LLM-based Grant Proposal Structuring via Classification, Extraction, and Compression

Zhe Yang[1,2], Wenjing Chang[1,2], Yue Wang[1], and Xiaojun Zhou[1,*]
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
Email: yangzhe@cnic.cn, changwenjing@cnic.cn, yuewang@cnic.cn, xjzhou@cnic.cn

***Abstract*—Research grant proposals contain rich scientific ideas and technical methodologies, serving as a critical resource for AI for Science. However, their multi-disciplinary nature, scattered distribution of core information, and heterogeneous writing styles pose significant challenges for neural document understanding and long-context reasoning. To address these challenges, we propose GRIST (Grant pRoposal Intelligent STructuring), a structure- and semantic-aware LLM-based framework that transforms lengthy proposals into structured, high-density representations. GRIST integrates instruction-driven discipline classification, metadata extraction, and semantic compression, enabling automated analysis through in-context learning without task-specific training. To facilitate research in this domain, we construct and release CGP-Bench, a Chinese grant proposal benchmark comprising 300 anonymized proposals curated for privacy compliance while preserving scientific logic. We conduct a comprehensive empirical analysis of mainstream LLMs on CGP-Bench, providing insights into their generalization boundaries and reasoning capabilities in complex grant proposal processing.**

***Index Terms*—LLMs, Grant Proposals, Information Extraction, Semantic Compression**

I. Introduction

Research grant proposals are critical documents that convey scientific ideas and expected outcomes to funding agencies. Unlike standardized published papers, grant proposals are prospective, encompassing state-of-the-art reviews, proposed technical roadmaps, and anticipated deliverables [1]. However, as global research investment surges, the sheer volume and complexity of these documents have created a bottleneck [2]. Review committees struggle to extract key information from proposals exceeding 10k words, a cognitively demanding and time-sensitive process [3], [4].

Recent breakthroughs in Large Language Models (LLMs) have demonstrated remarkable capabilities in neural document understanding for complex scientific literature, spanning tasks from semantic summarization and fine-grained information extraction to reasoning-intensive question answering. These advances offer promising opportunities for automating high-stakes and labor-intensive tasks in research management [5]–[8]. Yet, deploying general-purpose LLMs to grant proposal analysis is far from trivial, facing two primary challenges.

*Corresponding author: xjzhou@cnic.cn.
Supported by the Informatization Plan of Chinese Academy of Sciences, Grant CAS-WX2025GC-07.

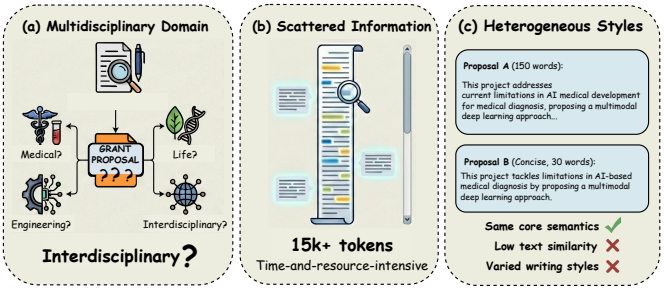

Fig. 1. An illustration of the intrinsic structural and semantic complexity of grant proposals.

The first challenge arises from the intrinsic structural and semantic complexity of grant proposals, as illustrated in Fig. 1. Grant proposals span multiple disciplines, necessitating dynamic adaptability for models to interpret content correctly without task-specific fine-tuning. Besides, key information is sparsely distributed across documents exceeding $15k$ tokens, where establishing logical links between distant sections severely tests the context retention and attention mechanisms of current models. Additionally, highly personalized writing styles and diverse linguistic expressions for similar scientific concepts challenge robust semantic alignment. ***The second challenge is the scarcity of domain-specific resources for benchmarking.*** Existing scientific text datasets primarily focus on published papers [9]–[13], leaving grant proposals largely unexplored due to their sensitive and unpublished nature. This lacks specialized benchmarks for fine-grained extraction or reasoning over prospective narratives, hindering systematic assessment of LLMs in this high-stakes domain and leaving it unclear if generic models can reliably support human experts.

To bridge these gaps, we propose **GRIST** (Grant pRoposal Intelligent STructuring), a structure- and semantic-aware framework that automates the parsing of lengthy grant proposals into structured, high-density representations. First, to handle multidisciplinary ambiguity, we introduce a *context-aware discipline classification* module that performs global semantic reasoning. Second, to alleviate the long-range dependency problem, we design an *instruction-driven metadata extraction* module that leverages the hierarchical structure to locate and retrieve metadata fields. Third, to mitigate semantic

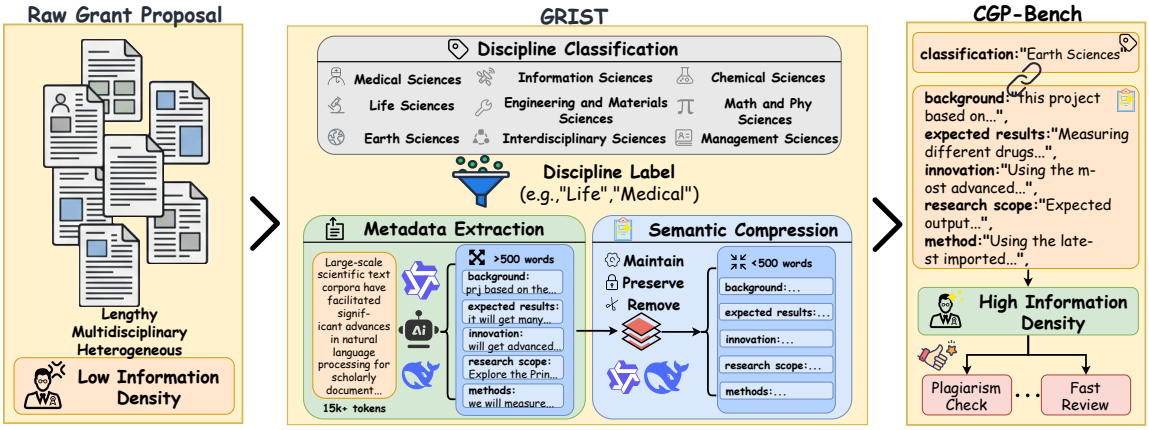

Fig. 2. Overview of GRIST and CGP-Bench. GRIST transforms raw grant proposals into structured, high-density representations through discipline classification, metadata extraction, and semantic compression, supporting downstream tasks such as plagiarism detection and rapid review.

heterogeneity, we devise a *semantic compression* module that maps diverse writing styles into concise representations for standardized cross-proposal comparison. To facilitate research in this domain, we construct and release CGP-Bench, a Chinese grant proposal benchmark comprising 300 manually collected and anonymized proposals that preserve scientific coherence, along with a subset of high-quality expert-annotated samples for evaluation. We evaluate five LLMs on GRIST, revealing performance variations across different task types and providing empirical insights into model capabilities for processing complex scientific documents.

The main contributions of this work are as follows:

- We propose GRIST, a structure- and semantic-aware framework that automates the extraction of scientific intelligence from lengthy, unstructured grant proposals.
- We release CGP-Bench, the first dedicated benchmark for Chinese grant proposal analysis, filling the gap in publicly available resources for this domain.
- We conduct a comprehensive empirical analysis of mainstream LLMs on CGP-Bench, revealing generalization boundaries of current models and highlighting performance trade-offs between extraction precision and semantic compression fidelity.

Dataset available at https://github.com/Makiroooooo/GRIST

II. Related Work

A. Scientific Text Datasets

Large-scale scientific text corpora have facilitated significant advances in natural language processing for scholarly documents. S2ORC [11] provides over 81 million academic papers with structured text and citation graphs. Similarly, PubMed [14] and arXiv [15] hosts millions of papers across medicine and STEM domains. For Chinese scientific literature, CSL [12] offers a large-scale corpus of paper abstracts and keywords. However, these datasets focus exclusively on published papers. In contrast, grant proposals remain largely unexplored due to their unpublished nature and restricted accessibility.

B. LLM for Scientific Document Analysis

LLMs have demonstrated strong capabilities in various scientific document processing tasks. For information extraction, Dagdelen et al. [16] showed that LLMs can perform named entity recognition and relation extraction from scientific literature through fine-tuning and in-context learning. Zhang et al. [17] proposed a multi-agent framework for extracting key insights from scientific articles. For text summarization, LLMs exhibit competitive zero-shot performance on scientific articles [18], and models with extended context windows such as Longformer [19] have been applied to long-document summarization. Beyond summarization, semantic compression aims to condense text while preserving core semantics under high compression ratios [20]. However, these methods primarily target published papers that report completed results, whereas grant proposals describe planned research with prospective elements such as innovation points and technical roadmaps.Moreover, their heterogeneous styles and diverse logical structures pose challenges conceptually related to robust representation learning across distributions [21].

III. Methodology

This section presents GRIST, a pipeline that transforms lengthy, unstructured grant proposals into compact and comparable representations, supporting downstream tasks such as plagiarism detection and rapid review. As illustrated in Fig. 2, given a raw proposal document, GRIST first performs document preprocessing to convert PDF or Word files into structured Markdown, preserving the logical hierarchy while removing noise such as headers and watermarks. The preprocessed document then flows through three core modules. First, the discipline classification module takes the full document as input and assigns it to the appropriate category in the NSFC taxonomy for efficient retrieval. The metadata extraction module then identifies five key fields, namely research background, innovation points, research content, technical approaches, and expected outcomes, organizing scattered information into structured records. Finally, the semantic compression module

condenses the extracted content into concise representations that preserve the core research logic regardless of writing style.

A. Problem Formulation

We formulate structuring as multi-stage inference. Given raw documen d, we reconstruct a structured tuple $\mathcal{S}^* = (c, \mathbf{m}, \mathbf{s})$ (discipline, metadata, compressed semantics) by maximizing the joint posterior, decomposed via logical dependencies:

$$\mathcal{S}^* \approx \underset{c,\mathbf{m},\mathbf{s}}{\operatorname{argmax}} \underbrace{P(c|d)}_{\text{Anchoring}} \cdot \underbrace{P(\mathbf{m}|d,c)}_{\text{Extraction}} \cdot \underbrace{P(\mathbf{s}|\mathbf{m},\Omega)}_{\text{Compression}}, \quad (1)$$

where Ω represents logic-preserving constraints. We approximate these sub-problems using an LLM as a reasoning agent.

B. Document Preprocessing

Raw grant proposals are typically stored in PDF or Word formats, where inconsistent encoding and complex layouts often disrupt the logical flow of content during parsing. To preserve document structure, we convert all proposals into Markdown format using an open-source document parsing tool [22]. Formally, we define a structural parsing operator $\mathcal{T}$ that maps the raw document space $\mathcal{D}$ to a hierarchical token sequence space $\tilde{\mathcal{D}}$: $\tilde{d}_i = \mathcal{T}(d_i) = \langle (h_k, x_k) \rangle_{k=1}^{L_i}$, where x_k represents the text content of the k-th segment and $h_k \in \mathbb{N}$ denotes its hierarchical level (e.g., section, subsection). This representation $\tilde{d}_i$ serves as the unified input for the subsequent inference steps. We choose Markdown as the intermediate representation because its hierarchical heading syntax explicitly encodes section boundaries, facilitating structural understanding by LLMs. The conversion also filters out semantically irrelevant elements such as headers, footers, and watermarks, yielding clean input for downstream modules.

C. Discipline Classification

Grant proposals span diverse research domains, and accurate categorization is essential for routing documents to appropriate reviewers and enabling discipline-specific retrieval. Following the taxonomy of the National Natural Science Foundation of China, we define nine target categories: Mathematical and Physical Sciences, Chemical Sciences, Life Sciences, Medical Sciences, Engineering and Materials Sciences, Information Sciences, Earth Sciences, Management Sciences, and Interdisciplinary Sciences.

Keyword-based methods often fail for classification due to interdisciplinary keyword co-occurrence and the necessity to evaluate objectives and methodology holistically rather than via isolated terms. To address these limitations, we formulate discipline classification as estimating the posterior distribution of the category variable c conditioned on the global document context. The optimal category c_i^* is obtained by maximizing the likelihood over the predefined taxonomy $\mathcal{C}$:

$$c_i^* = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P_\Theta(c \mid \tilde{d}_i, \mathcal{I}_{cls}), \quad (2)$$

where Θ represents the LLM parameters and $\mathcal{I}_{cls}$ is the task-specific instruction acting as a reasoning guide. This step anchors the global semantic scope for the extraction process. As shown in Table I(a), the LLM receives the full preprocessed document along with category definitions and identifies the dominant discipline based on the overall research focus.

TABLE I
PROMPT FOR DISCIPLINE CLASSIFICATION, METADATA EXTRACTION AND SEMANTIC COMPRESSION.

(a) Prompt:

 # Task
Classify the given research grant proposal into exactly one of the following nine disciplines based on its overall research objectives, methodology ...
 # Input Grant Proposal
< Full proposal content... >

LLM Outputs:
One of the nine discipline categories.

(b) Prompt:

 # Task
Extract the following five metadata fields from the given research grant proposal. Locate and return the relevant content directly from the source text.
1. **Background**: *Current state of the field and unresolved scientific problems that motivate the proposed work.*
2. .....
 # Input Grant Proposal
< Full proposal content... >

LLM Outputs:
{"research_background": "...", "novelty": "...", "research_scope": "...", "methodology": "...", "expected_outcomes": "..."}

(c) Prompt:

 # Task
Compress the following text while preserving its core semantic content.
 # Compression Principles
1. **Remove**: Verbose modifiers, excessive background elaboration, ...
2. **Preserve**: The logical chain connecting research motivation, ...
3. **Maintain**: Factual accuracy and argumentative structure of the original text.
 # Input Text
< Extracted metadata content... >

LLM Outputs:
Compressed text that retains core semantics with reduced length.

D. Metadata Extraction

Discipline classification provides a coarse-grained view of proposals, i.e., the research domain. However, domain-specific reviewers also need to quickly locate specific information from lengthy proposals, such as research gaps or key problems to be addressed, which is critical for effective evaluation. Thus, we design a metadata extraction module that transforms unstructured text into structured records.

The extraction schema is derived from the cognitive focus of domain experts during proposal review. Through consultation with experienced reviewers, we identify five fields that capture the essential aspects of a grant proposal. *Background* describes the current state of the field and the unresolved scientific problems that motivate the proposed work. *Innovation* articulates the core theoretical or methodological contributions that distinguish this proposal from existing studies. *Research Scope* specifies the concrete research tasks and their logical organization. *Methods* details the approaches, experimental designs, or algorithmic frameworks employed to achieve the research objectives. *Expected Results* outlines the anticipated outcomes and deliverables upon project completion.

We define the extraction process as a schema-constrained conditional generation task. Let $\mathcal{M} = \{\phi_1, \ldots, \phi_J\}$ be the semantic definitions of the target metadata fields. Assuming conditional independence between fields given the document context, we maximize the joint probability:

$$\mathbf{m}_i^* = \underset{\mathbf{m}}{\operatorname{argmax}} \prod_{j=1}^{J} P_\Theta(m^{(j)} \mid \tilde{d}_i, \phi_j, \mathcal{I}_{ext}). \tag{3}$$

Here, the LLM acts as a neural extractor where ϕ_j directs the attention mechanism to relevant local distinct regions within $\tilde{d}_i$. As shown in Table I(b), the LLM receives the full proposal along with field definitions and locates relevant content directly from the source text.

E. Semantic Compression

The extracted metadata retains the original wording from proposals, which varies substantially in length and writing style across different authors. Such heterogeneity hinders direct comparison between proposals, particularly for plagiarism detection where surface-level similarity can be low despite shared underlying ideas. To address this, we introduce a semantic compression module that condenses the extracted content into concise representations.

We formulate semantic compression as an optimization problem balancing information density and semantic fidelity. For metadata field $m_i^{(j)}$, we seek a compressed representation $s_i^{(j)}$ that minimizes a composite loss function $\mathcal{L}_{cmp}$:

$$s_i^{(j)*} = \underset{s}{\operatorname{argmin}} \left(\ell(s) - \lambda \cdot \mathrm{Sim}(s, m_i^{(j)}) \right), \quad \text{s.t.} \quad \Omega(s) = 1, \tag{4}$$

where $\ell(\cdot)$ measures token length, $\mathrm{Sim}(\cdot,\cdot)$ denotes semantic similarity, λ is a balancing hyperparameter, and Ω enforces logical coherence. Since directly optimizing Eq. (4) is intractable for black-box LLMs, we approximate this objective through instruction-driven generation via the prompt $\mathcal{I}_{cmp}$:

$$s_i^{(j)} \approx \mathrm{LLM}_\Theta(m_i^{(j)}, \mathcal{I}_{cmp}). \tag{5}$$

As shown in Table I(c), $\mathcal{I}_{cmp}$ operationalizes optimization objective by eliminating verbose expressions to minimize ℓ (lossy compression on redundancy) and preserving logical chain to maximize Sim (lossless compression on core logic). The compression is applied to each metadata field independently, producing compact representations for downstream tasks such as similarity-based retrieval and plagiarism detection.

IV. Experiments

In this section, we evaluate the capability of mainstream LLMs in processing grant proposals on the proposed GRIST pipeline. We investigate the following questions:

- ***RQ1****: How do LLMs perform on discipline classification for research grant proposals?*
- ***RQ2****: Which metadata fields pose greater challenges for LLMs, and how do models differ in fine-grained information extraction?*

TABLE II
Statistical characteristics of CGP-Bench.

Discipline	Ratio(%)	Avg. Length
Medical Sciences	36.00	16,945
Life Sciences	28.00	16,042
Engineering and Materials Sciences	16.33	18,236
Earth Sciences	7.00	16,452
Information Sciences	4.67	19,336
Interdisciplinary Sciences	3.33	17,726
Chemical Sciences	2.33	16,765
Mathematical and Physical Sciences	1.33	16,861
Management Sciences	1.00	19,717

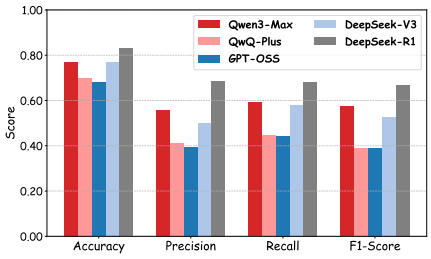

Fig. 3. Performance comparison on discipline classification task.

- ***RQ3****: Can current LLMs compress extracted content while preserving core research logic under high compression ratios?*

A. Experimental Setup

1) CGP-Bench Dataset: Due to the scarcity of publicly available grant proposal resources, we collect 300 Chinese research proposals from publicly accessible sources. The proposals span nine disciplines following the NSFC taxonomy. Table II summarizes the dataset statistics. All documents are processed through the GRIST pipeline to obtain structured annotations, including discipline labels, metadata fields, and compressed representations. To evaluate the quality of LLM-generated annotations, we choose 60 proposals for manual annotation as ground truth. We invited researchers with relevant disciplinary backgrounds for discipline classification and metadata extraction. Two annotators with relevant disciplinary backgrounds independently label discipline categories and metadata fields, with expert arbitration for disagreements.

2) Evaluation models: Processing grant proposals requires both long-context comprehension and complex reasoning capabilities. In these experiments, we select five models: DeepSeek-V3-671B [23], Qwen-Max [24], GPT-oss-120B [25], DeepSeek-R1-671B [26], and QwQ-Plus [27]. This selection includes both base and reasoning-enhanced architectures, as well as models with Chinese-specific pretraining.

B. Discipline Classification

We formulate discipline classification as a nine-way task following the NSFC taxonomy. We evaluate the performance using Accuracy, Precision, Recall, and F1-score. The results are shown in Fig. 3. We have the following observations.

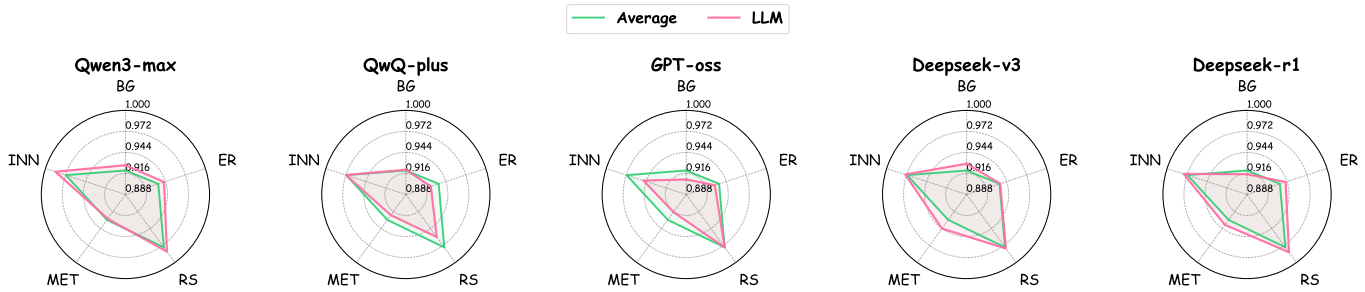

Fig. 4. Comparison of semantic similarity for metadata extraction. Each radar chart displays performance on five fields: background (BG), expected results (ER), innovation (INN), methods (MET), and research scope (RS). The green line indicates the average performance across all models.

TABLE III
PERFORMANCE COMPARISON ON METADATA EXTRACTION. THE BEST AND RUNNER-UP RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE.

Model	OSS	AFSS
Qwen3-max	0.973	0.948
QwQ-plus	0.968	0.930
GPT-oss	0.972	0.947
DeepSeek-V3	0.959	0.944
DeepSeek-R1	**0.974**	**0.967**

TABLE IV
PERFORMANCE COMPARISON ON SEMANTIC COMPRESSION. THE BEST AND RUNNER-UP RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE.

Model	CS	CR	ERE	SRE
Qwen3-max	**0.946**	0.418	0.618	0.391
QwQ-plus	0.919	0.606	0.679	0.555
GPT-oss	0.897	0.633	0.701	0.562
Deepseek-v3	0.883	0.654	0.719	0.577
DeepSeek-R1	0.898	**0.699**	**0.734**	**0.627**

LLMs demonstrate promising performance, with DeepSeek-R1 achieving the highest score, indicating that reasoning enhancement can significantly improve classification accuracy. However, QwQ-Plus underperforms Qwen3-Max, suggesting that while reasoning capabilities facilitate classification, the strength of the base model remains critical. These results imply that both top-tier reasoning models and robust general-purpose models are effective for grant proposal classification.

C. Metadata Extraction

This experiment evaluates the model's capability to extract fine-grained metadata from proposal texts. Specifically, the model is provided with target field names along with their semantic descriptions, and is required to extract the corresponding content from the input document.

We adopt three complementary metrics for evaluation. *Overall Semantic Similarity (OSS)* measures global coverage by comparing concatenated fields to ground truth. *Average Field Semantic Similarity (AFSS)* averages individual field similarities for fine-grained accuracy. *Per-field Semantic Similarity* reports scores for each of the five fields separately, enabling detailed analysis across information types.

Table III presents the metadata extraction performance of different models in terms of OSS and AFSS. All models achieve OSS scores above 0.95, indicating that current LLMs can reliably capture the overall content of grant proposals. AFSS scores range from 0.93 to 0.97, showing more variation in fine-grained extraction.

To further investigate where these differences come from, we visualize the per-field semantic similarity in Fig. 4. We can observe that INN consistently yields high scores across all models, while BG proves notably more difficult. This disparity likely stems from the nature of the content. Innovation statements are typically presented in concise, self-contained sentences with explicit markers, making them easier to locate and extract. Background information is often scattered throughout the document and mixed with other descriptions, making it harder for models to extract cleanly.

D. Semantic Compression

This experiment evaluates the model's ability to preserve logical coherence under high-ratio semantic compression. We select metadata segments exceeding 500 characters from the original proposals and require the model to condense them while retaining key information.

Following [20], we employ four metrics to balance compression ratio and information retention. *Compression Ratio (CR)* quantifies the degree of text reduction, i.e., $CR = 1 - L_c/L_o$ where L_c and L_o denote the length of compressed and original text. *Cosine Similarity (CS)* computes the semantic similarity between the compressed and original text embeddings. *Exact Reconstructive Effectiveness (ERE)* measures lexical-level preservation under compression, i.e., $ERE = (\log(CR) \times ED)^{-1}$ where ED is the edit distance between compressed and original text. *Semantic Reconstruction Effectiveness (SRE)* captures whether core semantics are retained at a given compression level, i.e., $SRE = CR \times CS$. The results are shown in Table IV.

The results reveal a trade-off between compression ratio and semantic similarity. Qwen3-Max achieves the highest CS but the lowest CR, indicating a tendency to preserve the original text rather than condensing it. This conservative compression strategy maintains semantic fidelity at the expense of informa-

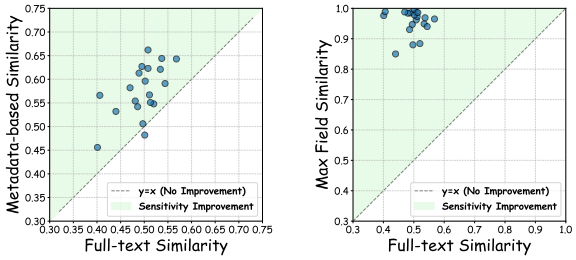

Fig. 5. Metadata-based vs. full-text similarity on paraphrased plagiarism samples. Left: average metadata similarity vs. full-text similarity. Right: maximum single-field similarity vs. full-text similarity.

tion density, resulting in lower ERE and SRE scores. Besides, while all models maintain high CS, CR varies substantially, propagating to ERE and SRE. This indicates that LLMs have largely converged in semantic preservation but diverge in compression behavior, reflecting different underlying strategies for balancing fidelity and conciseness.

E. Plagiarism Detection

Full-text similarity methods for detecting paraphrased plagiarism are vulnerable to surface-level manipulation such as rewriting background descriptions or reorganizing paragraphs, while the core research logic often remains intact. To provide a preliminary investigation of this scenario, we construct a controlled dataset of 20 source-suspect document pairs. Each suspect document is substantially rewritten to reduce overall textual similarity to approximately 50%, while preserving the logical structure of a specific metadata field. While this synthetic setup cannot fully capture the complexity of real-world plagiarism, it enables controlled evaluation of whether GRIST-extracted metadata can detect semantic-level similarities that surface-level methods may overlook.

The results are shown in Fig. 5, where the x-axis denotes full-text similarity and the y-axis denotes metadata-based similarity. We can observe that data points predominantly cluster in the upper-left region, indicating that metadata-based similarity detection consistently yields higher scores than full-text-based detection. This confirms that GRIST effectively captures preserved logical structures even when surface-level similarity has been deliberately reduced.

V. Conclusion

This paper proposes GRIST, a structure- and semantic-aware framework addressing multidisciplinary ambiguity, long-range dependencies, and stylistic heterogeneity in grant proposals. By integrating discipline classification, metadata extraction, and semantic compression, GRIST reconstructs structured scientific intelligence. We also release CGP-Bench, the first benchmark of 300 Chinese proposals, filling a critical data gap and laying the foundation for intelligent grant review systems.

References

[1] M. Cargill and P. O'Connor, *Writing scientific research articles: Strategy and steps*, 3rd ed. John Wiley & Sons, 2021, see Chapter: Writing research grant proposals.

[2] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2215–2222, 2015.

[3] W. Liang, Y. Zhang, H. Cao *et al.*, "Can large language models provide useful feedback on research papers? a large-scale empirical analysis," *Nature Human Behaviour*, 2024.

[4] A. Recio Saucedo, K. Crane, K. Meadmore *et al.*, "What works for peer review and decision-making in research funding: a realist synthesis," *Research Integrity and Peer Review*, vol. 7, no. 1, 2022.

[5] A. Checco, L. Bracciale, P. Loreti *et al.*, "Ai-assisted peer review," *Humanities and social sciences communications*, vol. 8, no. 1, 2021.

[6] L. A. Cox Jr, "An ai assistant to help review and improve causal reasoning in epidemiological documents," *Global epidemiology*, 2023.

[7] W. Chang, G. Qi, Y. Wang, and J. Yu, "Llm-enhanced heterogeneous graph neural networks for research project conflict detection," in *2025 International Joint Conference on Neural Networks*, 2025, pp. 1–8.

[8] Y. Wang, N. Yi, W. Chang, and J. Yu, "Leveraging llms for automated compliance risk identification in business processes," in *International Conference on Neural Information Processing*, 2024, pp. 456–471.

[9] Y. Luan, L. He, M. Ostendorf *et al.*, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3219–3232.

[10] S. Jain, M. van Zuylen, H. Hajishirzi *et al.*, "Scirex: A challenge dataset for document-level information extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[11] K. Lo, L. L. Wang, M. Neumann *et al.*, "S2orc: The semantic scholar open research corpus," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 4969–4983.

[12] Y. Li, Y. Zhang, Z. Zhao *et al.*, "CSL: A large-scale chinese scientific literature dataset," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 3917–3923.

[13] Y. Zhang, V. Zhong, D. Chen *et al.*, "Position-aware attention and supervised data improve slot filling," in *Conference on Empirical Methods in Natural Language Processing*, 2017.

[14] K. Canese and S. Weis, "Pubmed: The bibliographic database," *The NCBI Handbook*, 2013.

[15] C. B. Clement, M. Biber, J. P. Kelly *et al.*, "On the use of arxiv as a dataset," in *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, 2019.

[16] J. Dagdelen, A. Dunn, S. Lee *et al.*, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, 2024.

[17] Z. Song, G.-Y. Hwang, X. Zhang *et al.*, "A scientific-article key-insight extraction system based on multi-actor of fine-tuned open-source large language models," *Scientific Reports*, vol. 15, no. 1, p. 1608, 2025.

[18] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding *et al.*, "Evaluating large language models on medical evidence summarization," *NPJ Digital Medicine*, vol. 6, no. 1, p. 158, 2023.

[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[20] H. Gilbert, M. Sandborn, D. C. Schmidt *et al.*, "Semantic compression with large language models," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security*, 2023, pp. 1–8.

[21] K. Shi, J. Lu, S. Ye, G. Zhang, and Z. Fang, "Mirage: Multimodal discriminative representation learning for generalizable ai-generated image detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 353–361.

[22] B. Wang, C. Xu *et al.*, "Mineru: An open-source solution for precise document content extraction," *arXiv preprint arXiv:2409.18839*, 2024.

[23] A. Liu, B. Feng, B. Xue *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[24] A. Yang, A. Li, B. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[25] S. Agarwal, L. Ahmad, J. Ai *et al.*, "gpt-oss-120b & gpt-oss-20b model card," *arXiv preprint arXiv:2508.10925*, 2025.

[26] D. Guo, D. Yang, H. Zhang *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[27] Qwen Team, "QwQ: Reflecting deeply on reasoning," Alibaba Group, Technical Report, 2024. [Online]. Available: https://qwenlm.github.io/blog/qwq-32b-preview/