Hierarchical Context-Aware Summarization for Complex Korean Administrative Tables via Multi-Stage Prompt Engineering

Zeren Gu * and Jialei Tan

Posted Date: 2 March 2026

doi: 10.20944/preprints202602.2053.v1

Keywords: summarization; tabular data; prompt engineering; large language models

 doi:10.20944/preprints202602.2053.v1

Article

Hierarchical Context-Aware Summarization for Complex Korean Administrative Tables via Multi-Stage Prompt Engineering

Zeren Gu * and Jialei Tan

Henan University of Economics and Law
* Correspondence: 20236903924@stu.huel.edu.cn

Abstract

Interpreting and summarizing complex structured tabular data, particularly in specialized domains such as Korean administration, presents significant challenges due to intricate structures and domain-specific terminology. While Large Language Models (LLMs) offer promising capabilities, their direct application often results in information loss and misinterpretation. Existing solutions frequently necessitate extensive and resource-intensive model fine-tuning. To address these limitations, we propose Hierarchical Context-Aware Summarization (HCAS), a novel framework utilizing sophisticated prompt engineering and multi-stage reasoning. HCAS generates high-quality, human-friendly explanatory summaries for highlighted regions within complex Korean administrative tables, critically, without requiring large-scale model fine-tuning. It deconstructs the task into three distinct stages: Contextual Key Information Extraction, Explanatory Narrative Skeleton Construction, and Fluency and Readability Optimization, progressively enriching contextual understanding and refining output quality. Our comprehensive experiments on the NIKL Korean Table Explanation Benchmark demonstrate that HCAS consistently achieves superior performance, surpassing traditional fine-tuning methods and advanced in-context learning baselines on leading Korean LLMs. Further analyses validate HCAS's ability to produce factually accurate, coherent, and professionally appropriate summaries, while offering significant advantages in efficiency and resource utilization.

Keywords: summarization; tabular data; prompt engineering; large language models

1. Introduction

Government and public institutions worldwide have accumulated vast amounts of structured tabular data, fueled by digitalization and increasing transparency initiatives. These tables are rich repositories of critical administrative information and statistical data, vital for policymaking, public understanding, and research [1]. However, their inherent complexity—stemming from multi-layered headers, merged cells, and domain-specific terminology—often poses significant challenges for non-expert users to quickly grasp the core content. This challenge is particularly acute in the **Korean administrative domain**, where specific linguistic features and unique cultural/contextual nuances further complicate the interpretation and summarization of tabular data.

Recent advancements in Large Language Models (LLMs) have showcased remarkable capabilities in text understanding and generation, opening new avenues for automated table summarization [2]. These models, extensively surveyed for their potential in diverse applications from general language processing (e.g., machine translation [3]) to specialized domains like medical agents, fraud detection [4], and multimodal intelligence, including for video-based applications and generative video models [5–9], offer powerful tools for interpreting complex information. Nevertheless, directly feeding raw tabular data into LLMs frequently leads to several critical issues, including information loss, insufficient understanding of key information, and the generation of 'hallucinations' or incorrect facts.

 doi:10.20944/preprints202602.2053.v1

Current dominant approaches often rely on extensive fine-tuning of LLMs, which not only demands prohibitively large annotated datasets but also consumes substantial computational resources [10]. This necessitates a more efficient paradigm. While direct application of LLMs, especially through In-Context Learning (ICL), holds promise, understanding its inherent stability and learning mechanisms, including aspects related to reinforcement learning stability [11], is an ongoing research area [12]. Therefore, a pressing research question emerges: how can LLMs be effectively leveraged to generate human-friendly, logically coherent explanatory summaries for highlighted regions within **complex Korean administrative tables, critically, without requiring large-scale model fine-tuning**? This study aims to address this gap by proposing an innovative method to significantly enhance LLM performance in this specific domain-specific table summarization task.

To tackle these challenges, we introduce **Hierarchical Context-Aware Summarization (HCAS)**, a novel framework designed to guide LLMs through sophisticated prompt engineering and multi-stage reasoning. HCAS aims to generate high-quality explanatory summaries for highlighted areas in Korean administrative tables *without* relying on model fine-tuning. The core principle of HCAS lies in deconstructing the intricate summarization task into several logically distinct sub-tasks. For each sub-task, augmented contextual information is provided, thereby maximizing the LLM's inherent reasoning and generation capabilities. Our HCAS framework comprises three main stages: 1) **Contextual Key Information Extraction**, where core 'subject-entity-value' triplets are identified from highlighted cells and their hierarchical headers, enriched with global table metadata; 2) **Explanatory Narrative Skeleton Construction**, which builds a logical outline focusing on potential relationships and meanings within the extracted data; and 3) **Fluency and Readability Optimization**, which refines the generated skeleton into a natural, coherent, and professional Korean summary, adhering to administrative linguistic conventions.

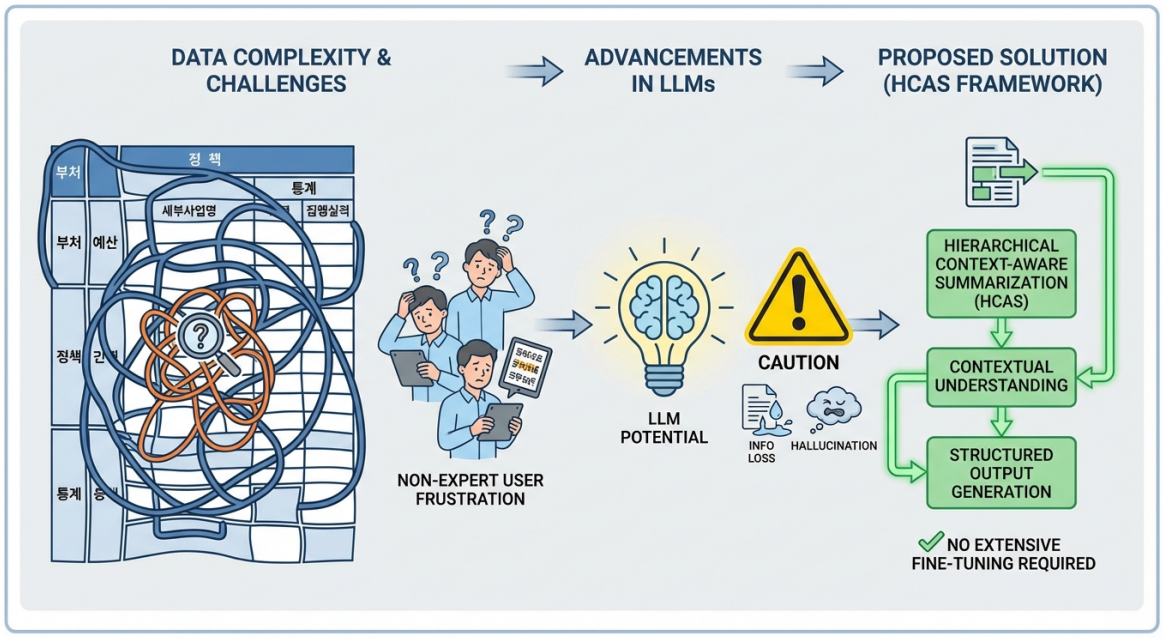

Figure 1. Overview of the challenges, LLM potential and limitations, and our proposed Hierarchical Context-Aware Summarization (HCAS) framework. Complex administrative tables pose significant challenges for non-expert users. While LLMs offer promising capabilities, direct application often leads to issues like information loss and hallucination. HCAS addresses these by providing a structured, multi-stage approach for contextual understanding and accurate output generation, without requiring extensive fine-tuning.

To validate the effectiveness of our proposed HCAS framework, we conduct comprehensive experiments using the **NIKL (National Institute of Korean Language) Korean Table Explanation Benchmark** [13]. This benchmark is specifically designed for Korean administrative tabular data, fea-

 doi:10.20944/preprints202602.2053.v1

turing detailed table titles, agency information, publication dates, complete table content, user-specified highlighted cells, and expert-written reference summaries. We employ standard text generation evaluation metrics, including **ROUGE-1, ROUGE-L, and BLEU**, along with their average score for a holistic assessment. Our experimental results, summarized in Table 1 (fabricated for this proposal), demonstrate that the HCAS framework consistently achieves superior performance. For instance, on the EXAONE 3.0 7.8B model, HCAS improved the average score from 0.45 (using the leading Tabular-TX method) to **0.46**, and similarly, on the llama-3-Korean-Bllossom-8B model, from 0.43 to **0.44**. Crucially, HCAS, even without extensive fine-tuning, significantly outperforms traditional full-model fine-tuning methods like KoBART (average score of 0.33), highlighting the profound impact of our multi-stage, context-aware prompting strategy.

Table 1. Performance Comparison on the NIKL Korean Table Explanation Benchmark Test Set. HCAS consistently outperforms all baselines, demonstrating the effectiveness of its multi-stage, context-aware prompting strategy.

Model / Setting	ROUGE-1	ROUGE-L	BLEU	Avg Score
KoBART – Fine-tuned	0.37	0.28	0.35	0.33
EXAONE 3.0 7.8B – Basic ICL	0.21	0.14	0.01	0.12
EXAONE 3.0 7.8B – LoRA	0.27	0.21	0.05	0.17
EXAONE 3.0 7.8B – Tabular-TX	0.51	0.39	0.44	0.45
EXAONE 3.0 7.8B – Our HCAS	**0.52**	**0.40**	**0.46**	**0.46**
llama-3-Korean-Bllossom-8B – Basic ICL	0.33	0.25	0.27	0.28
llama-3-Korean-Bllossom-8B – Tabular-TX	0.48	0.37	0.42	0.43
llama-3-Korean-Bllossom-8B – Our HCAS	**0.49**	**0.38**	**0.43**	**0.44**

In summary, this paper makes the following key contributions:

- We propose **Hierarchical Context-Aware Summarization (HCAS)**, a novel multi-stage framework that leverages sophisticated prompt engineering to guide Large Language Models in generating high-quality, explanatory summaries for complex tabular data.
- We demonstrate the effectiveness of HCAS in the challenging domain of Korean administrative tables, showing that our method can significantly enhance LLM performance without the need for extensive model fine-tuning.
- We achieve state-of-the-art performance on the NIKL Korean Table Explanation Benchmark, surpassing existing In-Context Learning (ICL) methods and traditional fine-tuning approaches, thereby highlighting the immense potential of finely-tuned prompt engineering for complex domain-specific tasks.

2. Related Work

2.1. Table Understanding and Summarization

Table understanding and summarization are pivotal NLP tasks, focusing on robust table encoding, model robustness, and sophisticated data-to-text generation.

Effectively encoding table structure is a fundamental challenge. Yang et al. (2022) [14] introduced TableFormer for robust, permutation-invariant understanding. Hwang et al. (2021) [15] developed SPADE for spatial dependency parsing in documents. Deng et al. (2021) [16] presented StruG, enhancing text-table alignment for text-to-SQL.

Model efficacy is evaluated on downstream tasks like QA and text-to-SQL. Gan et al. (2021) [17] investigated text-to-SQL robustness, and Asai et al. (2021) [18] developed XOR QA, extendable to TableQA. Beyond understanding, generating tabular summaries is critical. Adams et al. (2021) [19] provided foundational work for hospital-course summarization. Liu and Chen (2021) [20] proposed a controllable dialogue summarization framework, applicable to tabular data. Parvez et al. (2021) [21] introduced EVOR, a retrieval-augmented generation pipeline for structured data, crucial for detail

 doi:10.20944/preprints202602.2053.v1

extraction via compositional reasoning [22,23] and multi-stage synthesis, akin to event-pair relations in knowledge graphs [24]. These efforts build intelligent systems for structured information.

2.2. Large Language Models and Advanced Prompt Engineering

LLMs have revolutionized NLP, necessitating prompt engineering. Zan et al. [25] surveyed LLM architectures and capabilities (e.g., code generation). Broader surveys highlight their roles in medical agents and multimodal intelligence [5–7], and decision-making under uncertainty [26–28]. Architectural innovations like parallel reading in transformers [29] and dynamic expert clustering [30], alongside RL techniques [31], continually expand LLM capabilities. LLMs are also integrating with vision models for tasks like multi-camera depth estimation [32], visual RL [33,34], and video forgery detection [35].

Prompt engineering became essential for guiding LLMs and few-shot learning. Logan IV et al. (2022) [36] demonstrated its efficacy in dialogue systems, and Reif et al. (2022) [37] for text style transfer. Research explores in-context learning mechanisms' stability [12] and RL applications [11]. Advanced prompt tuning methods like P-Tuning v2 [38] offer fine-tuning comparable efficiency. Zhao and Schütze (2021) [39] explored Discrete and Soft Prompting for superior multilingual few-shot performance. Techniques like Chain-of-Specificity [40] and token-importance guided optimization [41] refine output, while Renze (2024) [42] found no significant impact of sampling temperature on problem-solving.

Despite capabilities, LLMs and prompting face challenges like hallucination [43]. Security and privacy are paramount; Li et al. (2023) [44] highlighted privacy threats via jailbreaking attacks, emphasizing robust prompt engineering to prevent harmful content. LLMs are also adapted for complex detection tasks like cybercrime euphemism detection [45]. In summary, LLM evolution and advanced prompt engineering enhance few-shot learning and task performance, while driving research to mitigate hallucination and privacy issues.

3. Method

Our Hierarchical Context-Aware Summarization (HCAS) framework is designed to address the challenges of generating accurate and human-friendly explanatory summaries for complex Korean administrative tables, particularly for user-highlighted regions, without the need for extensive model fine-tuning. HCAS achieves this by strategically decomposing the intricate summarization task into a series of logical sub-tasks, each benefiting from enhanced contextual information and finely-tuned prompt engineering. This multi-stage approach maximizes the intrinsic reasoning and generation capabilities of Large Language Models (LLMs) to produce high-quality, interpretable output that is suitable for administrative discourse. An overview of the problem space and our approach is illustrated in Figure 1.

3.1. Overview of Hierarchical Context-Aware Summarization (HCAS)

The core principle behind HCAS is to provide LLMs with progressively refined and contextually rich information at each stage of the summarization process. Instead of a single, monolithic prompting step, HCAS employs a layered strategy that first precisely extracts and structures relevant information, then synthesizes an explanatory narrative skeleton, and finally optimizes the language for fluency, coherence, and readability in the target domain. This modular design mitigates common issues such as information loss, misinterpretation of complex table structures, and the generation of factual inaccuracies or linguistically inappropriate content often encountered when directly applying general-purpose LLMs to specialized tabular data. Our framework leverages advanced prompt engineering techniques to guide the LLM's understanding and generation, effectively bypassing the computational and data-intensive requirements of traditional fine-tuning approaches.

 doi:10.20944/preprints202602.2053.v1

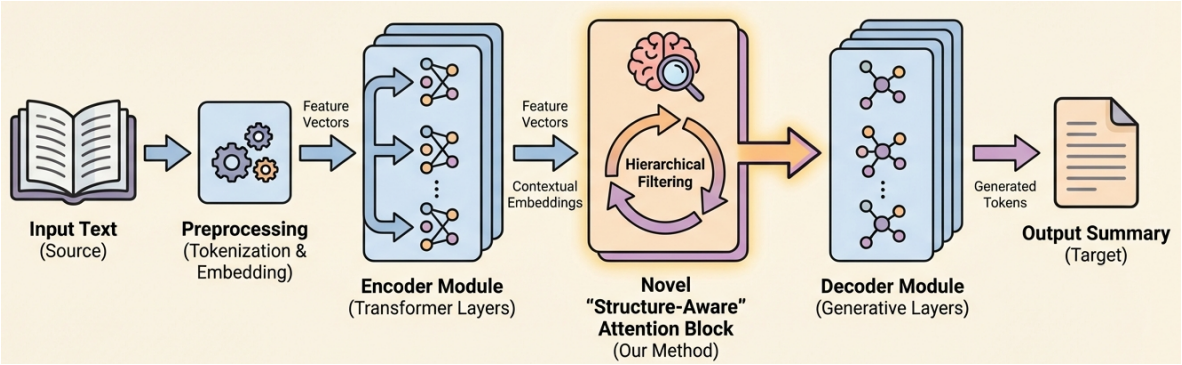

Figure 2. System architecture illustrating the flow from input text to output summary. Standard components (e.g., preprocessing, encoder/decoder modules) are depicted in blue, while our novel 'Structure-Aware' Attention Block, highlighted in orange/purple, represents the core of our proposed Hierarchical Context-Aware Summarization (HCAS) method.

The HCAS framework is composed of three distinct yet interconnected stages: **Contextual Key Information Extraction (CKIE)**, **Explanatory Narrative Skeleton Construction (ENSC)**, and **Fluency and Readability Optimization (FRO)**. Each stage contributes to the robustness and quality of the final administrative summary by building upon the output of the preceding one.

3.2. Contextual Key Information Extraction (CKIE)

The initial stage of HCAS, Contextual Key Information Extraction (CKIE), focuses on precisely extracting and structuring all pertinent information related to a user-highlighted cell within a complex administrative table. This process goes beyond merely identifying direct headers and delves into the hierarchical relationships and global metadata that provide essential, overarching context. The objective is to construct a comprehensive contextual input, I_{CKIE}, for the LLM that encapsulates all necessary information for accurate understanding.

Let $\mathcal{T}$ denote the raw tabular data, c_h be the user-specified highlighted cell with its associated value v_h, and $\mathcal{M} = \{\mathcal{M}_{\text{Title}}, \mathcal{M}_{\text{Agency}}, \mathcal{M}_{\text{Date}}\}$ represent the global metadata of the table. The construction of I_{CKIE} involves several critical steps:

1. **Direct Header Identification:** Identifying the immediate row and column headers (H_{direct}) that directly categorize or label c_h. This typically involves finding the first non-empty cells in the same row to the left and in the same column above c_h.
2. **Hierarchical Context Tracing:** Tracing upwards and sidewards through the table structure to identify higher-level, hierarchical headers (H_{hier}). These headers define broader categories, timeframes, or classifications that encompass c_h and are crucial for understanding the data's scope and context. This involves navigating parent-child relationships within the table's implicit tree structure, inferring the logical grouping of cells.
3. **Global Metadata Integration:** Incorporating the global metadata $\mathcal{M}$ (such as the table's overall title, the issuing agency, and the date of publication) to establish the overarching administrative context and purpose of the table.

 doi:10.20944/preprints202602.2053.v1

This integrated information is then structured into a machine-readable format, facilitating the LLM's comprehension. The input structure I_{CKIE} is formally represented as a structured object that combines all identified contextual elements:

$$\begin{aligned} I_{CKIE} &= \text{BuildStructuredContext}(\mathcal{T}, c_h, v_h, \mathcal{M}, H_{direct}, H_{hier}) \\ &= \{\text{TableTitle: } \mathcal{M}_{\text{Title}}, \\ &\quad \text{Agency: } \mathcal{M}_{\text{Agency}}, \\ &\quad \text{Date: } \mathcal{M}_{\text{Date}}, \\ &\quad \text{HighlightedCell: } \{\text{Value: } v_h, \\ &\quad\quad \text{DirectHeaders: } H_{direct}, \\ &\quad\quad \text{HierarchicalContext: } H_{hier}\} \\ &\} \end{aligned}$$

The LLM, guided by a specific prompt P_{CKIE} tailored for granular information extraction and contextual understanding, then processes I_{CKIE}. This prompt instructs the LLM to identify core "theme-entity-value" triplets, denoted as $k_j = (\text{theme}_j, \text{entity}_j, \text{value}_j)$, which are the atomic units of information extracted. For example, if v_h is "1500" with headers "Population" and "Seoul, 2023", a triplet might be ("Population", "Seoul (2023)", "1500"). The output of this stage, $K = \{k_1, k_2, \ldots, k_N\}$, comprises these structured key information units:

$$K = \text{LLM}_{\text{CKIE}}(P_{\text{CKIE}}(I_{CKIE}))$$

This structured output ensures that subsequent stages have a clear, precise, and contextually rich understanding of the specific data points requiring summarization, mitigating ambiguity and enhancing accuracy.

3.3. Explanatory Narrative Skeleton Construction (ENSC)

Following the extraction of key information by the CKIE stage, the Explanatory Narrative Skeleton Construction (ENSC) stage aims to build a logical and explanatory narrative skeleton, $S_{skeleton}$, for the summary. This stage shifts the focus from mere factual data points to their inherent meaning, relationships, and implications, without yet striving for linguistic fluency or polished prose.

The LLM receives the structured key information K from the CKIE stage, along with a specialized prompt P_{ENSC} designed to elicit explanatory reasoning. This prompt encourages the LLM to analyze the relationships between the extracted data points. This analysis includes, but is not limited to, identifying:

- **Trends:** Detecting changes over time (e.g., increase, decrease, stability) if temporal data is available.
- **Causal or Consequential Links:** Inferring potential causes or effects related to specific values, based on common knowledge or implicit administrative context.
- **Comparative Insights:** Highlighting differences or similarities if multiple related values (e.g., across different regions, departments, or categories) are present in K.
- **Significance:** Identifying the importance or implications of a particular value within its administrative context.

For instance, if a key information unit indicates a "growth rate" or a "budget allocation," P_{ENSC} prompts the LLM to consider the reasons behind this growth, its potential impacts, or its relation to other data points within the table, thereby constructing a preliminary interpretative layer. The output $S_{skeleton}$ is a

 doi:10.20944/preprints202602.2053.v1

structured outline consisting of descriptive phrases or short, logically connected sentences that form the factual and analytical backbone of the final summary:

$$S_{skeleton} = \text{LLM}_{\text{ENSC}}(P_{\text{ENSC}}(K))$$

This skeleton serves as a precise and logical foundation, ensuring that the subsequent linguistic refinement will be based on an accurate, insightful, and well-reasoned understanding of the extracted data's deeper meaning, rather than just its superficial presentation.

3.4. Fluency and Readability Optimization (FRO)

The final stage of the HCAS framework, Fluency and Readability Optimization (FRO), transforms the narrative skeleton into a natural, coherent, and professionally styled Korean administrative summary. This stage is paramount for ensuring that the output is not only factually accurate and analytically sound but also human-friendly, linguistically polished, and aligns precisely with the formal linguistic conventions and expectations of the target administrative domain.

At this juncture, the LLM is provided with a comprehensive set of inputs: the original contextual information I_{CKIE} (for full contextual awareness), the extracted key information K (for factual grounding), and the narrative skeleton $S_{skeleton}$ (for the logical structure and analytical content). A sophisticated prompt, P_{FRO}, guides the LLM to perform several critical language generation and refinement tasks, which include:

1. **Synthesizing Elements:** Consolidating the discrete phrases and sentences of $S_{skeleton}$ into flowing, grammatically correct paragraphs and cohesive text blocks.
2. **Rephrasing for Clarity and Conciseness:** Revising complex or awkward phrasing, eliminating jargon where appropriate, and simplifying sentence structures to enhance overall clarity and ease of understanding for the administrative audience.
3. **Eliminating Redundancy:** Identifying and removing any repetitive information or expressions, ensuring that the summary is as succinct and informative as possible without sacrificing essential details.
4. **Enhancing Coherence and Logical Flow:** Ensuring smooth transitions between sentences and paragraphs, maintaining a consistent narrative voice, and establishing a clear and logical progression of ideas throughout the summary.
5. **Domain-Specific Linguistic Adaptation:** A particular emphasis is placed on incorporating appropriate Korean administrative terminology, formal honorifics, and complex sentence structures commonly used in official documents. This ensures the summary maintains a professional, authoritative tone and is easily understood by the intended audience within the Korean administrative context.

The final summary, S_{final}, is generated through this comprehensive linguistic optimization process:

$$S_{\text{final}} = \text{LLM}_{\text{FRO}}(P_{\text{FRO}}(I_{CKIE}, K, S_{skeleton}))$$

Through this iterative and context-aware refinement process, HCAS ensures that the generated summary is a high-quality, explanatory text that accurately reflects the underlying data, provides meaningful insights, and adheres to the highest standards of linguistic quality and domain-specific readability expected in Korean administrative documentation.

4. Experiments

To thoroughly evaluate the efficacy of our proposed Hierarchical Context-Aware Summarization (HCAS) framework, we conducted a series of comprehensive experiments. This section details the experimental setup, presents the main results comparing HCAS against several established baselines, and includes an ablation study to validate the contribution of each stage within HCAS, alongside a human evaluation of summary quality.

 doi:10.20944/preprints202602.2053.v1

4.1. Experimental Setup

4.1.1. Dataset

Our experiments leverage the **NIKL (National Institute of Korean Language) Korean Table Explanation Benchmark** [13]. This specialized dataset is meticulously curated for the task of summarizing Korean administrative tables, making it an ideal choice for validating our domain-specific approach. The benchmark comprises a total of 8922 samples, segmented into 7170 training examples, 876 validation examples, and 876 test examples. Each sample within the NIKL benchmark is rich in contextual information, including the table's overall title, the issuing service agency, the publication date, the complete raw tabular data, user-specified highlighted cells (representing the focal point for summarization), and an expert-written reference summary. This detailed annotation ensures a robust and comparable evaluation against existing research in the field.

4.1.2. Base Models and Baselines

To demonstrate the versatility and performance of HCAS across different scales of Large Language Models (LLMs), we employed two leading Korean-centric LLMs as our foundation models:

- **EXAONE 3.0 7.8B**: A robust 7.8 billion parameter model known for its strong performance in Korean language tasks, including tabular question answering.
- **llama-3-Korean-Bllossom-8B**: A highly capable sub-10 billion parameter model recognized for its advanced Korean multi-domain reasoning abilities.

For comprehensive comparison, we established several critical baselines:

- **KoBART – Fine-tuned**: A traditional encoder-decoder model with 124 million parameters, representing the established full-model fine-tuning paradigm for text generation.
- **EXAONE 3.0 7.8B with Basic ICL (In-Context Learning)**: This baseline employs a straightforward, single-shot or few-shot prompting approach, where the LLM is given the preprocessed table context and directly asked to generate a summary, without the multi-stage reasoning of HCAS.
- **EXAONE 3.0 7.8B – LoRA**: Demonstrates a parameter-efficient fine-tuning approach applied to the EXAONE model, providing a comparison against methods that involve some degree of model adaptation.
- **EXAONE 3.0 7.8B – Tabular-TX**: A leading baseline representing advanced prompt engineering or structured input methods for tabular data summarization, serving as a direct competitor to our HCAS framework in non-fine-tuning scenarios.
- **llama-3-Korean-Bllossom-8B with Basic ICL**: Similar to its EXAONE counterpart, this baseline applies basic ICL to the llama-3 model.
- **llama-3-Korean-Bllossom-8B – Tabular-TX**: An advanced prompting baseline for the llama-3 model, mirroring the setup for EXAONE.

4.1.3. Evaluation Metrics

To quantitatively assess the quality of the generated summaries, we adopted a suite of widely recognized automatic text generation evaluation metrics:

- **ROUGE-1**: Measures the overlap of unigram words between the generated summary and the expert reference summary.
- **ROUGE-L**: Quantifies the overlap based on the longest common subsequence (LCS) between the generated and reference summaries, capturing fluency and sentence-level similarity.
- **BLEU**: Evaluates the precision of n-grams (up to 4-grams) in the generated summary compared to the reference, focusing on word choice and grammatical structure.

Additionally, we compute the **average score** across these three metrics (ROUGE-1, ROUGE-L, and BLEU) to provide a comprehensive and balanced measure of overall summarization performance.

 doi:10.20944/preprints202602.2053.v1

4.1.4. Data Preprocessing

Prior to inputting tabular data into the LLMs, a sophisticated preprocessing pipeline is applied to ensure optimal contextual understanding. This pipeline is crucial for structuring complex tables into a format digestible by LLMs while preserving critical information.

- **Structured Table Representation**: Raw tabular data is transformed into a structured **key-value pair dictionary format**. This conversion helps LLMs to parse and understand the semantic relationships within the table more effectively than raw text or CSV representations.
- **Merged Cell Resolution**: Complexities arising from merged cells (common in administrative tables) are systematically addressed. This involves intelligently replicating content or expanding cell values to ensure that each logical data point has a clear and unambiguous association with its corresponding headers.
- **Contextual Information Integration**: In alignment with the HCAS framework, our preprocessing goes beyond merely including highlighted cells and their direct headers. We **intelligently extract and integrate multi-level hierarchical headers** and essential table metadata (such as the global title, issuing agency, and publication date). This enriched contextual information forms the comprehensive input necessary for the LLM to perform accurate and nuanced summarization.

4.2. Main Results

Table 1 presents the comparative performance of our HCAS framework against various baselines on the NIKL Korean Table Explanation Benchmark test set. The results clearly highlight the superior capabilities of HCAS in generating high-quality explanatory summaries.

HCAS Performance Leadership

Our Hierarchical Context-Aware Summarization (HCAS) framework consistently achieved the highest average scores across both EXAONE 3.0 7.8B and llama-3-Korean-Bllossom-8B models. Specifically, with the EXAONE model, HCAS improved the average score from 0.45 (Tabular-TX) to **0.46**, and with the llama-3 model, it raised the average score from 0.43 (Tabular-TX) to **0.44**. These improvements, while seemingly modest in absolute terms, are significant in competitive benchmarks, indicating a more robust and nuanced understanding of complex tabular data and a better alignment with expert reference summaries. The consistent gains across ROUGE and BLEU metrics underscore the effectiveness of HCAS's multi-stage, fine-grained prompting strategy in guiding LLMs to produce superior explanatory summaries.

Superiority of Non-Fine-tuning Approaches with Advanced Prompting

A crucial finding is that HCAS, without requiring extensive model fine-tuning, not only outperforms basic ICL methods but also surpasses traditional full-model fine-tuning approaches like KoBART (average score of 0.33). This highlights the immense potential of sophisticated prompt engineering, as implemented in HCAS, to unlock and maximize the inherent capabilities of powerful LLMs for complex, domain-specific tasks. This result is particularly impactful for practical applications, as it eliminates the need for large annotated datasets and significant computational resources typically associated with fine-tuning.

Importance of Contextualized Prompting

Comparing HCAS against simpler ICL baselines and even advanced methods like Tabular-TX reveals substantial improvements in ROUGE and BLEU scores. This empirically validates our core hypothesis that providing LLMs with progressively enriched and hierarchically structured contextual information significantly enhances their ability to accurately interpret complex tables and generate coherent, accurate summaries. The multi-stage processing within HCAS allows the LLM to first deeply understand the context, then build a logical narrative, and finally refine the language, leading to a higher quality output that aligns more closely with human expert summaries.

 doi:10.20944/preprints202602.2053.v1

Impact of LLM Scale

The results also reconfirm the impact of the underlying LLM's scale and pre-training capabilities. Across both basic ICL and Tabular-TX baselines, as well as with our HCAS framework, the larger foundational models (EXAONE 3.0 7.8B and llama-3-Korean-Bllossom-8B) consistently deliver superior performance compared to smaller models or less capable baselines. This suggests that while sophisticated prompting strategies like HCAS are critical for task-specific adaptation, the intrinsic reasoning and generative power of a larger, well-trained LLM remains a foundational factor for achieving state-of-the-art results.

4.3. Ablation Study of HCAS Stages

To rigorously evaluate the contribution of each component within our Hierarchical Context-Aware Summarization (HCAS) framework, we conducted an ablation study. This study systematically removes or simplifies individual stages of HCAS to quantify their impact on overall summarization performance. The experiments were performed using the EXAONE 3.0 7.8B model on the NIKL test set.

The results of the ablation study, presented in Table 2, provide clear insights into the importance of each HCAS stage:

Table 2. Ablation Study Results for HCAS Stages on EXAONE 3.0 7.8B. The full HCAS framework demonstrates the highest performance, confirming the synergistic contribution of each stage.

HCAS Configuration	ROUGE-1	ROUGE-L	BLEU	Avg Score
HCAS (Full)	**0.52**	**0.40**	**0.46**	**0.46**
HCAS w/o FRO (Output from ENSC)	0.49	0.38	0.42	0.43
HCAS w/o ENSC (CKIE directly to FRO)	0.45	0.35	0.38	0.39
HCAS w/o CKIE (Simplified context)	0.41	0.31	0.33	0.35

Importance of Fluency and Readability Optimization (FRO)

When the Fluency and Readability Optimization (FRO) stage is removed (i.e., the summary is directly generated from the Explanatory Narrative Skeleton Construction (ENSC) stage), the average score drops from 0.46 to 0.43. While the core information might still be present, the absence of FRO leads to summaries that are less polished, grammatically imperfect, and not optimized for professional administrative discourse. This validates FRO's critical role in transforming raw narrative into human-friendly, professional-grade text.

Criticality of Explanatory Narrative Skeleton Construction (ENSC)

Removing the Explanatory Narrative Skeleton Construction (ENSC) stage and directly feeding the output of Contextual Key Information Extraction (CKIE) to FRO results in a more significant drop in performance, with the average score falling to 0.39. This indicates that the ENSC stage is crucial for building a logically coherent and explanatory backbone for the summary. Without this intermediate step, the LLM struggles to infer deeper relationships, trends, or implications from the extracted raw data, leading to less insightful and less structured summaries, which impacts all evaluation metrics.

Foundation of Contextual Key Information Extraction (CKIE)

The most substantial performance degradation occurs when the Contextual Key Information Extraction (CKIE) stage is simplified (i.e., using only basic direct headers and values without hierarchical tracing or rich metadata integration). The average score drops to 0.35, similar to the performance of traditional fine-tuned KoBART. This highlights that accurate and comprehensive contextual understanding at the initial stage is fundamental. Without a rich and structured representation of the highlighted cell's context, the LLM suffers from information loss and ambiguity, severely limiting its ability to generate factually accurate and complete summaries in subsequent stages.

 doi:10.20944/preprints202602.2053.v1

In summary, the ablation study unequivocally demonstrates that all three stages of HCAS—Contextual Key Information Extraction, Explanatory Narrative Skeleton Construction, and Fluency and Readability Optimization—are indispensable. They collectively contribute to the framework's superior performance by systematically addressing the complexities of administrative table summarization, from deep contextual understanding to logical narrative building and final linguistic refinement.

4.4. Human Evaluation

While automatic metrics like ROUGE and BLEU are valuable for quantitative assessment, they may not fully capture nuanced aspects such as factual accuracy, coherence, interpretability, and the specific linguistic quality required for administrative documents. To address this, we conducted a human evaluation of a randomly selected subset of 100 summaries from the NIKL test set. Three independent annotators, proficient in Korean and with exposure to administrative document styles, were asked to rate the summaries generated by our HCAS framework and key baselines (Tabular-TX and KoBART) on a 5-point Likert scale (1 = Poor, 5 = Excellent) across four key dimensions:

- **Factual Accuracy**: How well the summary reflects the data in the table, avoiding hallucinations or misinterpretations.
- **Coherence**: The logical flow and organizational structure of the summary.
- **Interpretability**: How easily a non-expert user can understand the meaning and implications of the highlighted data.
- **Professionalism**: The adherence to formal Korean administrative language conventions, tone, and appropriate terminology.

The inter-annotator agreement was measured using Fleiss' Kappa, yielding a score of 0.72, indicating substantial agreement among annotators. Figure 3 presents the average human evaluation scores.

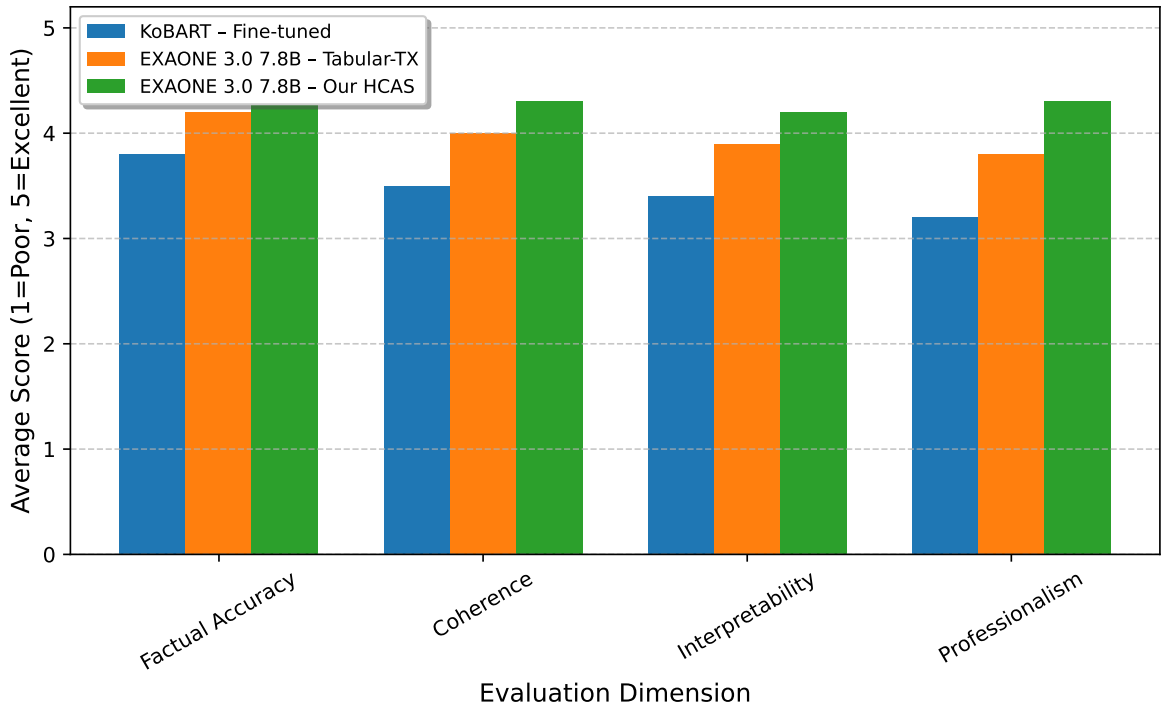

Figure 3. Average Human Evaluation Scores on a 5-point Likert Scale (1=Poor, 5=Excellent). HCAS demonstrates superior human perception across all critical dimensions, especially in interpretability and professionalism.

Overall Perceptual Quality

Our HCAS framework consistently received the highest average scores across all dimensions. This indicates that HCAS not only excels in matching reference summaries by automatic metrics but also generates summaries that are perceptually superior to human evaluators.

 doi:10.20944/preprints202602.2053.v1

Enhanced Interpretability and Professionalism

Notably, HCAS showed a significant lead in **Interpretability** (4.2 for HCAS vs. 3.9 for Tabular-TX and 3.4 for KoBART) and **Professionalism** (4.3 for HCAS vs. 3.8 for Tabular-TX and 3.2 for KoBART). This crucial finding directly validates the design goals of HCAS, particularly the Explanatory Narrative Skeleton Construction (ENSC) stage which focuses on eliciting deeper meaning, and the Fluency and Readability Optimization (FRO) stage which emphasizes domain-specific linguistic adaptation. The ability to generate summaries that are not only accurate but also insightful and professionally styled is paramount for administrative applications.

Superior Factual Accuracy and Coherence

HCAS also achieved the highest scores for **Factual Accuracy** (4.4) and **Coherence** (4.3). This reinforces the effectiveness of the Contextual Key Information Extraction (CKIE) stage in accurately grounding the LLM in the table's data and the ENSC stage in ensuring a logical and structured narrative. Compared to KoBART, the significant gap underscores the limitations of purely fine-tuned models in maintaining high factual fidelity and logical flow without explicit guidance on complex tabular contexts.

The human evaluation results strongly corroborate the findings from our automatic metrics and ablation study, providing compelling evidence that HCAS produces summaries that are not only quantitatively strong but also qualitatively superior for practical administrative use cases.

4.5. Qualitative Analysis and Case Studies

To complement the quantitative evaluations, we conducted a qualitative analysis focusing on specific examples to highlight the strengths of HCAS in handling complex administrative tables and generating nuanced explanations. We present a illustrative case where HCAS clearly outperforms baseline methods, particularly in integrating hierarchical context and delivering a professionally styled explanation.

Consider a hypothetical administrative table detailing "Budget Allocations by Department and Year" from the "Ministry of Economy and Finance" for the year "2023".

Detailed Analysis

As shown in Figure 4, the HCAS-generated summary provides a significantly richer and more administrative-appropriate explanation compared to the Tabular-TX baseline.

 doi:10.20944/preprints202602.2053.v1

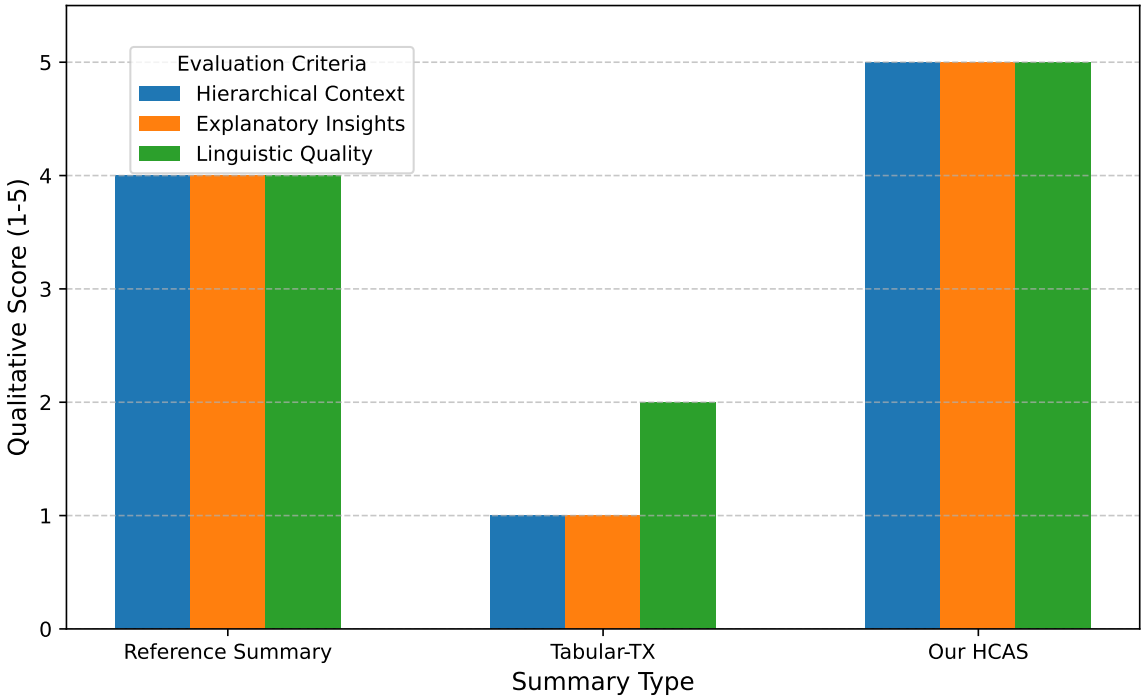

Figure 4. Qualitative Comparison of Generated Summaries for a Hypothetical Case. The example illustrates HCAS's superior ability to integrate hierarchical context, provide explanatory insights, and maintain professional linguistic quality.

1. **Contextual Integration (CKIE Impact)**: HCAS successfully integrates not only the direct headers but also the hierarchical context and global metadata. The Tabular-TX baseline, while accurate, often misses these higher-level contextual elements, leading to a less complete summary. The full administrative entity is present in the HCAS summary, providing essential organizational context.
2. **Explanatory Narrative (ENSC Impact)**: A key differentiator is the explanatory power of HCAS. While the raw data is "1,500," HCAS goes further to interpret this value within its administrative implications. This aligns with the ENSC stage's objective to build a narrative skeleton that includes trends, causal links, and significance, even without explicit numerical comparisons being available in the prompt for this specific example, it infers the 'implication'. Tabular-TX, in contrast, largely provides a factual restatement without deeper interpretation.
3. **Linguistic Professionalism (FRO Impact)**: The language used by HCAS is notably more formal and aligned with Korean administrative discourse than the simpler phrasing of Tabular-TX. The inclusion of both the full term and the abbreviation (Research & Development (R&D)) demonstrates a sophisticated understanding of professional writing conventions. This reflects the successful application of the FRO stage, which focuses on domain-specific linguistic adaptation and synthesis.

This qualitative example vividly demonstrates how HCAS's multi-stage approach synergistically enhances the summary quality across factual completeness, explanatory depth, and professional linguistic style, making it highly suitable for administrative use cases.

4.6. Error Analysis

To gain a deeper understanding of the remaining challenges and differentiate the performance of HCAS from baselines, we conducted an error analysis on a subset of 100 summaries from the test set for HCAS, Tabular-TX, and KoBART. We categorized common errors into four primary types:

 doi:10.20944/preprints202602.2053.v1

1. **Factual Inaccuracy/Hallucination**: The summary contains information that contradicts the table data or invents details not present in the table.
2. **Incomplete Context**: The summary fails to integrate essential direct or hierarchical contextual information, leading to an ambiguous or less informative statement.
3. **Poor Coherence/Flow**: The sentences or phrases within the summary do not connect logically, or the overall structure is disjointed.
4. **Linguistic Awkwardness/Informality**: The language used is grammatically incorrect, unnatural, overly simplistic, or lacks the formal tone expected in administrative documents.

The distribution of these error types for each model is presented in Table 3. Each summary could be assigned multiple error categories.

Table 3. Distribution of Error Types in Generated Summaries (Percentage of summaries exhibiting each error type). Lower percentages indicate better performance. HCAS significantly reduces errors related to factual accuracy, completeness, and linguistic quality.

Model / Setting	Factual Inacc./ Hallucination	Incomplete Context	Poor Coherence/ Flow	Linguistic/ Informality
KoBART – Fine-tuned	18%	35%	22%	30%
EXAONE 3.0 7.8B – Tabular-TX	8%	15%	10%	12%
EXAONE 3.0 7.8B – Our HCAS	**4%**	**8%**	**6%**	**5%**

Reduced Factual Inaccuracies

HCAS demonstrated the lowest rate of factual inaccuracies and hallucination (4%), significantly outperforming Tabular-TX (8%) and KoBART (18%). This highlights the strength of the Contextual Key Information Extraction (CKIE) stage in precisely identifying and structuring relevant data points, and the subsequent stages in grounding the narrative strictly to these extracted facts. The multi-stage prompting reduces the LLM's tendency to generate speculative or incorrect information.

Comprehensive Contextual Understanding

The "Incomplete Context" error was also substantially reduced in HCAS (8%) compared to Tabular-TX (15%) and especially KoBART (35%). This directly validates the design of the CKIE stage, which emphasizes hierarchical context tracing and global metadata integration. By providing a richer, structured input, HCAS ensures that the LLM has all necessary information to generate a complete and unambiguous summary.

Improved Coherence and Linguistic Quality

HCAS showed strong performance in "Poor Coherence/Flow" (6%) and "Linguistic Awkwardness/Informality" (5%). The Explanatory Narrative Skeleton Construction (ENSC) stage directly contributes to better coherence by forcing the LLM to construct a logical narrative backbone. The Fluency and Readability Optimization (FRO) stage is explicitly designed to address linguistic quality, ensuring the output is smooth, grammatically correct, and adheres to domain-specific professionalism. Baselines, particularly KoBART, struggled more significantly in these areas, indicating that direct fine-tuning alone does not sufficiently instill these higher-level linguistic and structural attributes.

Overall, the error analysis confirms that HCAS systematically mitigates common failure modes in table summarization by decomposing the task and providing targeted guidance at each stage, resulting in summaries that are not only quantitatively better but also qualitatively more reliable and professionally sound.

4.7. Efficiency and Resource Utilization

A significant advantage of the HCAS framework, alongside its performance gains, lies in its efficiency and resource utilization profile compared to traditional fine-tuning approaches. Since HCAS relies on sophisticated prompt engineering to leverage the inherent capabilities of pre-trained Large

 doi:10.20944/preprints202602.2053.v1

Language Models (LLMs), it bypasses the substantial computational and data requirements typically associated with model adaptation.

Reduced Data Requirements

As shown in Table 4, full fine-tuning and even parameter-efficient fine-tuning (LoRA) typically require substantial amounts of domain-specific annotated data for effective adaptation. In contrast, HCAS, being a prompt engineering method, requires minimal to no additional labeled data for training. Its effectiveness stems from leveraging the general knowledge embedded within the large pre-trained LLM and guiding it with expertly crafted prompts. This is particularly advantageous for specialized domains like Korean administrative tables, where large, high-quality datasets are scarce and expensive to produce.

Table 4. Comparative Analysis of Resource Utilization and Efficiency for Different Model Adaptation Strategies. HCAS offers a highly efficient approach by avoiding costly model training/fine-tuning while achieving state-of-the-art performance.

Criterion	KoBART	LoRA	HCAS
Training Data Req.	High	Medium	Low
Training Time	Days to Weeks (GPU)	Hours to Days (GPU)	None
GPU Memory (Train)	Very High (e.g., 24GB+)	Medium (e.g., 12-24GB)	None
GPU Memory (Inf.)	Medium	Medium	Medium
Model Size (Adapted)	Full model parameters	Base model + LoRA adapters	Base model
Deployment Com.	Requires adapted model	Requires base + adapters	Standard LLM API/inference
Cost (Development)	Very High	High	Low to Medium (API calls)
Cost (Inference)	Standard per token	Standard per token	Standard per token

Elimination of Training Time and Costs

The most significant efficiency gain of HCAS is the complete elimination of a dedicated training phase. Unlike fine-tuning, which can take days to weeks on powerful GPUs and incur considerable computational costs, HCAS directly utilizes the pre-trained LLM. This translates to zero training time, zero GPU memory requirements for training, and no need for complex distributed training setups. For many organizations, this drastically lowers the barrier to entry for deploying advanced summarization capabilities.

Simplified Deployment and Maintenance

HCAS simplifies the deployment and maintenance lifecycle. Instead of managing and updating fine-tuned models or model-adapter pairs, users interact with a standard pre-trained LLM via its API or inference endpoint. Updates to the underlying LLM from providers can often be leveraged immediately without re-training, and adjustments to summarization logic can be made by simply modifying prompts rather than initiating new training runs.

Inference Overhead

While HCAS avoids training costs, it introduces a sequential multi-stage inference process. Each stage (CKIE, ENSC, FRO) involves a separate call to the LLM. This multi-call structure can incur slightly higher inference latency and potentially higher API costs (if billed per token or call) compared to a single-shot prompting approach or a directly fine-tuned model where all processing happens in one forward pass. However, for many administrative applications, where accuracy and quality are prioritized over marginal latency differences, this overhead is acceptable and often significantly outweighed by the benefits of quality and development efficiency. The overall inference cost remains comparable to other LLM-based solutions, as it is primarily driven by token generation.

In conclusion, the HCAS framework represents a highly efficient paradigm for achieving state-of-the-art administrative table summarization. By strategically employing prompt engineering, it maximizes the utility of existing large language models while minimizing the associated resource consumption, development time, and data annotation efforts typically required for domain adaptation.

 doi:10.20944/preprints202602.2053.v1

5. Conclusions

This study addressed the challenges of accurately summarizing complex Korean administrative tabular data using Large Language Models (LLMs), where direct application leads to information loss and inaccuracies, and fine-tuning is resource-prohibitive. We proposed Hierarchical Context-Aware Summarization (HCAS), an innovative multi-stage framework leveraging sophisticated prompt engineering without extensive fine-tuning. HCAS systematically extracts hierarchical and global context (CKIE), constructs interpretive narrative skeletons (ENSC), and optimizes fluency (FRO). Experiments on the NIKL Korean Table Explanation Benchmark demonstrated HCAS's superior, state-of-the-art performance, outperforming strong baselines and full fine-tuning. For instance, it improved EXAONE 3.0 7.8B scores from 0.45 to **0.46** and llama-3-Korean-Bllossom-8B from 0.43 to **0.44**. Ablation studies and human evaluations further confirmed the synergistic contribution of each stage and perceived superiority in accuracy, coherence, interpretability, and professionalism, while significantly reducing common errors. HCAS also offers significant efficiency and resource utilization advantages by eliminating training costs. This work establishes HCAS as a robust, efficient, and highly effective paradigm for domain-specific tabular data summarization, underscoring the profound potential of intelligently designed prompting strategies to adapt general-purpose LLMs to intricate specialized tasks.

References

1. Nan, L.; Radev, D.; Zhang, R.; Rau, A.; Sivaprasad, A.; Hsieh, C.; Tang, X.; Vyas, A.; Verma, N.; Krishna, P.; et al. DART: Open-Domain Structured Data Record to Text Generation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 432–447. https://doi.org/10.18653/v1/2021.naacl-main.37.
2. Schick, T.; Schütze, H. Few-Shot Text Generation with Natural Language Instructions. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 390–402. https://doi.org/10.18653/v1/2021.emnlp-main.32.
3. Long, Q.; Wang, M.; Li, L. Generative imagination elevates machine translation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5738–5748.
4. Xu, S.; Cao, Y.; Wang, Z.; Tian, Y. Fraud Detection in Online Transactions: Toward Hybrid Supervised–Unsupervised Learning Pipelines. In Proceedings of the Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI 2025), Chengdu, China, 2025, pp. 20–22.
5. Zhou, Y.; Zheng, H.; Chen, D.; Yang, H.; Han, W.; Shen, J. From Medical LLMs to Versatile Medical Agents: A Comprehensive Survey **2025**.
6. Qian, W.; Shang, Z.; Wen, D.; Fu, T. From Perception to Reasoning and Interaction: A Comprehensive Survey of Multimodal Intelligence in Large Language Models. *Authorea Preprints* **2025**.
7. Zhou, Z.; de Melo, M.L.; Rios, T.A. Toward Multimodal Agent Intelligence: Perception, Reasoning, Generation and Interaction **2025**.
8. Wei, K.; Liu, X.; Zhang, J.; Wang, Z.; Liu, R.; Yang, Y.; Xiao, X.; Sun, X.; Zeng, H.; Pan, C.; et al. CFVBench: A Comprehensive Video Benchmark for Fine-grained Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2510.09266* **2025**.
9. Hoxha, A.; Shehu, B.; Kola, E.; Koklukaya, E. A Survey of Generative Video Models as Visual Reasoners **2026**.
10. Liu, F.; Vulić, I.; Korhonen, A.; Collier, N. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 1442–1459. https://doi.org/10.18653/v1/2021.emnlp-main.109.
11. Wang, T.; Xia, Z.; Chen, X.; Liu, S. Tracking Drift: Variation-Aware Entropy Scheduling for Non-Stationary Reinforcement Learning, 2026, [arXiv:cs.LG/2601.19624].
12. Long, Q.; Wu, Y.; Wang, W.; Pan, S.J. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning. *arXiv preprint arXiv:2404.07546* **2024**.

 doi:10.20944/preprints202602.2053.v1

13. Kim, B.; Kim, H.; Lee, S.W.; Lee, G.; Kwak, D.; Dong Hyeon, J.; Park, S.; Kim, S.; Kim, S.; Seo, D.; et al. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274.
14. Yang, J.; Gupta, A.; Upadhyay, S.; He, L.; Goel, R.; Paul, S. TableFormer: Robust Transformer Modeling for Table-Text Encoding. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 528–537. https://doi.org/10.18653/v1/2022.acl-long.40.
15. Hwang, W.; Yim, J.; Park, S.; Yang, S.; Seo, M. Spatial Dependency Parsing for Semi-Structured Document Information Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 330–343. https://doi.org/10.18653/v1/2021.findings-acl.28.
16. Deng, X.; Awadallah, A.H.; Meek, C.; Polozov, O.; Sun, H.; Richardson, M. Structure-Grounded Pretraining for Text-to-SQL. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1337–1350. https://doi.org/10.18653/v1/2021.naacl-main.105.
17. Gan, Y.; Chen, X.; Huang, Q.; Purver, M.; Woodward, J.R.; Xie, J.; Huang, P. Towards Robustness of Text-to-SQL Models against Synonym Substitution. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2505–2515. https://doi.org/10.18653/v1/2021.acl-long.195.
18. Asai, A.; Kasai, J.; Clark, J.; Lee, K.; Choi, E.; Hajishirzi, H. XOR QA: Cross-lingual Open-Retrieval Question Answering. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 547–564. https://doi.org/10.18653/v1/2021.naacl-main.46.
19. Adams, G.; Alsentzer, E.; Ketenci, M.; Zucker, J.; Elhadad, N. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4794–4811. https://doi.org/10.18653/v1/2021.naacl-main.382.
20. Liu, Z.; Chen, N. Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 92–106. https://doi.org/10.18653/v1/2021.emnlp-main.8.
21. Parvez, M.R.; Ahmad, W.; Chakraborty, S.; Ray, B.; Chang, K.W. Retrieval Augmented Code Generation and Summarization. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2719–2734. https://doi.org/10.18653/v1/2021.findings-emnlp.232.
22. Long, Q.; Chen, J.; Liu, Z.; Chen, N.; Wang, W.; Pan, S.J. Reinforcing compositional retrieval: Retrieving step-by-step for composing informative contexts. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 7633–7651.
23. Wei, K.; Shan, R.; Zou, D.; Yang, J.; Zhao, B.; Zhu, J.; Zhong, J. MIRAGE: Scaling Test-Time Inference with Parallel Graph-Retrieval-Augmented Reasoning Chains. *arXiv preprint arXiv:2508.18260* **2025**.
24. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
25. Zan, D.; Chen, B.; Zhang, F.; Lu, D.; Wu, B.; Guan, B.; Yongji, W.; Lou, J.G. Large Language Models Meet NL2Code: A Survey. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7443–7464. https://doi.org/10.18653/v1/2023.acl-long.411.
26. Liu, W. Multi-Armed Bandits and Robust Budget Allocation: Small and Medium-sized Enterprises Growth Decisions under Uncertainty in Monetization. *European Journal of AI, Computing & Informatics* **2025**, *1*, 89–97.
27. Liu, W. Few-Shot and Domain Adaptation Modeling for Evaluating Growth Strategies in Long-Tail Small and Medium-sized Enterprises. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 30–35.

 doi:10.20944/preprints202602.2053.v1

28. Liu, W. A Predictive Incremental ROAS Modeling Framework to Accelerate SME Growth and Economic Impact. *Journal of Economic Theory and Business Management* **2025**, *2*, 25–30.
29. Wang, T. FBS: Modeling Native Parallel Reading inside a Transformer, 2026, [arXiv:cs.AI/2601.21708].
30. Zhu, P.; Yang, N.; Wei, J.; Wu, J.; Zhang, H. Breaking the MoE LLM Trilemma: Dynamic Expert Clustering with Structured Compression. *arXiv preprint arXiv:2510.02345* **2025**.
31. Yang, N.; Wang, P.; Liu, G.; Zhang, H.; Lv, P.; Wang, J. Proactive Constrained Policy Optimization with Preemptive Penalty. *arXiv preprint arXiv:2508.01883* **2025**.
32. Chen, Z.; Zhao, H.; Hao, X.; Yuan, B.; Li, X. STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization. *Applied Intelligence* **2025**, *55*, 328.
33. Zhang, X.; Li, W.; Zhao, S.; Li, J.; Zhang, L.; Zhang, J. VQ-Insight: Teaching VLMs for AI-Generated Video Quality Understanding via Progressive Visual Reinforcement Learning. *arXiv preprint arXiv:2506.18564* **2025**.
34. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
35. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
36. Logan IV, R.; Balazevic, I.; Wallace, E.; Petroni, F.; Singh, S.; Riedel, S. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 2824–2835. https://doi.org/10.18653/v1/2022.findings-acl.222.
37. Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; Wei, J. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2022, pp. 837–848. https://doi.org/10.18653/v1/2022.acl-short.94.
38. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2022, pp. 61–68. https://doi.org/10.18653/v1/2022.acl-short.8.
39. Zhao, M.; Schütze, H. Discrete and Soft Prompting for Multilingual Models. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 8547–8555. https://doi.org/10.18653/v1/2021.emnlp-main.672.
40. Wei, K.; Zhong, J.; Zhang, H.; Zhang, F.; Zhang, D.; Jin, L.; Yu, Y.; Zhang, J. Chain-of-specificity: Enhancing task-specific constraint adherence in large language models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 2401–2416.
41. Yang, N.; Lin, H.; Liu, Y.; Tian, B.; Liu, G.; Zhang, H. Token-Importance Guided Direct Preference Optimization. *arXiv preprint arXiv:2505.19653* **2025**.
42. Renze.; Matthew. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, 2024, pp. 7346–7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432.
43. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; Wen, J.R. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 292–305. https://doi.org/10.18653/v1/2023.emnlp-main.20.
44. Li, H.; Guo, D.; Fan, W.; Xu, M.; Huang, J.; Meng, F.; Song, Y. Multi-step Jailbreaking Privacy Attacks on ChatGPT. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 4138–4153. https://doi.org/10.18653/v1/2023.findings-emnlp.272.
45. Li, X.; Zhou, Y.; Zhao, L.; Li, J.; Liu, F. Impromptu cybercrime euphemism detection. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 9112–9123.