**Journal of Computer Science and Electrical Engineering**
**Print ISSN: 2663-1938**
**Online ISSN: 2663-1946**
**DOI: https://doi.org/10.61784/jcsee3080**

# HUMAN-AI CO-CREATION SYSTEM FOR KNOWLEDGE WORK BASED ON MULTI-AGENT APPROACH

WeiJing Zhu[1], RunTao Ren[2*], Wei Xie[1], CenYing Yang[2]
[1]*Guangxi Science and Technology Information Network Center, Nanning 530022, Guangxi, China*
[2]*City University of Hong Kong, Kowloon Tong, Hong Kong region 999077, China.*
*Corresponding Author: RunTao Ren, Email: runtaoren2-c@my.cityu.edu.hk*

**Abstract:** The task of writing a scientific research proposal is complex and highly structured, yet traditional writing methods are typically inefficient. To address this challenge, this study introduces an intelligent writing system leveraging a large language model (LLM). The core contribution is a modular proposal drafting framework that automatically generates multi-chapter application content based on user-specified disciplines and topic questions. The system first creates outlines and content overviews for each chapter through intent recognition, then composes detailed chapter content guided by these outlines. Finally, the system consolidates these components into a complete proposal draft. This modular architecture not only ensures logical consistency across the document but also empowers users to independently refine and optimize individual chapters. To evaluate the system's efficacy, we invited multiple researchers for assessment. Benchmarking against a single-agent LLM demonstrates that our multi-agent system produces proposals with superior content coverage, logical coherence, and user satisfaction, while significantly improving writing efficiency. The proposed modular prompt framework exhibits broader applicability and can be readily extended to other funding application contexts, offering a novel technological approach for advancing intelligent research support systems.
**Keywords:** Human–AI collaboration; Generative AI; Large Language Model; Intelligent system

## 1 INTRODUCTION

With the exponential growth of interdisciplinary research, collaborative knowledge creation has become an increasingly essential yet challenging endeavor in scientific communities. From national science foundations to international cooperation programs, researchers must synthesize diverse domain insights into cohesive, high-quality deliverables within strict space and time constraints [1]. At the same time, the number of submissions to major funding agencies worldwide has surged, while success rates continue to decline[2]. This contradiction has also prompted the writing quality of research proposals to meet almost stringent requirements. For example, researchers must accurately condense scientific problems, design rigorous technical routes, and clearly demonstrate the value and feasibility of research within a limited space[3]. However, for the traditional manual writing method, it takes an average of 38 working days for researchers to prepare a new proposal, and 28 working days for a resubmitted proposal, with an overall average time of 34 days per proposal[4]. This dual dilemma of efficiency and quality has become an important obstacle to improving scientific research productivity.

In recent years, automated writing tools based on natural language processing (NLP) have begun to emerge, but their applications are mostly limited to shallow tasks such as grammar checking and text polishing[5]. However, breakthroughs in generative AI (GenAI) have provided new possibilities for text creation. Large language models (LLMs) such as GPT-4 and Claude have demonstrated near-human-level fluency and coherence in general writing tasks, but their application in professional scientific research scenarios still faces severe challenges [6][7][8]. Tasks such as scientific proposal drafting demand not only a deep integration of fragmented ideas but also adherence to specialized discipline standards [9]. These characteristics also expose two gaps in existing AI support systems about writing in the task of generating research proposals: (1) High knowledge barriers: applicants for interdisciplinary research projects often lack an understanding of the writing standards of specific disciplines; (2) Efficiency bottlenecks: inexperienced researchers need to spend a lot of time learning the structure and expression of applications.

A variety of AI-assisted writing paradigms have attempted to address these difficulties[10][11][12]: (1) Template-based Systems (e.g., Research Rabbit, Grantable) rely on pre-designed outlines but lack substantial content-generation functionality; (2) Component-focused Systems solve local issues such as grammar checks (e.g., BERT-based GEC) or literature recommendations, rather than providing a holistic drafting process; (3) General-purpose Systems (e.g., GPT-4, Claude) can generate text across many topics but may struggle with specialized domain accuracy and advanced structural needs. Although these paradigms have made some progress, they have failed to establish a generative paradigm of the "domain-content-structure" trinity, that is, to improve creation efficiency through intelligent interaction while ensuring domain accuracy, content innovation, and structural integrity. Motivated by these gaps, our research focuses on human-AI

collaborative creation: How can we design a system that harnesses GenAI to assist users in producing complex, domain-centered documents with greater speed and consistency?

To validate this co-creation approach, we develop a GenAI support system—tentatively named Proposal AI—and test it on the domain of research proposals. We choose proposals as a prime example because they require strict structural adherence and deep disciplinary insight, offering a rigorous testbed for human–AI collaboration. Our system, however, is not confined to proposals alone; the underlying framework can be generalized to other forms of structured academic or technical writing. Specifically, we implement three core designs: (1) Adaptive Prompt Engineering: Customizing generation strategies for different disciplines and sections to enhance terminological accuracy and logical flow; (2) Human–Machine Collaborative Workflow: Enabling real-time user edits within a multi-agent division of labor, thereby optimizing content quality while providing oversight; (3) Structured Generation Framework: Dynamically creating outlines and guided content to ensure outputs conform to desired standards; users input a project name and discipline, receive a multi-tier outline generated by LLMs, then refine each chapter through specialized agents. By demonstrating these features in the proposal-writing scenario, we showcase a human–AI knowledge co-creation methodology that balances user direction with automated text generation. Our key contributions include:

(1) We propose a co-creative workflow supporting complex organizational tasks. By integrating domain cues, user interactions, and hierarchical structuring, the system helps LLMs manage the controllability challenges.

(2) We design a template-driven approach to tailor prompts. This strategy enhances LLM outputs by aligning each section with discipline requirements, ensuring greater logical consistency.

(3) We introduce a multi-agent architecture wherein different specialized agents handle various brainstorming. This division of labor significantly reduces overall drafting time, facilitating a human–AI synergy that preserves content depth and clarity.

## 2 RELATED WORK

Recent advances in artificial intelligence have given rise to diverse paradigms for AI-assisted writing, each providing partial solutions to knowledge creation tasks while revealing limitations that become critical in complex or domain-specific contexts. Although we use research proposal writing as a test scenario for our multi-agent framework, the approaches summarized here have broader applicability—and corresponding shortcomings—in tasks requiring structured, domain-aware co-creation.

### 2.1 Template-Based Systems

Template-based systems rely on pre-defined document structures to guide users in organizing content. They prioritize structural compliance over dynamic content generation, typically employing rule-based mechanisms to validate headings, word counts, and institutional formats. For instance, Mohammad et al. generated literature reviews by extracting citation sentences from academic papers[13], and Jha et al. assembled relevant text fragments to form review sections[14]. These approaches produce academic text via pattern matching and segment extraction but often yield less-cohesive passages. Sun and Zhuge further introduced a template tree to generate literature reviews recursively, organizing multi-document content via dimension and topic nodes [15]. Subsequent systems combined template methods with machine learning—for example, Liu et al. used a BERT classifier to label sentences as background, objectives, methods, and results, then concatenated these segments in a predefined order [16]. Although such strategies introduce partial automation, they essentially extend the template paradigm, which can become rigid or inadequate when the topic or input data exceed the template's scope[17]. Users may also need to inject domain expertise into template structures, raising the barrier to adoption. Furthermore, the often static nature of these templates hinders flexibility and stifles creative freedom, limiting the systems' potential for broader knowledge co-creation tasks.

### 2.2 Component-Focused Systems

Component-focused systems target specific aspects of the writing process, such as grammar correction, terminology optimization, or literature recommendation. Component-focused systems treat documents as collections of localized components rather than holistic artifacts, focusing on improving specific elements without addressing document-level problems. For example, Kaneko et al. incorporated the pre-trained language model into the encoding-decoding error correction framework and then used its output as additional features for the error correction model[18]. Omelianchuk et al. proposed to treat error correction as a sequence labeling problem, directly predicting the modification operations required for each word, thereby simplifying training and reaching a leading level[19]. These grammar checkers of component-focused Systems based on pre-trained models can already provide grammar-polishing suggestions for application writing that are close to human editing. In addition, scientific proposals require rigorous and unified terminology to avoid inappropriate wording or inconsistency. Component-focused Systems can also verify whether the terminology is used correctly and consistently by using scientific databases or domain corpora. For example, academic writing assistants (such as Writefull) helped authors check whether the manuscript covered key terms by extracting domain keywords from

academic papers[20]. In addition to the above functions, component-focused systems can also be extended to cover support for format, structure, and persuasiveness. For example, SWAN (Scientific Writing AssistaNt) is a type of component-focused system for scientific research paper writing, with built-in expert-developed indicators, which can check and provide feedback on each part of the manuscript from the title, abstract, introduction to the conclusion[21]. The feedback provided by SWAN includes marking where sentences are too long or the wording is inappropriate, and giving suggestions on how to enhance the persuasiveness of the article, such as using more powerful wording, ensuring that the title is consistent with the content, etc. Similar ideas can be applied to research proposal writing: by analyzing whether the various components of the application are complete and coordinated, reminding the author of missing information or disordered structure[22]. Some methods also use machine learning to evaluate the readability and persuasiveness of the text, helping applicants to better meet the expectations of reviewers in terms of wording and argumentation[23]. However, these functions of component-focused systems are usually still based on predefined rules or shallow NLP analysis, with limited understanding of semantics. When the author's creative ideas conflict with the built-in rules of the model, the tool's suggestions may not apply. Therefore, although component-level system improves the local quality of scientific research writing, they are independent of each other: grammar tools do not understand the meaning of the content, and terminology tools do not understand the research background. This separation limits their help to the overall writing quality and logical coherence.

### 2.3 General-Purpose Systems

General-purpose systems leverage large language models (LLMs) to generate text across a wide spectrum of topics and domains, treating document creation as an end-to-end sequence prediction task. GPT-3 notably demonstrated near-human fluency with minimal prompts in translation, question-answering, and text continuation[24], showcasing LLMs' potential for multi-task generation. Within scientific writing, Wang et al. introduced PaperRobot, which sequentially generated abstracts, conclusions, and even potential future-paper titles starting from a single paper topic [25]. Other researchers experimented with domain-specialized LLMs such as Galactica [26] to automate sections of academic writing. While these approaches display striking fluency and adaptability, their integration with discipline-specific rigor and structured control remains underexplored[27]. Text generated by general LLMs often strays from the hierarchical organization expected in formal documents (e.g., proposals), misuses technical terms, or neglects domain conventions[28]. Consequently, human creators still bear the responsibility of injecting critical scientific insights, reinforcing logic, and ensuring that proposals or other highly structured texts meet standards for academic discourse.

Subsequent work has explored adding explicit retrieval or multi-agent coordination to regain structural control. AutoSurvey divides survey-paper drafting into four LLM-mediated stages—retrieval & outline building, subsection drafting by specialized agents, integration & refinement, and automatic evaluation—thereby overcoming context-window limits and parametric-knowledge gaps when synthesising rapidly expanding literatures [29]. In the adjacent domain of English-for-Academic-Purposes (EAP) writing, AcademiCraft employs a multi-agent architecture to iteratively correct, enrich and explain revisions to scholarly prose, outperforming leading commercial grammar tools on coherence, cohesion and context-sensitive word choice [30].

Despite these advances, tensions remain between flexibility and discipline-specific rigour. Even with retrieval pipelines (as in AutoSurvey) or role-specialised agent teams (as in AcademiCraft), generated text can stray from the hierarchical organisation demanded by formal documents—misplacing methodological details, misusing technical terms, or ignoring disciplinary conventions. Human authors therefore still shoulder ultimate responsibility for injecting critical scientific insight, reinforcing logic, and ensuring compliance with community standards.

As synthesized in Table 1, existing paradigms each address different facets of structured writing. Template-based systems excel at structural control but struggle with topic diversity and creativity; component-focused solutions enhance local quality yet lack global coherence; and general-purpose LLMs provide versatility while falling short on domain precision and robust structuring. These trade-offs extend well beyond proposal drafting to any advanced knowledge-creation task, highlighting a fundamental tension between rigid templates, piecemeal enhancements, and unconstrained text generation.

**Table 1** Capability Matrix of Existing Approaches

| Paradigm | Structural Control | Content Quality | Domain Accuracy | Logical Coherence |
|---|---|---|---|---|
| Template-based | High | Medium | Low | Medium |
| Component-focused | Medium | Low | High | Low |
| General-purpose Systems | Low | High | Medium | Medium |

This capability matrix reveals the need for an integrated approach that unifies domain alignment, flexible content generation, and chapter-level organization. In the context of our study, we focus on proposal writing as a stringent, real-world application scenario for human–AI collaborative knowledge creation. Our multi-agent system, tested on proposals but

equally relevant to other complex writing tasks, aims to overcome the trilemma of structural rigidity, piecewise optimization, and generic LLM output by embedding domain constraints and structured prompts directly into the generation pipeline. As subsequent sections detail, this approach leverages funding agency (or similarly formal) templates as latent constraints while providing multi-phase prompt engineering to balance automated efficiency with scholarly precision.

## 3 METHOD

This section describes our human–AI co-creative framework for generating complex, domain-specific documents. The method is designed to support knowledge creation in various structured writing contexts requiring domain compliance, multi-chapter organization, and user oversight as shown in figure 1. The approach comprises three key phases: Dynamic Outline Generation, Personalized Refinement, and Multi-Agent Specialized Writing, each formalized with mathematical notation to ensure clarity and reproducibility. Below, we detail each phase and then summarize the overall system architecture.

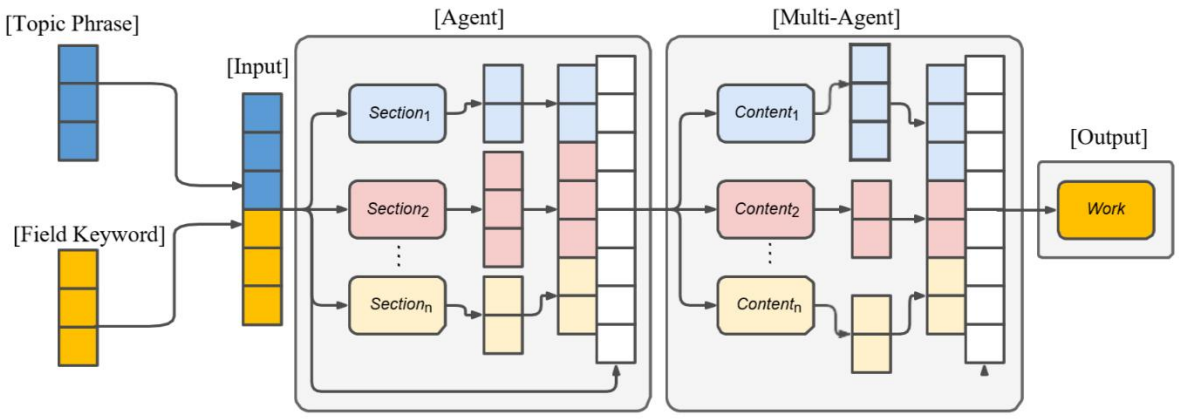


**Figure 1** Multi-Agent Method

### 3.1 Dynamic Outline Generation

Our system begins by producing an initial hierarchical outline based on user inputs: topic phrase $T$ and a field keyword $K$, where $T$ refers to an overarching subject and $K$ refers to a specific discipline, angle, or methodological focus. These inputs are applicable to various structured content scenarios (e.g., technical reports, detailed literature reviews). Given the input $T$ and $K$, the system constructs a hierarchical outline $O^{(0)}$ as follows:

**Sections**: The LLM generates a set of section titles $S_1 = \{s_1^1, s_2^1, ..., s_m^1\}$, where each $s_i^1$ corresponds to a major section of the idea.

**Subsections**: For each subsection title of sections, the LLM generates a set of subsections $S_2^i = \{s_1^2, s_2^2, ..., s_n^1\}$, where $s_j^2$ represents a subsection title under $s_i^1$.

**Content Instructions**: For each subsection $s_j^2$, the LLM generates a preliminary content instruction $c_j$, which provides a brief description of the content to be written in that subsection.

Hence, the outline generation process can be formalized as:

$$O^{(0)} = LLM(T, K) = \bigcup_{i=1}^{m} \left( s_i^1 \times \bigcup_{j=1}^{n_i} (s_j^2, c_j) \right) \tag{1}$$

where $m$ is the number of Level-1 sections, $n_i$ is the number of Level-2 subsections under $s_i^1$. This phase ensures a domain-sensitive skeletal structure, providing a coherent basis for subsequent content creation.

### 3.2 Personalized Refinement

Next, users refine the outline $O^{(0)}$ through an interactive interface, infusing domain knowledge or personal preferences. This phase allows users to modify the structure and adjust content instructions. The system ensures that user edits are seamlessly integrated into the outline while maintaining structural and logical consistency. The refinement process is driven by user edits $\Delta U$, which include:

**Structural Edits**: Adding, removing, or revising sections and subsections.

**Semantic Edits**: Rewriting content instructions $c_j$ to better reflect the user's intent.

This stage embodies human–AI collaborative knowledge creation—users integrate deep expertise into the system-generated framework. The updated outline $O^{(final)}$ is produced via:

$$O^{(final)} = LLM\left(O^{(0)}, \Delta U\right) \tag{2}$$

The final outline $O^{(final)}$ is then passed to the next phase for content generation.

### 3.3 Multi-Agent Specialized Writing

Once the refined outline $O^{(final)}$ is set, the system employs a **multi-agent** architecture $A_1 = \{a_1, a_2, ..., a_n\}$ to compose each section's text. For each section $s_i^1$ and subsection $s_j^2$ in the final outline $O^{(final)}$, the agent $a_i$ generates content through the following two designs:

**System Prompt**: The system prompt $P_s^i$ is preconfigured to ensure global or organizational guidelines

**User Prompt**: The user prompt $P_u^i$ dynamically constructed from the outline, including the section/subsection title and the custom instructions $c_j$.

The agent $a_i$ combines $P_s^i$ and $P_u^i$ to generate the final content for the section and subsection. The content generation process for section $s_i^1$ and subsection $s_j^2$ is formalized as:

$$Content_{s_i^1 or s_j^2} = a_i\left(P_s^i, P_u^i\right) \tag{3}$$

The outputs of these specialized agents are then integrated to form the complete document:

$$D = \bigcup_{i=1}^{m} \left( \text{Content}_{s_i^1} \times \bigcup_{j=1}^{n_i} \text{Content}_{s_j^2} \right) \tag{4}$$

## 4 EVALUATION

### 4.1 Experimental Setup

To evaluate the efficacy of our human–AI co-creation approach for knowledge work, we designed an experimental study using research proposal writing as the test scenario. We chose proposals because they demand structured organization, domain-specific rigor, and creative synthesis—attributes reflective of broader complex knowledge-creation tasks.

Our experiments employed Qwen-Turbo, an LLM developed by Alibaba, as the core text generator. We set the model's temperature to 0.7, aiming to balance creativity with coherence. A total of 20 researchers (15 graduate students and 5 professors) participated, representing three diverse academic fields: Management Science & Engineering, Economics, and Computer Science. This disciplinary spread enabled us to observe how different knowledge domains interact with the system's multi-agent design. Before the evaluation began, each participant received a short training session on how to operate our co-creative platform.

We adopted a multi-faceted evaluation methodology to rigorously test how well our multi-agent co-creation framework supports users in a demanding knowledge-creation scenario—namely, writing an extended research proposal. The method combined a user study, which collected subjective feedback through questionnaires, and a baseline comparison against a more traditional single-agent LLM.

Each of the 20 participants used our co-creative system to draft a full research proposal aligned with their expertise or area of interest. When the system finished generating a draft, participants reviewed the content, focusing on key aspects such as logical structure, domain accuracy, and originality. Once a draft proposal was generated, participants reviewed the output. They were then asked to fill out a detailed questionnaire evaluating the system and the generated proposal. The questionnaire captured the participants' ratings on various aspects of the text quality (fluency, coherence, readability), the structural integrity of the proposal, and the originality of the content. It also included items on the usability of the system (how easy and intuitive it was to use) and overall satisfaction with the experience. Participants completed the questionnaire immediately after using the system, ensuring their feedback was based on their fresh experience. This user-centric evaluation allowed us to gather insights into how well the system meets the needs of researchers in practice and how comfortable they are with the automatically generated content.

To benchmark the effectiveness of a multi-agent, structured approach, we compared our system against a single-agent method (e.g., ChatGPT4 and DeepResearch) with straightforward prompts. After finishing their interaction with our co-creative platform, each participant examined a ChatGPT-generated draft on the same project. They then rated that version using the same questionnaire items, enabling direct comparisons across metrics such as structural completeness, coherence, or user satisfaction.

### 4.2 Evaluation Metrics

We defined a set of evaluation metrics covering both the quality of the generated text and the user experience of the system. These metrics together provide a comprehensive evaluation of the system's performance, balancing output quality and process quality. The key evaluation metrics include:

**Quality**: Measures the linguistic and narrative quality of the proposal [31]. This includes fluency (naturalness of language and absence of grammatical errors), coherence (logical consistency and flow of ideas throughout the proposal), and readability (clarity and ease of understanding for the reader).

**Completeness**: Assesses the organizational quality of the proposal [32]. We evaluate completeness in terms of whether all essential sections of a standard research proposal are present and logical structure, meaning the content is well-organized with a clear progression of ideas and a sound argument structure.

**Innovativeness**: Evaluates the originality and creativity of the content[33]. This reflects whether the proposed research ideas and approaches appear novel. All participants and experts considered if the automatically generated proposal offers fresh insights or interesting research directions.

**Usability**: Captures the ease of use and user-friendliness of the system[34]. This metric reflects the user experience: how easy it was for participants to interact with the system, understand its prompts, and steer the generation process. It also covers the satisfaction of users with the system's interface and functions.

**Efficiency**: Measures the time and effort saved by using the system[35]. We looked at two aspects: time to generate (how quickly a complete draft was produced by the system, from the moment the user input the initial idea to the time the final draft was ready), and time to revise (how much additional time the participant needed to revise or polish the AI-generated draft to reach a submission-ready state). This metric is important to gauge whether the system actually speeds up the proposal writing process compared to writing a proposal manually or using simpler tools.

Although tested here on proposals, these metrics reflect broader dimensions of knowledge work—evaluating both content quality and user interaction in high-complexity writing tasks.

### 4.3 Results And Analysis

We collected questionnaire data from 20 participants who each used two systems: our ProposalAI multi-agent platform and baselines (ChatGPT4 and DeepResearch). While proposals anchored the experiment, the findings illustrate how structured co-creation compares to a more generic LLM workflow in terms of coverage, creativity, and user experience.

Figure 2 shows the average scores for Quality, Completeness, Innovativeness, Usability, and Efficiency. Across all metrics, ProposalAI outperforms the baseline, reflecting the effectiveness of incremental outline-building, section-by-section refinement, and domain-oriented prompts.

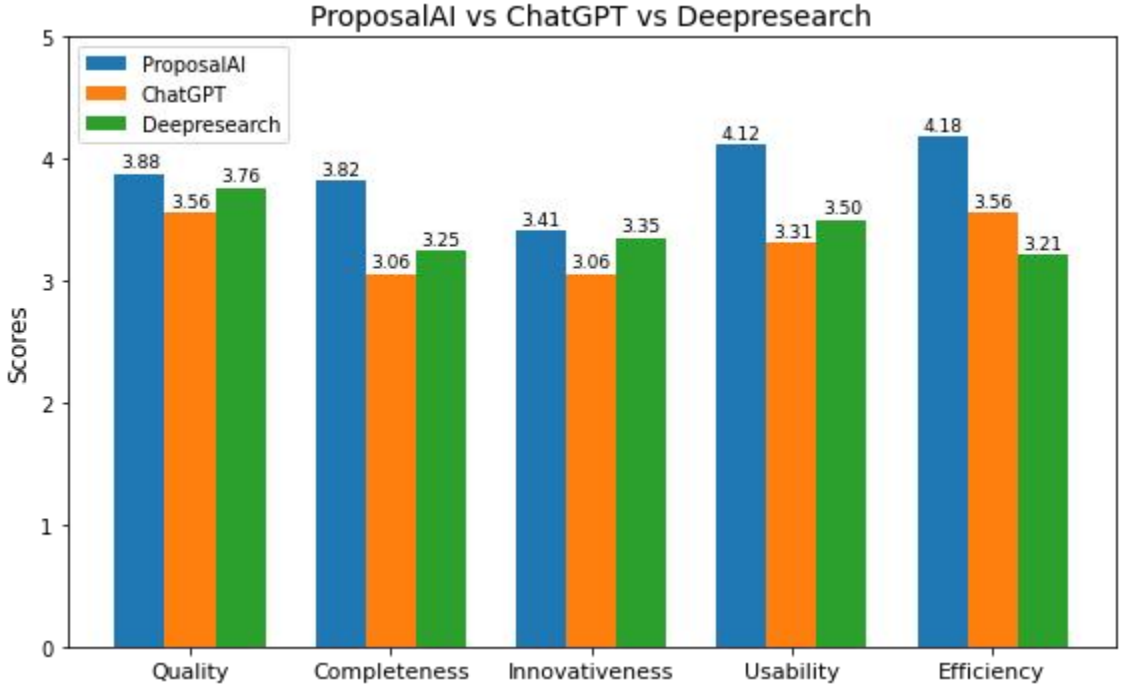


**Figure 2** The Evaluation Results

#### *4.3.1 Quality*

ProposalAI again tops the chart (3.88). DeepResearch (3.76) narrows the gap to ProposalAI and outperforms ChatGPT (3.56). Participants attributed DeepResearch's edge over ChatGPT to its explanation-driven corrections, yet noted that ProposalAI's section-specific prompting still produced the clearest narrative flow. This difference reflects two primary factors:

**Fluency and Coherence**: Users commented that, because ProposalAI breaks down each section with domain-specific or section-specific prompts, the model can maintain a better overall flow. In contrast, ChatGPT sometimes introduced minor redundancy or abrupt shifts when attempting to handle everything in one prompt.

**Readability**: Even though both systems occasionally used generic phrasing, participants said ProposalAI's text was typically clearer and more to-the-point. This is likely due to the multi-phase approach, which reduces the chance of meandering or repeating irrelevant content.

Interestingly, a subset of participants with extensive proposal-writing experience rated both systems lower on language sophistication, indicating that extremely specialized or discipline-specific styles may require more refined prompts or domain-tailored lexical knowledge. Overall, the difference in Quality suggests that scaffolding the writing process via multiple agents effectively boosts clarity and flow, although further fine-tuning or domain adaptation might be needed to reach near-expert human writing levels in specialized fields.

#### *4.3.2 Completeness*

Completeness exhibited the widest performance gap: ProposalAI (3.82) markedly outperformed DeepResearch (3.25) and ChatGPT (3.06). Three observations explain this pattern:

**Outline Enforcement**: ProposalAI's comprehensive outline reduced omissions. DeepResearch, lacking mandatory outline checks, sometimes omitted auxiliary elements but still covered more sections than ChatGPT. Because the system begins by generating and refining a comprehensive outline with mandatory sections (e.g., introduction, literature review, methods, anticipated results), participants were less likely to overlook crucial parts of a standard proposal. This structural scaffolding makes omissions less probable.

**Progressive Detailing**: The multi-agent approach iteratively fills in sections, ensuring each receives adequate attention. By contrast, ChatGPT's single-pass generation sometimes skipped or glossed over essential content (e.g., feasibility analysis or budget justification) if the prompt was not explicit enough.

**Participant Guidance**: The outline phase encouraged participants to add custom sub-sections or expand on domain-specific concerns (e.g., "Ethical Considerations"), thus pushing the final output toward greater comprehensiveness.

Qualitative comments from participants confirm that the clarity and forced coverage of the Outline Generation step were key reasons for higher Completeness scores. Several participants mentioned feeling more confident that "nothing important was missing." In short, a well-structured, multi-step pipeline appears crucial for producing thorough research proposals that meet standard academic or funding agency expectations.

#### *4.3.3 Innovativeness*

ProposalAI's Innovativeness score (3.41) modestly surpasses DeepResearch (3.35) and ChatGPT (3.06). Our analysis suggests two main factors:

**Opportunity for Customization**: By prompting users to refine the outline and letting them inject new ideas at different stages, ProposalAI can incorporate domain insights that lead to less generic or formulaic text. DeepResearch or ChatGPT's single prompt sometimes defaulted to "safe" or "boilerplate" suggestions.

**Limitations of AI Creativity**: Despite the advantage, participants generally felt that neither system truly substitutes for a researcher's unique intellectual contribution. Users with advanced domain knowledge noted that if they only provided cursory instructions, the content would still be somewhat formulaic. This indicates that while multi-agent scaffolding can facilitate creative thinking, genuine scientific innovation still heavily depends on the user's active input.

Multiple participants also pointed out that the system can only reassemble known concepts. This feedback implies that further improvements (e.g., deeper domain integration, synergy with recent literature or specialized databases) could yield even more innovative proposals.

#### *4.3.4 Usability*

Usability scores reveal ProposalAI (4.12) > DeepResearch (3.50) > ChatGPT (3.31). From user feedback, we identified several design elements that contributed to higher satisfaction:

**Stepwise Interaction**: Rather than having everything happen in one large output, participants found the multi-phase approach more transparent and manageable. This process "felt natural," mirroring the mental steps of outlining, drafting, and refining.

**Fine-Grained Control**: Breaking the proposal into sections allowed users to intervene more precisely. They could refine each part (e.g., method, background) without risking the rest of the text. ChatGPT's single-shot approach often needed repeated re-prompts to fix local issues in one section without inadvertently modifying correct parts elsewhere. DeepResearch improved over ChatGPT yet lacked ProposalAI's slot-specific guidance.

**Guided Prompts**: Clear instructions for each section (e.g., "research objectives," "anticipated results") gave participants confidence they were focusing on the right aspects. Some praised the interface for "not letting me forget a key element."

Nevertheless, about one-fifth of participants wished for even more direct integration of external documents (e.g., references, prior proposals) into the multi-agent workflow. They felt such a feature would further streamline the writing process.

Overall, high Usability scores reaffirm that a well-structured interface and guided prompts can significantly enhance user experience beyond what a general-purpose LLM alone can offer.

#### *4.3.5 Innovativeness*

Efficiency likewise favored ProposalAI (4.18). ChatGPT (3.56) slightly exceeded DeepResearch (3.21) because DeepResearch's explanatory cycle added turnaround time. Users reported noticeable reductions in:

**Initial Draft Time**: Thanks to the forced outline stage, many participants stated that they overcame "writer's block" quickly. The system auto-populated a skeleton with relevant subheadings, so participants did not have to figure out organizational flow from scratch. DeepResearch's explanations lengthened iteration time despite yielding clearer text than ChatGPT.

**Revision Effort**: Because the generated drafts tended to be more logically organized and complete, the subsequent editing or fine-tuning stage required less time. In the ChatGPT condition, participants indicated they often had to revisit the prompt multiple times and manually add missing sections, leading to more iterative overhead.

In some open-ended responses, participants acknowledged that if a proposal was highly specialized or if they needed extensive references to advanced literature, the AI needed more custom prompts or expansions. However, even in these scenarios, the foundational structure and partial content still saved them time relative to starting with a blank document or a one-shot generation from ChatGPT. Overall, the multi-agent design gave them fewer "back-and-forth loops," thereby boosting perceived efficiency.

## 5 SUMMARY AND CONCLUSION

Overall, the multi-agent system scored notably higher in Completeness, Usability, and Efficiency—key dimensions for any collaborative knowledge work. Although improvements in innovativeness and deeper domain adaptation remain open areas for future research, our study underscores the promise of dividing content generation across structured prompts and specialized agents.

In conclusion, the user study suggests that ProposalAI outperforms baselines in most critical dimensions of research proposal writing. Notably, the largest gains appear in Completeness (thanks to structured outlines), Usability (stepwise guidance and better user control), and Efficiency (less revision time). While there is still room for growth in fostering truly novel content, the multi-agent methodology evidently provides a more reliable framework for generating comprehensive, coherent proposals. This underscores the importance of combining domain-specific structuring with user-driven refinement when applying large language models to specialized tasks like grant writing. DeepResearch consistently surpasses ChatGPT in Quality, Completeness, Innovativeness and Usability, confirming the benefit of explanation-oriented agent collaboration, though its longer feedback loop reduces perceived efficiency. These results underscore the efficacy of template-aware, multi-phase prompting for complex scholarly writing. Future work should explore combining DeepResearch-style explanatory feedback with ProposalAI's structural scaffolding, alongside deeper domain adaptation and automated citation support.

Future enhancements also might involve deeper integration with domain knowledge bases, real-time citation management, or iterative refinement loops that incorporate external critiques. Nonetheless, current findings validate that the multi-agent approach holds promise for improving research proposal drafting, potentially reducing the cognitive overhead and time investment traditionally associated with this complex writing task.

## 6 DISCUSSION AND FUTURE WORK

While the multi-agent architecture demonstrably improves structural guidance and user control, the data layer remains a bottleneck in three respects. First, the system relies exclusively on the frozen parametric knowledge of the underlying LLM; it cannot query up-to-date bibliographic databases or domain-specific corpora during generation. As a result, references must be inserted manually, and citation accuracy is susceptible to hallucination. Second, all empirical results were obtained on a single, moderately sized evaluation set (n = 20 proposals) drawn from three academic fields. This dataset is insufficient to capture the full diversity of research domains, styles, and disciplinary conventions, limiting external validity. Future work will address both issues by (i) integrating authenticated retrieval modules for real-time access to scholarly indices and domain-specific databases, and (ii) conducting a multi-institutional study with a substantially larger and more diverse participant pool encompassing senior academics, industry practitioners, and non-English speakers. Such expansions will enable finer-grained error analysis, strengthen external validity, and inform domain-tailored prompt engineering.

## COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

## REFERENCES

[1] Lindgreen A, Di Benedetto CA, Verdich C, et al. How to write really good research funding applications. Industrial Marketing Management, 2019, 77: 232-239.

[2] Ren R, Ma J, Zheng Z. Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. Expert Systems with Applications, 2025, 278: 127330.
[3] Locke LF, Spirduso WW, Silverman SJ. Proposals that work: A guide for planning dissertations and grant proposals. Sage Publications, 2013.
[4] Herbert DL, Barnett AG, Clarke P, et al. On the time spent preparing grant proposals: an observational study of Australian researchers. BMJ Open, 2013, 3(5): e002800.
[5] Russo F. Automated content writing tools and the question of objectivity. Digital Society, 2023, 2(3): 50.
[6] Wu J, Yang S, Zhan R, et al. A survey on LLM-generated text detection: Necessity, methods, and future directions. Computational Linguistics, 2025: 1-66.
[7] Ren R, Ma J, Luo J. Large language model for patent concept generation. Advanced Engineering Informatics, 2025, 65: 103301.
[8] Wang Y, Guo Q, Yao W, et al. Autosurvey: Large language models can automatically write surveys. Advances in Neural Information Processing Systems, 2024, 37: 115119-115145.
[9] Kallet RH. How to write the methods section of a research paper. Respiratory Care, 2004, 49(10): 1229-1232.
[10] Sharma R, Gulati S, Kaur A, et al. Research discovery and visualization using ResearchRabbit: A use case of AI in libraries. COLLNET Journal of Scientometrics and Information Management, 2022, 16(2): 215-237.
[11] Zhou YC, Zheng Z, Lin JR, et al. Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. Computers in Industry, 2022, 142: 103746.
[12] Pal S, Bhattacharya M, Islam MA, et al. AI-enabled ChatGPT or LLM: A new algorithm is required for plagiarism-free scientific writing. International Journal of Surgery, 2024, 110(2): 1329-1330.
[13] Mohammad S, Dorr B, Egan M, et al. Using citations to generate surveys of scientific paradigms. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, June: 584-592.
[14] Jha R, Finegan-Dollak C, King B, et al. Content models for survey generation: A factoid-based evaluation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, July: 441-450.
[15] Sun X, Zhuge H. Automatic generation of survey paper based on template tree. 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), 2019, September: 89-96.
[16] Liu S, Cao J, Yang R, Wen Z. Generating a structured summary of numerous academic papers: Dataset and method. International Joint Conferences on Artificial Intelligence, 2022.
[17] Zhu K, Feng X, Feng X, et al. Hierarchical catalogue generation for literature review: A benchmark. Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, December: 6790-6804.
[18] Kaneko M, Mita M, Kiyono S, et al. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. arXiv preprint, arXiv:2005.00987, 2020.
[19] Omelianchuk K, Atrasevych V, Chernodub A, et al. GECToR–Grammatical error correction: Tag, not rewrite. Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2020, July: 163-170.
[20] Mitchell P, Riedlinger M, Goldenfein J, et al. Research GenAI: Situating generative AI in the scholarly economy. AoIR Selected Papers of Internet Research, 2024.
[21] Kinnunen T, Leisma H, Machunik M, et al. SWAN-scientific writing AssistaNt: A tool for helping scholars to write reader-friendly manuscripts. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, April: 20-24.
[22] Adhi P. Exploring the use of ChatGPT as a supporting tool in writing research proposals: EFL students' perspectives. Doctoral dissertation, UIN Sunan Gunung Djati Bandung, 2024.
[23] Bai X, Stede M. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. International Journal of Artificial Intelligence in Education, 2023, 33(4): 992-1030.
[24] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
[25] Wang Q, Huang L, Jiang Z, et al. PaperRobot: Incremental draft generation of scientific ideas. arXiv preprint, arXiv:1905.07870, 2019.
[26] Taylor R, Kardas M, Cucurull G, et al. Galactica: A large language model for science. arXiv preprint, arXiv:2211.09085, 2022.
[27] Huang J, Tan M. The role of ChatGPT in scientific communication: Writing better scientific review articles. American Journal of Cancer Research, 2023, 13(4): 1148.
[28] Seckel E, Stephens BY, Rodriguez F. Ten simple rules to leverage large language models for getting grants. PLOS Computational Biology, 2024, 20(3): e1011863.
[29] Wang Y, Guo Q, Yao W, et al. Autosurvey: Large language models can automatically write surveys. Advances in Neural Information Processing Systems, 2024, 37: 115119-115145.

[30] Du Z, Hashimoto K. AcademiCraft: Transforming writing assistance for English for academic purposes with multi-agent system innovations. Information, 2025, 16(4).
[31] Daudaravicius V. Automated evaluation of scientific writing: AESW shared task proposal. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, June: 56-63.
[32] Cai Y, Ziad M. Evaluating completeness of an information product. AMCIS 2003 Proceedings, 2003, 294.
[33] Dean DL, Hender J, Rodgers T, et al. Identifying good ideas: Constructs and scales for idea evaluation. Journal of Association for Information Systems, 2006, 7(10): 646-699.
[34] Davis FD. Technology Acceptance Model: TAM. In: Al-Suqri MN, Al-Aufi AS, eds. Information Seeking Behavior and Technology Adoption. Hershey, PA: IGI Global; 2015: 205–219.
[35] Michailidis A, Rada R, Gouma P. A study of efficiency in computer-supported collaborative writing. Journal of Intelligent Systems, 1994, 4(1-2): 133-162.