIN(words[win_start : win_end])

Step 9 : split_info ← LLM_SUGGESTSPLIT(candidate)

Step 10 : local_split ← MAPSENTENCESTOWORDINDEX(candidate, split_info)

Step 11 : split_idx ← win_start + local_split

Step 12 : chunk ← JOIN(words[i : split_idx])

Step 13 : APPEND(chunks, chunk)

Step 14 : i ← split_idx

Step 15 : end while

Step 16 : summaries ← []

Step 17 : for all c in chunks do

Step 18 : s ← LLM_SUMMARIZE(c)

Step 19 : APPEND(summaries, s)

Step 20 : end for

Step 21 : final_summary ← LLM_MERGE(summaries)

Step 22 : return final_summary

2.3.6. Summary of Findings

Overall, the experiments demonstrated that the model's summarization capability does not scale proportionally with input length. Beyond a certain threshold, longer inputs consistently produced compressed or incomplete summaries, regardless of model size or configuration. Fixed-length segmentation generated usable partial summaries but failed during the merging stage, where the model re-summarized rather than concatenated content. The context-preserving sequential method improved coherence but suffered from rapid memory accumulation and OOM errors. Structural decomposition techniques, such as character-position segmentation and main subclaim extraction, introduced interpretability but proved unstable due to inconsistent boundary detection and variable claim granularity. In contrast, the semantic chunking approach addressed these limitations collectively by aligning split points with semantic boundaries instead of fixed lengths or character indices. It produced coherent, detailed, and reproducible summaries across petitions while remaining computationally stable. As a result, semantic chunking was adopted as the standard strategy for handling long petition texts in subsequent stages of the project. A comparison of chunking and summarization strategies is shown in Table 2 below.

Table 2. Comparison of chunking and summarization strategies evaluated for petition summarization.

Method	Core Idea and Implementation	Observed Issues	Outcome / Stability
Fixed-Length Segmentation and Merging	Petitions were split into equal-length parts processed independently. Each part was summarized separately, and the partial summaries were merged into one complete output.	The model summarized instead of concatenating, producing shorter merged summaries and losing detail.	Conceptually simple but inconsistent; discarded.
Context-Preserving Sequential Summarization	Each segment was summarized using the previous part's final 500 characters and its summary to maintain continuity.	Improved coherence, but cumulative context caused high GPU memory usage and frequent OOM errors.	Conceptually strong but computationally unstable.
Character-Position Claim Segmentation	The model was asked to identify the start and end character indices of each claim, and these ranges were summarized separately.	Character boundaries varied across models and runs, producing misaligned and fragmented segments.	Structurally elegant but non-reproducible.

Table 2. (Continued)

Main–Subclaim Decomposition	The entire petition was analysed in one pass to extract hierarchical pairs of main claims and subclaims. Each main subclaim unit was summarized independently and merged.	The number and depth of extracted claims varied; over- and under segmentation caused inconsistent lengths and detail levels.	Provided structural insight but unstable for large-scale use.
Semantic Chunking (Final Approach)	Petitions were divided into meaning-based chunks within model limits (approx. 16k input / 4k output tokens). Each chunk was summarized independently and merged sequentially.	No significant computational issues; minimal information loss.	Most balanced and reliable; adopted as standard procedure.

Note: OOM = out-of-memory error.

2.4. Inference-Time Prompt Design

This section describes the development and refinement of the inference-time prompt used for summarizing long petitions. While the fine-tuning stage focused on learning the general summarization behaviour, inference required a standardized instruction that balanced factual completeness, legal tone, and stylistic consistency across multi-part summaries. The process involved approximately 30 prompt candidates, iteratively refined through evaluation with large-scale models such as ChatGPT, Claude, and internal deployments of LLaMA-3.3-70B-Instruct. The final prompt served as the foundation for all experiments presented in Section 4.

2.4.1. Inference Objectives

At inference time, the model was required to generate outputs that conformed to formal Turkish legal administrative style while preserving the full factual and legal content of the petitions. Despite task-aligned fine-tuning, output quality remained sensitive to instruction phrasing for long, multi-claim inputs. Accordingly, the inference objectives were defined along five axes:

- **Factual and Legal Completeness:** Every factual claim, legislative citation, specification article, date, percentage, monetary value, annex reference (e.g., Ek-1, Ek-2), and procedural justification had to be retained. No evidentiary or contextual element was to be omitted or paraphrased to the point of losing meaning.
- **Structural Uniformity:** The summary had to begin with the exact phrase “Başvuru sahibinin dilekçesinde özetle;” and end with a single formal closing line. Claims were to be presented as a single continuous numbered list (1), 2), 3), … without restarting, with each claim occupying exactly one paragraph.

- **Passive-Voice Enforcement:** The entire text was required to use passive constructions consistent with official administrative style. Each numbered paragraph ended with a neutral passive clause (e.g., "… uygun olmadığı," "… geçersiz olduğu," "… tespit edildiği," "… değerlendirme dışı bırakılması gerektiği"). Endings such as "iddia edilmektedir" or "verilmektedir" were disallowed within the claim paragraphs.
- **Linguistic and Stylistic Fidelity:** Outputs were expected to reflect the tone and rhythm of official Public Procurement Authority summaries: formal, objective, and free of colloquial or advocative phrasing. Long, multi-clause passive sentences were acceptable so long as supporting facts were integrated clearly and consistently.
- **Context Preservation Within Claims:** When a claim included justifications, evidence, examples, or consequences, these components were to be included within the same numbered paragraph to maintain causal and evidentiary coherence without fragmenting the argument.

These objectives guided the final inference instructions toward a balance of completeness, structural reliability, and stylistic consistency, enabling reproducible summaries across petitions of varying length and complexity. Across approximately thirty iterations, prompt design experiments explored different stylistic and structural formulations. The main design dimensions included:

- **Instruction Framing:** Whether the model was instructed to "summarize" vs. "rewrite intro claims".
- **Output Structuring:** Numbered vs. unnumbered, single sentence vs. multi-sentence per claim.
- **Stylistic Control:** Passive voice enforcement, tone formality, or explicit mention of Turkish language.
- **Length Regulation:** Inclusion or omission of token, paragraph, or claim count limits.
- **Factual Anchoring:** Explicit emphasis on preserving percentages, dates, and references to legislation.
- **Each Prompt Variant Was Tested on a Representative Subset of Long Petitions.** Native Turkish legal experts qualitatively assessed the outputs for factual accuracy, linguistic formality, and consistency with the official summary style.

2.4.2. Final Prompt and Rationale

The final inference process employed a two-stage prompting strategy designed to preserve factual completeness while enforcing grammatical and stylistic consistency in Turkish legal-administrative language. This approach was introduced after observing that large models often struggled to maintain fully passive and nominalized sentence endings (-dığı/-diği/-duğu/-düğü) when summarizing long legal petitions in a single pass. To address this, inference was separated into a content-generation stage followed by a linguistic normalization stage.

2.4.2.1. Stage 1 – Argumentative Summarization

In the first step, the model was instructed to act as a legal summarization specialist, producing a detailed, argumentative summary of the petition. The goal was to extract all factual claims, legal citations, evidence, and quantitative details, with each numbered paragraph representing a coherent argumentative unit. This stage reliably preserved evidential reasoning and claim structure but required subsequent grammatical standardization to align with passive-voice norms.

2.4.2.2. Stage 2 – Linguistic Normalization

The second step reformatted the Stage 1 output into a legally consistent linguistic form. Here, the model acted as a legal text formatter, transforming the preliminary summary without altering its content. Each claim was rewritten as a single Turkish sentence ending with an appropriate nominalized passive form (-dığı/-diği/-duğu/-düğü), while varied endings (e.g., belirtildiği, tespit edildiği, ifade edildiği, öne sürüldüğü) were encouraged for naturalness and stylistic authenticity. This stage also enforced fixed structural conventions: the standardized opening "Başvuru sahibinin dilekçesinde özetle;", numbered one-sentence claims, and the closing line "İddia edilmekte ve itirazen şikâyet başvurusunun Kurumumuzca incelenerek gereğinin yapılması istenmektedir." In operational terms, the pipeline treats Stage 1 as a high-recall extraction step and Stage 2 as a constrained rewriting step with strict invariants. Specifically, Stage 2 is instructed to preserve the original ordering, claim count, and all factual entities (dates, quantities, legal references, and identifiers) while only regularizing syntax and surface form. To minimize error propagation, Stage 2 explicitly prohibits adding, removing, merging, or splitting claims, and it enforces one sentence-per-claim formatting with a fixed opening and closing line. Normalization makes the method robust to long-document effects: even when the input is processed in multiple semantic chunks (Algorithm 5), the normalization step yields a uniform legal style across chunks without sacrificing evidentiary content.

2.4.2.3. Rationale

The two-step procedure was essential for achieving both semantic fidelity and linguistic precision. Stage 1 ensured complete preservation of claims and evidentiary relationships, while Stage 2 guaranteed syntactic regularity and conformity to institutional writing norms. Separating reasoning from formatting mitigated the model's tendency toward mixed verb forms and yielded summaries that were consistent, grammatically controlled, and legally admissible. This two-stage prompt was applied uniformly to all semantically segmented chunks (Algorithm 5), minimizing stylistic drift across the final output.

2.4.2.4. Illustrative Example

The effectiveness of the approach is demonstrated by comparing the raw argumentative output of Stage 1 with the linguistically normalized version from Stage 2. Firm and personal identifiers have been anonymized for publication.

- **Stage 1 Output — Content-Focused Argumentative Summarization (Anonymized):** *Başvuru sahibinin dilekçesinde özetle;*

- İdari Şartnamenin 7.1(h) maddesine göre isteklilerin teklif kapsamında Gıda Üretim İzni Belgesi veya İşletme Kayıt Belgesi sunmaları gerektiği, ancak Firma A ve Firma B tarafından sunulan belgelerin faaliyet konularının "Hazır Yemek ve Tabldot Yemek Üretimi" kapsamını karşılamadığı ve belgelerde yer alan adres beyanlarının gerçeği yansıtmadığı, bu nedenle söz konusu isteklilerin teklif dosyalarının değerlendirme dışı bırakılması gerektiği ileri sürülmektedir.
- Firma B'nin işletme kayıt belgesinin Tarım ve Orman Bakanlığının resmi kayıtları ile uyumlu olmadığı, adres bilgilerinin gerçeği yansıtmadığı ve belgenin geçersiz olduğu, bu nedenle ilgili isteklinin teklifinin değerlendirme dışı bırakılması gerektiği belirtilmektedir.
- İdari Şartnamenin 7.4.1 maddesi uyarınca sunulması gereken banka referans mektuplarının, Firma A ve Firma B tarafından ilan tarihinden önce alındığı ve kredi tutarlarının şartnameye uygun olmadığı değerlendirilmektedir.
- İdari Şartnamenin 7.5.2 maddesi uyarınca sunulması gereken 1.000 kişilik yemek üretim kapasitesine ilişkin kapasite raporlarının, Firma A ve Firma B tarafından geçerli şekilde sunulmadığı ve adres beyanlarının diğer ihale belgeleri ile örtüşmediği ifade edilmektedir.
- Sonuç olarak, ihale sürecinde belirtilen aykırılıklar doğrultusunda düzeltici işlem yapılması gerektiği iddia edilmektedir.

- **Stage 2 Output — Passive-Voice Legal Normalization (Anonymized):** *Başvuru sahibinin dilekçesinde özetle;*
  - İdari Şartnamenin 7.1(h) maddesi gereği sunulması gereken Gıda Üretim İzni Belgesi veya İşletme Kayıt Belgesi kapsamında, Firma A ve Firma B tarafından sunulan belgelerde "Hazır Yemek ve Tabldot Yemek Üretimi" faaliyet konusunun yer almadığı ve adres beyanlarının gerçeği yansıtmadığı belirtilerek, ilgili isteklilerin teklif dosyalarının değerlendirme dışı bırakılması gerektiği belirtildiği,
  - Firma B tarafından sunulan işletme kayıt belgesinin resmi kayıtlarla uyumlu olmadığı ve belgenin geçersiz olduğu tespit edildiği,
  - İdari Şartnamenin 7.4.1 maddesi uyarınca sunulması gereken banka referans mektuplarının Firma A ve Firma B tarafından ilan tarihinden önce alındığı ve kredi tutarlarının şartname koşullarını karşılamadığı belirtildiği,
  - İdari Şartnamenin 7.5.2 maddesi uyarınca sunulması gereken 1000 kişilik yemek üretim kapasitesine ilişkin kapasite raporlarının Firma A ve Firma B tarafından geçersiz sunulduğu ve adres bilgilerinin ihale belgeleri ile örtüşmediği vurgulandığı,
  - İddia edilmekte ve itirazen şikâyet başvurusunun Kurumumuzca incelenerek gereğinin yapılması istenmektedir.

2.4.3. Empirical Evaluation of the Two-Stage Inference Design

To justify the proposed two-stage inference strategy, a comparative evaluation was conducted against a single-pass summarization baseline on 25 administrative petitions selected to capture structural and linguistic variability. Both approaches were applied under identical decoding settings and evaluated in collaboration with domain experts. Quantitative results are reported in Table 3, comparing single-pass (Stage 1) and two-stage (Stage 2) outputs. The two-stage approach consistently outperforms the baseline across all automatic metrics, including BERT Score and ROUGE-1/2/L (F1), indicating improved semantic alignment with reference summaries while preserving lexical and structural consistency. Error-type analysis revealed that single-pass summaries frequently exhibited non-compliant sentence endings and occasional non-Turkish lexical forms. These issues were consistently corrected in the second stage through linguistic normalization enforcing standardized passive constructions and Turkish lexical usage. Finally, blind pairwise human evaluations by domain experts showed a clear preference for the two-stage summaries in all 25 cases, citing higher legal accuracy, greater compactness, and closer alignment with expert-written administrative summaries. Overall, the results in Table 3 provide empirical justification for the proposed two-stage inference design.

Table 3. Quantitative comparison between single-pass (Stage 1) and two-stage (Stage 2) summarization.

Metric	Stage 1 Mean	Stage 1 Std	Stage 2 Mean	Stage 2 Std
BERT Precision	0.6672	0.0822	0.7010	0.0903
BERT Recall	0.6719	0.0899	0.6974	0.0955
BERT F1	0.6688	0.0834	0.6976	0.0898
ROUGE-1 F1	0.5349	0.1380	0.5779	0.1386
ROUGE-2 F1	0.3629	0.1390	0.3991	0.1404
ROUGE-L F1	0.3758	0.1422	0.4090	0.1363

Note: Stage 1 corresponds to single-pass summarization, while Stage 2 corresponds to the proposed two-stage summarization pipeline. Values are reported as mean and standard deviation across the evaluation set.

2.4.4. Common Failure Patterns and Lessons Learned

During prompt evaluation, several recurring failure types were observed:

- **Under-Summarization:** Some prompts caused the model to compress multi-claim petitions into overly brief outputs.
- **Over-Generation:** Others led to redundant restatements or the introduction of fabricated subclaims.

- **Inconsistent Formatting:** Minor template variations occasionally resulted in missing numbering or mixed sentence structures.
- **Style Drift:** Without strict phrasing, some models reverted to active-voice constructions or informal tone.

Through iterative refinement and domain-expert review, these issues were systematically reduced. The final inference prompt reflects an empirically optimized balance between content fidelity, structural consistency, and linguistic control. While earlier configurations produced acceptable results for some petitions, achieving stable consistency across all document types required multiple cycles of prompt adjustment and evaluation.

2.5. Human-in-the-Loop Evaluation

In addition to automated consistency checks, a human-in-the-loop evaluation process was integrated to ensure legal and linguistic adequacy of the generated summaries. A panel of procurement law specialists and linguistic reviewers examined approximately ten petition outputs produced under different inference configurations. These sessions provided continuous expert feedback on the model's handling of sentence endings, Turkish grammatical correctness, claim separation, and the overall conformity of tone to formal administrative language. The experts' feedback was systematically incorporated into prompt and model refinements across several iterations. Attention was given to achieving natural variation in passive endings, accurate reflection of legal claims, and balanced sentence structure without loss of factual completeness. Through this cycle, both the model's inference behaviour and the prompt design evolved toward higher linguistic precision and legal reliability. Ultimately, this collaborative process yielded a configuration that consistently met expert expectations for clarity, structural consistency, and adherence to institutional writing standards. It also established a feedback-driven evaluation loop that can be reused for future fine-tuning and validation cycles.

2.5.1. Data and Code Availability

The dataset utilized in this study consists of confidential petition texts that are not publicly available due to institutional data protection requirements. These documents contain legally sensitive information and are subject to internal confidentiality policies, which prohibit public release or unrestricted distribution. For this reason, the fine-tuned model, training corpus, preprocessing scripts, and intermediate training artifacts derived from the dataset cannot be shared externally.

2.5.2. Ethical, Privacy and Data Governance Considerations

Given the sensitive nature of administrative and legal petitions, careful ethical and privacy-aware data handling practices were applied throughout this study. All documents were obtained through authorized institutional channels and were used exclusively for research purposes. Full anonymization was not applied prior to modelling, as preserving the identities of the parties involved (e.g., contracting authority and bidder sides) was necessary to correctly interpret roles, claims, and contextual relationships required for accurate legal summarization. However, the

data was not modified, redistributed, or exposed beyond the authorized research environment. Access to the dataset was strictly limited to the research team, and all processing was conducted in secured, access-restricted systems. The study relies on retrospective administrative documents and does not involve human subjects or intervention; therefore, formal ethics committee approval was not required under applicable institutional regulations.

2.5.3. Main Contributions

The main contributions of this work can be summarized as follows:

- A domain-specific summarization framework for Turkish public procurement petitions. A complete, end-to-end summarization pipeline is presented, tailored to the structural, linguistic, and legal characteristics of administrative petition texts, addressing challenges such as inconsistent claim structure, long document length, and strict institutional writing requirements.
- A two-stage inference strategy separating legal reasoning from linguistic normalization. A two-stage inference design is introduced and empirically justified in which argumentative content extraction is decoupled from stylistic and grammatical normalization, enabling both factual completeness and strict adherence to Turkish legal-administrative passive-voice conventions.
- A systematic evaluation of long-document handling strategies for petition summarization. A detailed empirical analysis of multiple long-text processing approaches is conducted, including fixed segmentation, sequential context propagation, claim-based decomposition, and semantic chunking, and semantic chunking combined with staged inference is identified as the most stable and scalable solution.
- Domain-informed fine-tuning with controlled prompt diversification. It is demonstrated that supervised fine-tuning with LoRA, combined with prompt diversification rather than fixed instruction templates, improves generalization and reduces prompt overfitting in domain-specific summarization tasks.
- Expert-driven evaluation and error analysis beyond automatic metrics. A human-centred evaluation framework is provided involving domain experts, including quantified comparison, error-type analysis, and blind preference studies, showing that the proposed approach consistently outperforms single-stage summarization in legal admissibility, linguistic accuracy, and stylistic conformity.

3. Conclusion

This study presented a complete, domain-focused summarization framework for legal petition texts, combining dataset preparation, model training, algorithmic design, and expert-driven evaluation. A detailed analysis of petition and summary structures was first conducted, examining length distributions, variability in linguistic clarity, and the role of numbering and legal reference patterns. Based on these observations, a dataset was curated and filtered that

preserved realistic claim organization while reducing noise and structural ambiguity. On the model development side, supervised fine-tuning (SFT) was employed on LLaMA-3.3-70B-Instruct using parameter-efficient LoRA adaptation, supported by carefully diversified system prompts to prevent memorized prompting behaviour and ensure stylistic generalization. In parallel, multiple strategies were developed and tested for handling long petition texts, including fixed segmentation, sequential summarization with continuity signals, hierarchy-based claim decomposition, and character-position boundary heuristics. Through iterative evaluation, semantic chunking combined with a two-stage inference prompt emerged as the most stable approach, providing coherent claim separation, consistent passive-voice legal style, and improved factual retention without exceeding the model's operational context limits. Finally, legal experts reviewed model outputs to validate correctness, tone, and structural alignment with institutional summarization standards. The resulting system demonstrates that domain adaptation, prompt engineering, and structured long document processing can significantly improve the reliability and legal validity of summarization models in real-world procurement contexts. Future work will focus on extending the system beyond supervised summarization toward a preference-aligned, context-stable model. Future work will aim to improve long-context generation within a single inference window through attention-extension and continuity-aware training, reducing reliance on chunking. In addition, Reinforcement Learning from Human Feedback using Proximal Policy Optimization will be applied with a summarization-specific reward model targeting claim completeness, factual accuracy, and legal-administrative passive voice style. Finally, user-configurable summarization profiles will be explored to control compression level, evidential detail, and linguistic formality across institutional contexts.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This research was supported by HAVELSAN / Artificial Intelligence Group Leadership under internal R&D support. No external funding was received.

Acknowledgements

The authors would like to express their sincere gratitude to Dilek BİLİR for her expert review, critical evaluation, and invaluable guidance throughout the study. They also wish to acknowledge Burcu ŞIKLAR and İrem BURGAÇ PEKSAYIN for their detailed feedback and domain-specific assessments, which greatly contributed to enhancing the quality of the summarization output. Finally, the authors extend their appreciation to Necmettin YÜKSEL, Head of the Electronic Procurement Department, and Serdar HALICI, Deputy President of the Institution, for their continued support and constructive contributions during the research and implementation process.

Ethical Statements

The study involved analysis of petition-style documents and expert evaluation of anonymized model outputs. Any personal or institutional identifiers were removed or masked prior to experimentation and reporting, and results are presented only in aggregated or anonymized form. No human subject experimentation was conducted.

Data and Code Availability

Due to confidentiality constraints associated with domain-specific petition data and institutional usage, the dataset and trained model artifacts cannot be publicly released. An anonymized sample and/or evaluation templates may be made available from the corresponding author upon reasonable request, subject to approval and data-sharing policies.

Supplementary Materials

Supplementary materials include additional prompt variants, example anonymized outputs for Stage 1 and Stage 2, and extended qualitative evaluation notes. These materials can be provided upon request.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Chen, Y., Ma, Y., Mao, X., & Li, Q. (2019). Multi-task learning for abstractive and extractive summarization. *Data Science and Engineering,* 4(1), 14–23. https://doi.org/10.1007/s41019-019-0087-7

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., ... Hooker, S. (2024). *Aya Expanse: Combining research breakthroughs for a new multilingual frontier.* arXiv. https://doi.org/10.48550/arXiv.2412.04261

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research,* 22, 457–479. https://doi.org/10.1613/jair.1523

Guan, C., Chin, A., & Vahabi, P. (2024). *Enhancing news summarization with eLearnFit through efficient in-context learning and efficient fine-tuning*. arXiv. https://doi.org/10.48550/arXiv.2405.02710

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA of LLMs. arXiv. https://doi.org/10.48550/arXiv.2106.09685

Huang, L., Cao, S., Parulian, N., Ji, H., & Wang, L. (2021). Efficient attentions for long document summarization. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1419–1436). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.112

Liu, D., & Demberg, V. (2024). RST-LoRA: *A discourse-aware LoRA for long document abstractive summarization*. arXiv. https://doi.org/10.48550/arXiv.2405.00657

Liu, Y. (2019). *Fine-tune BERT for extractive summarization.* arXiv. https://doi.org/10.48550/arXiv.1903.10318

Meta AI. (2024). *The LlaMA 3 herd of models.* arXiv. https://doi.org/10.48550/arXiv.2407.21783

OpenAI. (2025). *gpt-oss-120b & gpt-oss-20b model card.* arXiv. https://doi.org/10.48550/arXiv.2508.10925

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://doi.org/10.48550/arXiv.2203.02155

Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., Luo, J., Liu, X., & Magno, M. (2024). *Accurate LoRA-finetuning quantization of LLMs via information retention*. arXiv. https://doi.org/10.48550/arXiv.2402.05445

Rallapalli, S., Gallagher, S., Mellinger, A. O., Ratchford, J., Sinha, A., Brooks, T., & Brown, B. (2025). *Fine-tuning LLMs for report summarization: Analysis on supervised and unsupervised data*. arXiv. https://doi.org/10.48550/arXiv.2503.10676

See, A., Liu, P. J., & Manning, C. D. (2017). *Get to the point: Summarization with pointer-generator networks*. arXiv. https://doi.org/10.48550/arXiv.1704.04368

Xiao, W., & Carenini, G. (2019). Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3011-3021). https://doi.org/10.18653/v1/D19-1298

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., ... Qiu, Z. (2025). *Qwen3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2505.09388

Zhang, Y., Ni, A., Mao, Z., Wu, C.-H., Zhu, C., Deb, B., & Zhang, R. (2022). SummN: A multi-stage summarization framework for long input dialogues and documents. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1592–1604). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.112

Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., & Rishi, D. (2024). *LoRA Land: 310 fine-tuned LLMs that rival GPT-4, a technical report*. arXiv. https://doi.org/10.48550/arXiv.2405.00732