

A Comparative Study of Multi-Agent Collaboration Structures for On-Premises LLM-Based Technical Document Generation

Jinkyung Kwon^a, Hogeon Seo^{a, b*}

^aKorea Atomic Energy Research Institute (KAERI), Daejeon, Republic of Korea

^bUniversity of Science and Technology (UST), Daejeon, Republic of Korea

E-mail: kwon0312@kaeri.re.kr

*Corresponding author: hogeony@hogeony.com

***Keywords :** On-premises LLM, Multi-agent system, Technical document generation, Controlled comparison, LLM evaluation

1. Introduction

Researchers prepare technical disclosure documents describing their inventions and submit them to patent attorneys for formal patent applications. However, the source material often consists of fragmented idea notes rather than systematically organized documents. Although large language models (LLMs) can reduce this documentation burden, connecting confidential invention information to external servers raises security concerns. Therefore, technical documents need to be generated in an on-premises environment, despite potential limitations in model performance and diversity.

To address this issue, this study compared four AI-agent collaboration structures with a non-agent baseline while applying the same document template to all conditions. The input dataset consisted of ten drafts drawn from the company's internal technical materials. Documents were generated from these drafts using the Gemma-4-31B model, which is also used by the company's internal chatbot. Five documents were generated from each input draft under each generation structure. Internal evaluation and review during the generation process were separated from the independent evaluation of the final documents, and both document quality and repeatability robustness were analyzed.

2. Comparative Framework

2.1 Common Task and Baseline

The same idea draft was provided under all experimental conditions. Each output document consisted of the invention title, prior art and related problems, objectives, configuration and operation, specific embodiments, and effects. The prompts, section divisions, document assembly process, and format validation rules were derived from a single common template definition. Test0 served as the baseline, generating documents from only the idea draft and template without agent roles or a collaboration process. The effectiveness of the collaboration processes was evaluated by comparing their results with this baseline.

2.2 Collaboration Structures

Test1 adopted the section-by-section drafting structure of AutoPatent [1]. A planner established the document tree according to the common template, a writer drafted each section using a fixed reference set, and a reviewer requested revisions to individual sections. The structural format was fixed in the code to prevent the model from altering or damaging it.

Test2 was inspired by Self-Refine [2]. A coordinator assigned the task, and a writer created or revised the entire document. A reviewer then provided information about defects that could be directly used in the next revision.

In Test3, two independent proposers drafted separate documents. Neither proposer could view the other's output, and only the document prepared by the synthesizer was scored and retained. This structure was based on an isolated-proposal variant of Mixture-of-Agents[3] using the same model. It was not intended to reproduce the effects of the original study, which combined different models.

Test4 was adapted from the independent generation and cross-referencing revision procedure used in the multi-agent debate framework proposed by Du et al. [4] for narrative technical document generation. As in Test3, two proposers were employed. In the first round, each proposer independently generated a draft, while in subsequent rounds, each proposer revised its own output by referring to the other proposer's draft from the preceding round. Because the majority-voting scheme used in the original study is difficult to apply to narrative documents, a separate Consensus Agent was introduced to construct the final document by retaining claims supported by the input data and selecting alternatives with sufficient evidence. Disagreements that could not be supported by the input data were not resolved arbitrarily but were explicitly identified in the document as insufficiently supported content.

Table I summarizes the five document generation architectures used in the comparative experiments and the key generation procedures of each architecture. Test0 serves as the baseline condition without agent collaboration, while Test1 through Test4 employ split writing, iterative revision, parallel proposal, and

discussion-based consensus architectures, respectively.

Table I: Comparison of Document Generation Structures

Condition	Core Procedure
Test0	Single-model generation without an agent structure
Test1	Tree planning and section-by-section drafting
Test2	Iterative revision by a coordinator, writer, and reviewer
Test3	Two isolated parallel proposals followed by synthesis
Test4	Two-proposer cross-review and grounded consensus

3. Experimental Design

3.1 Input Drafts and Repeated Generation

In the document generation experiments, Gemma-4-31B was accessed through an external API. The experiments were conducted as a preliminary evaluation to compare collaboration structures for future use in the technical document generation functionality of the institution’s internal chatbot. Ten drafts were used, each representing a different invention. The same draft content was provided to all five generation structures. Five successful outputs were collected for each combination of draft and structure, resulting in a total of 250 documents ($10 \times 5 \times 5$).

The target internal chatbot operates on an isolated network with no external connectivity, and Gemma-4-31B is served on a server equipped with two NVIDIA A100 80GB GPUs. This hardware configuration describes the current on-premises deployment and does not represent the server specifications used for the external API-based experiments reported in this paper.

The input draft was treated as the unit of analysis. The five outputs obtained from each draft–structure combination were used to measure variability across repeated executions of the entire pipeline. Treating all 250 outputs as independent samples would result in pseudoreplication. Therefore, statistical tests were conducted using the mean of the five repetitions for each draft–structure combination.

3.2 Independent Evaluation and Statistical Methods

The quality of the generated documents was evaluated using OpenAI GPT-5.1, accessed through an external API, as an independent LLM judge separate from the Gemma-4-31B model used for document generation. The LLM judge assessed whether each evaluation criterion was satisfied using item-level binary judgments, which were then converted by the evaluation code into scores for four dimensions and an overall score. The item-level binary evaluation approach was inspired by CheckEval [5].

The four evaluation dimensions were as follows: (1) Logical coherence assessed the relationships among the problem, objectives, configuration, and effects; (2) Structural consistency assessed content organization, unnecessary redundancy, and the evidential basis for stated effects; (3) Draft fidelity assessed whether the input draft was accurately reflected and whether unsupported information was introduced; and (4) Technical specificity assessed whether the key components and operations could be specifically identified from the document.

Because the same ten drafts were repeatedly evaluated under all generation structures, differences among the structures were examined using the Friedman test. Given the small sample size and the concentration of scores near the upper bound, p-values were calculated using a permutation-based approach. For post-hoc comparisons, a sign-flip permutation test of paired mean differences was used, with Holm correction for multiple comparisons. Variability across repeated executions was evaluated as the difference between the maximum and minimum valid scores collected for the same draft–structure combination.

3.3 Technical Value Evaluation

Document-based technology and intellectual property potential was evaluated using OpenAI GPT-5.1 and the company’s internal technology valuation criteria. The original evaluation framework consists of a total of 100 points: 40 points for technology, 25 points for intellectual property rights, and 35 points for marketability. However, because no separate market, commercialization, or prior-art data were provided in this experiment, the evaluation focused on the technology and intellectual property criteria that could be assessed, to a limited extent, from the generated documents. The technology dimension included innovativeness, applicability, and difficulty of imitation, while the intellectual property dimension included distinctiveness, strength of rights protection, and ease of detecting and proving infringement. Technology substitutability, intensity of technological competition, and marketability were excluded from the final score because their assessment requires external information.

The maximum possible score for the included criteria was 55 points, and the obtained score was divided by the maximum score of the criteria actually evaluated and converted to a 100-point scale. However, even the included criteria were assessed solely on the basis of information described in the generated documents, without external data on technology trends or prior patents. Therefore, these scores should not be interpreted as definitive measurements of actual technology value, but rather as limited estimates of the technology and intellectual property potential reflected in the documents.

4. Results

4.1 Document Quality

Document quality differed significantly across the generation structures ($p = .011$). Test4 achieved the highest mean score of 94.38 and was significantly better than Test1 in the post-hoc comparison (adjusted $p = .039$). However, no statistically significant difference was found between Test4 and Test3.

Figure 1 shows the distribution of the average Judge scores for the drafts obtained under each generation architecture. Each point represents the average score across repeated runs for a single input draft, while the gray lines indicate how the score of the same draft changes across different generation architectures. Test3 and Test4 show relatively high score distributions, whereas in Test0 and Test1, some drafts received substantially lower scores than others. The boxes represent the interquartile range from the first quartile (Q1) to the third quartile (Q3), and the line inside each Box indicates the median.

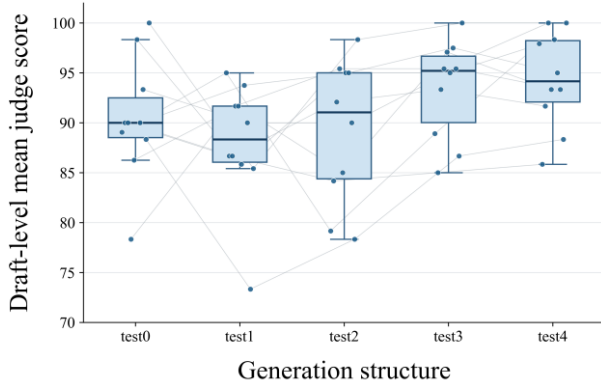


Fig. 1. Distribution of the mean Judge scores by generation structure. Each point represents one of the ten drafts, averaged over five repeated runs under each structure.

4.2 Analysis by Evaluation Dimension

Among the evaluation dimensions, only structural consistency differed significantly across the generation structures ($p < .001$), with Test1 showing the lowest mean score. Draft fidelity was higher for Test3 and Test4, but the difference was not statistically significant. Rather than demonstrating a dominant advantage in a particular dimension, Test4 produced balanced results without a major weakness across any of the evaluation dimensions.

Figure 2 shows the average scores of each generation architecture across four evaluation dimensions. Logical coherence and technical specificity were relatively high across most architectures, whereas larger differences were observed in structural consistency and draft fidelity. In particular, Test1 showed relatively low structural consistency, while Test3 and Test4 achieved high scores in draft fidelity.

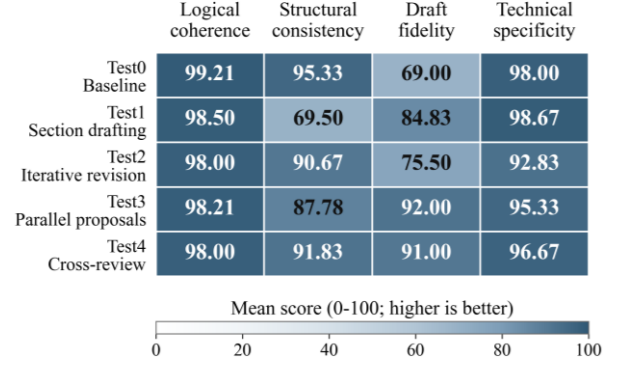


Fig. 2. Heatmap of the mean evaluation-dimension scores by generation structure. Each cell represents the mean across ten drafts, after averaging the five repeated runs for each draft. All evaluation dimensions use the same 0–100 scale, with darker colors indicating higher mean scores. Test0–Test4 correspond to the generation structures defined in Table I.

4.3 Repeatability Robustness

Repeatability robustness was evaluated using the difference between the maximum and minimum scores obtained from five repeated runs using the same draft. Test4 showed the smallest mean repetition range at 13.96 points, followed by Test3 at 16.80 points. Although the overall difference in repetition range across the generation structures was statistically significant ($p = .009$), the post-hoc tests did not identify any significant pairwise differences. Therefore, Test4 showed the most stable tendency, but it cannot be concluded that it was statistically more robust than the other structures.

Table II presents the mean and standard deviation of document quality scores for each generation architecture, along with the range across repeated runs. Test4 achieved the highest mean Judge score and the smallest run-to-run range, followed by Test3. In contrast, Test1 and Test2 showed relatively large run-to-run ranges, indicating greater score variability across repeated executions even when the same draft was used.

Table II: Document Quality and Repeatability Robustness

Condition	Judge Mean ± SD	Mean Repetition Range
Test0	90.36±6.08	23.12
Test1	88.00±6.17	27.50
Test2	89.25±7.15	27.71
Test3	93.43±4.95	16.80
Test4	94.38±4.84	13.96

4.4 Technical Value Evaluation

As shown in Table III, the document-based technology value scores were similar across all generation architectures, ranging from 77.61 to 78.70, with no statistically significant differences among the

architectures ($p = .997$). In contrast, differences were observed in document quality, suggesting that the effect of the generation architecture was more evident in the quality of document presentation than in the assessed technology value.

Table III: Technical value scores by generation structure

Condition	Technical Value Mean $\pm$ SD
Test0	78.70 ± 7.12
Test1	78.17 ± 7.96
Test2	77.61 ± 7.31
Test3	78.41 ± 6.43
Test4	78.33 ± 7.58

5. Discussion

Test4 achieved the highest mean document-quality score and the smallest mean repetition range. Test3 showed a similar level of document quality and the second-smallest repetition range. Across the evaluation dimensions, the main differences were the low structural consistency of Test1 and the high draft fidelity observed for Test3 and Test4. The high overall score of Test4 appears to reflect consistently strong performance in both structural consistency and draft fidelity rather than an exceptional score in a single evaluation dimension. The only statistically significant pairwise difference identified in the post-hoc tests was between Test4 and Test1.

This experiment used the same ten drafts drawn from actual internal technical materials, thereby controlling for differences among input drafts. However, both the preparation method and the technical domains of the input documents were limited. It remains uncertain whether the same results would be observed for other types of technical documents. In addition, the document-quality scores and repetition ranges reflect variability from both the generation model and the LLM judge. Further validation using a wider range of real-world technical documents and evaluations by domain experts is therefore required.

A primary security measure in an on-premises LLM environment is network isolation from external networks. The target chatbot performs model inference within the institution's internal network, which has no external connectivity, allowing technical materials and generated documents to be processed internally without transmission to external LLM services. This configuration prevents information exposure associated with transmitting data to external services and restricts access paths through external networks. However, addressing unauthorized internal access and threats introduced through imported files also requires user authentication and access control, protection of documents and logs, and integrity verification of models and software brought into the internal network.

6. Conclusions

Using ten internal technical drafts as inputs, this study analyzed documents generated by Gemma-4-31B under five generation structures. Document quality and repetition range varied across the generation structures. Test4 achieved the highest mean document-quality score and the smallest repetition range, while Test3 showed similar document quality and the second-smallest repetition range. However, after adjustment for multiple comparisons, a statistically significant difference was identified only between Test4 and Test1; no significant difference was found between Test3 and Test4.

No significant differences among the structures were found in the document-based technical value scores. Within the scope of this experiment, differences in generation structure were more apparent in document presentation quality and repeatability than in technical value judgments. Nevertheless, the limited number and technical scope of the drafts, as well as the reliance on an LLM judge, prevent Test4 from being generalized as a universally superior structure. Further validation using diverse real-world technical documents and expert evaluation is needed.

ACKNOWLEDGEMENTS

This work was supported by the Institute Strategic Development Project (ISD) funded by the Ministry of Science and ICT (No. 526140-26), the Substantiation Support Program through the Korea Innovation Foundation, funded by the Ministry of Science and ICT, and the Program for Strengthening Technological Capabilities in Nondestructive Testing through the National Research Foundation of Korea (No. RS-2026-25545949), funded by the Ministry of Science and ICT.

REFERENCES

- [1] Q. Wang et al., AutoPatent: A Multi-Agent Framework for Automatic Patent Generation, arXiv:2412.09796, 2024.
- [2] A. Madaan et al., Self-Refine: Iterative Refinement with Self-Feedback, *Advances in Neural Information Processing Systems*, Vol. 36, 2023.
- [3] J. Wang et al., Mixture-of-Agents Enhances Large Language Model Capabilities, *International Conference on Learning Representations*, 2025.
- [4] Y. Du et al., Improving Factuality and Reasoning in Language Models through Multiagent Debate, *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235, pp. 11733–11763, 2024.
- [5] Y. Lee et al., CheckEval: A Reliable LLM-as-a-Judge Framework for Evaluating Text Generation Using Checklists, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798, 2025.