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

Presenter

Jinkyung Kwon

Author

Jinkyung Kwon [a], Hogeon Seo [a,b*]

[a]KAERI · [b]UST

Contents

1. Introduction

2. Comparative framework

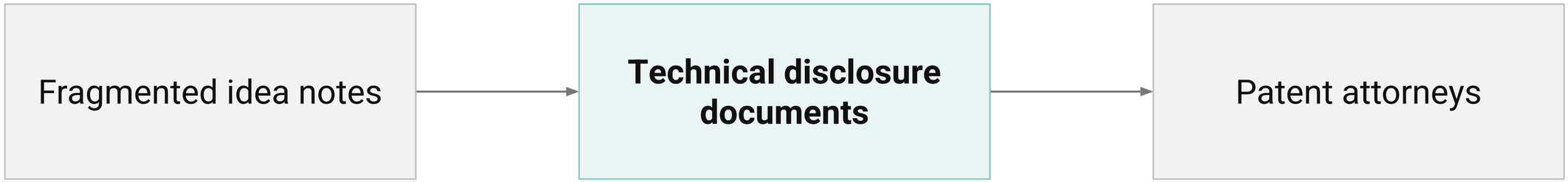
3. Experimental design

4. Results

5. Discussion

6. Conclusions

Introduction



LLMs (Large Language Models)

Documentation burden ↓

Security concerns

Confidential invention information → External servers

On-premises environment

Potential limitations in model performance and diversity

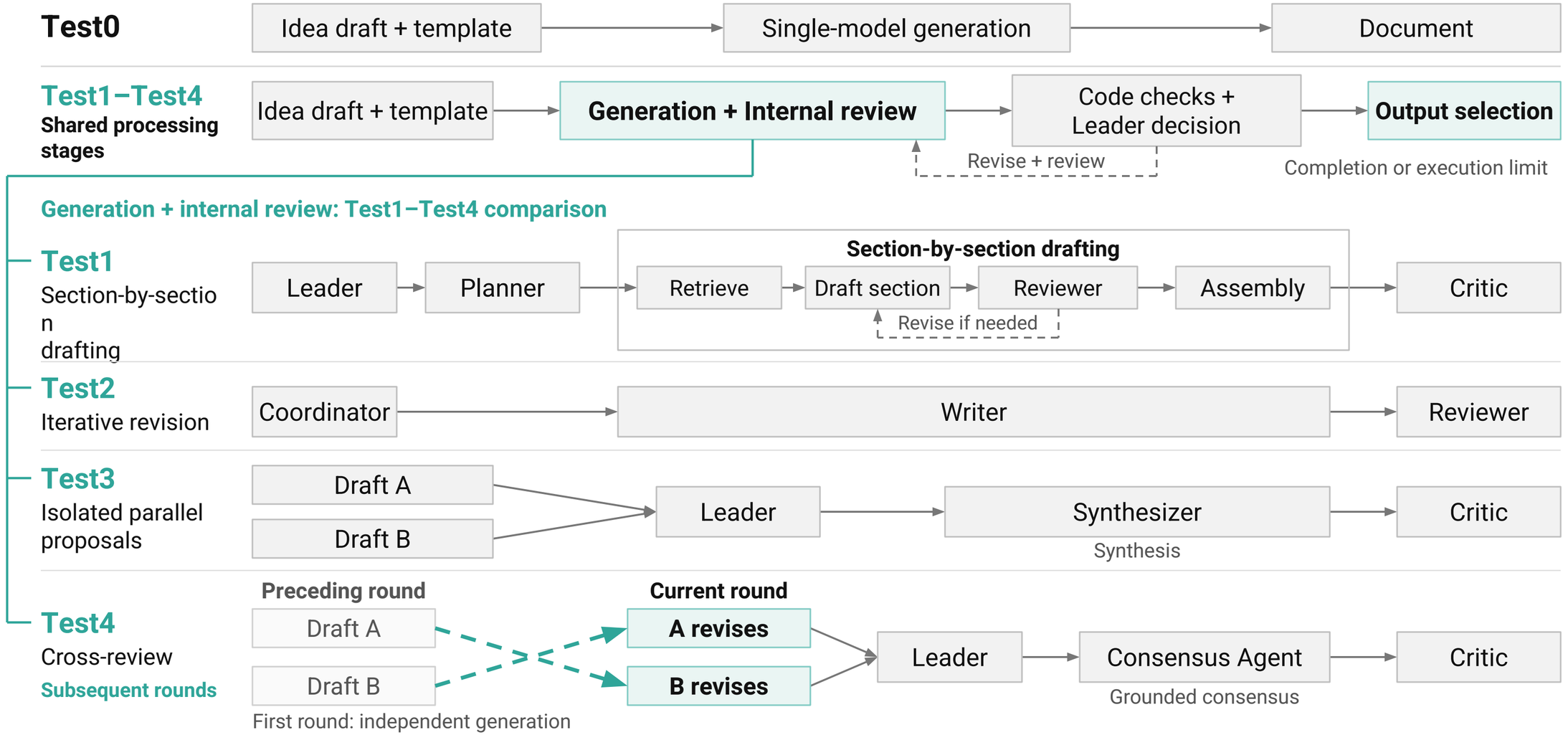
Controlled comparison

Effectiveness of the collaboration processes

Non-Agent baseline vs **Agent collaboration**

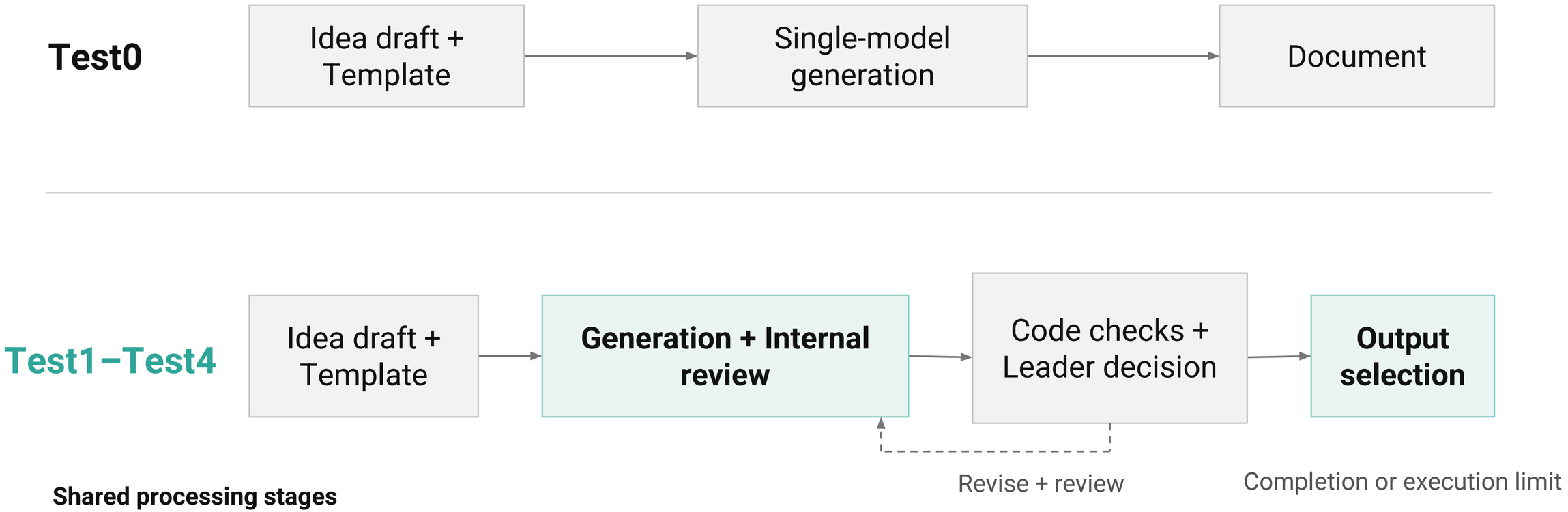
Document quality / Repeatability robustness

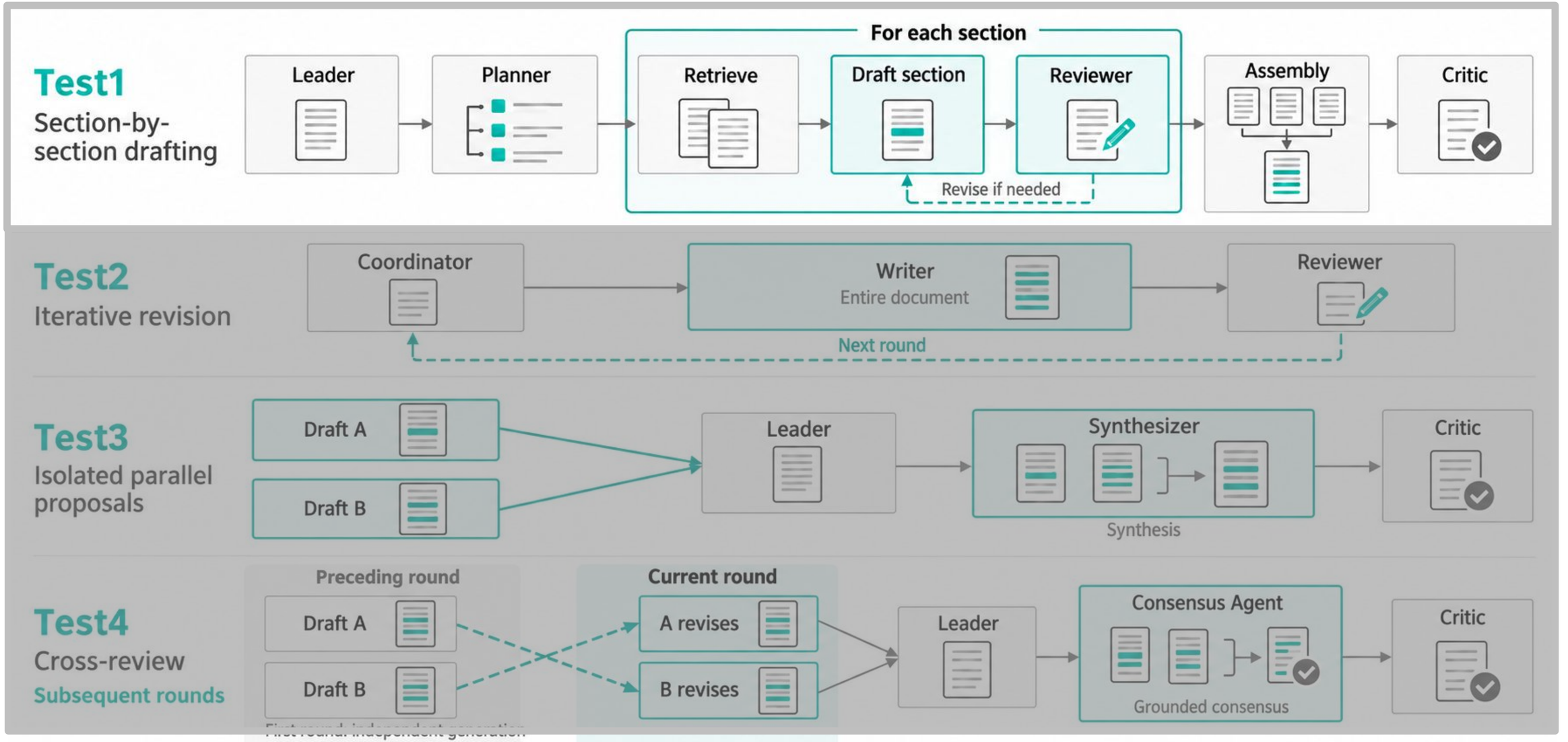
Document generation architectures

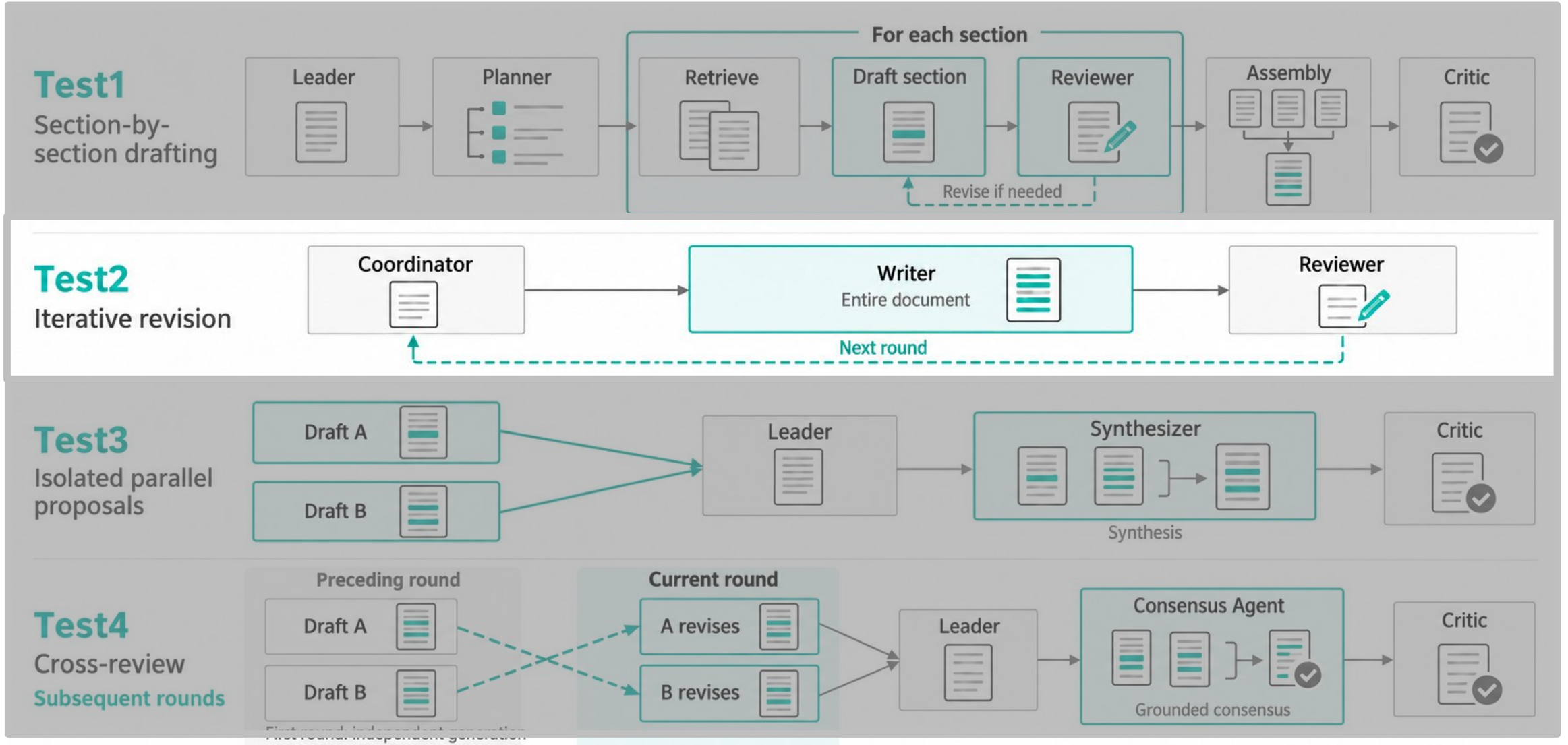


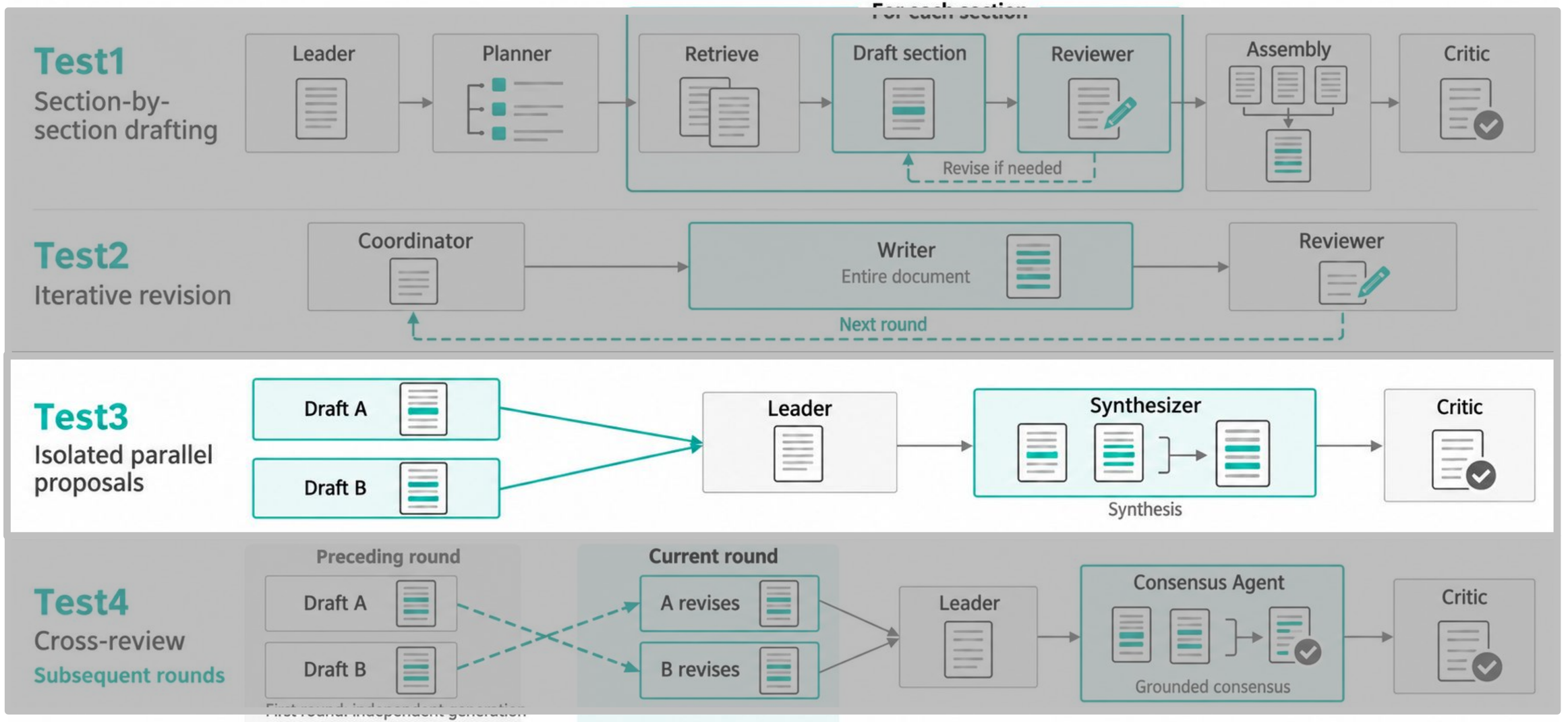
Independent evaluation: OpenAI GPT-5.1

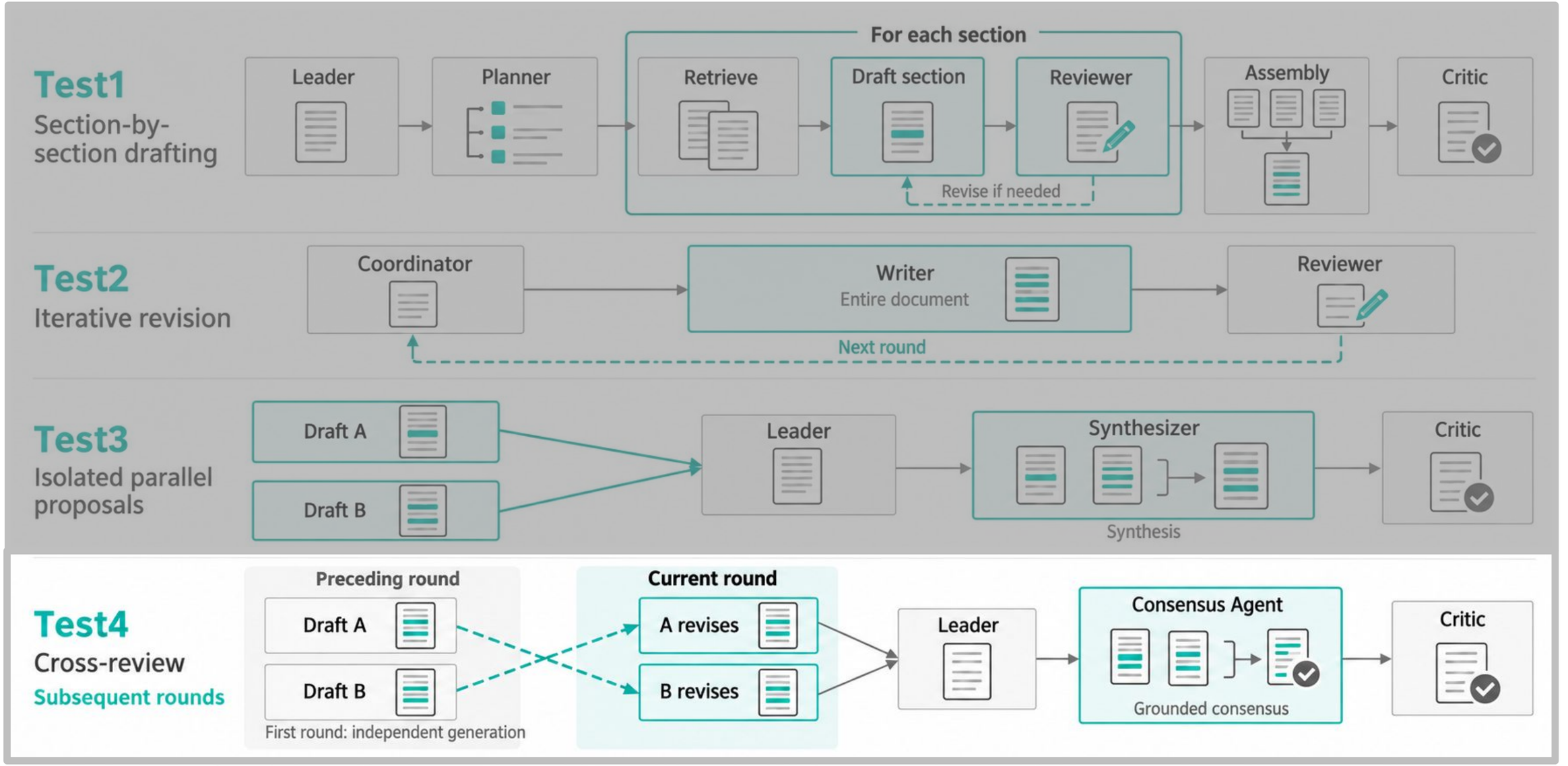
Document generation architectures




Test1
Section-by-section drafting
Leader
Planner
For each section
Retrieve
Draft section
Reviewer
Revise if needed
Assembly
Critic
Test2
Iterative revision
Coordinator
Writer
Entire document
Reviewer
Next round
Test3
Isolated parallel proposals
Draft A
Draft B
Leader
Synthesizer
Synthesis
Critic
Test4
Cross-review
Subsequent rounds
Preceding round
Draft A
Draft B
Current round
A revises
B revises
Leader
Consensus Agent
Grounded consensus
Critic


Test1
Section-by-section drafting
Leader
Planner
For each section
Retrieve
Draft section
Reviewer
Revise if needed
Assembly
Critic
Test2
Iterative revision
Coordinator
Writer
Entire document
Reviewer
Next round
Test3
Isolated parallel proposals
Draft A
Draft B
Leader
Synthesizer
Synthesis
Critic
Test4
Cross-review
Subsequent rounds
Preceding round
Draft A
Draft B
Current round
A revises
B revises
Leader
Consensus Agent
Grounded consensus
Critic


Test1
Section-by-section drafting
Leader
Planner
For each section
Retrieve
Draft section
Reviewer
Revise if needed
Assembly
Critic
Test2
Iterative revision
Coordinator
Writer
Entire document
Reviewer
Next round
Test3
Isolated parallel proposals
Draft A
Draft B
Leader
Synthesizer
Synthesis
Critic
Test4
Cross-review
Subsequent rounds
Preceding round
Draft A
Draft B
Current round
A revises
B revises
Leader
Consensus Agent
Grounded consensus
Critic


Test1
Section-by-section drafting
Leader
Planner
For each section
Retrieve
Draft section
Reviewer
Revise if needed
Assembly
Critic
Test2
Iterative revision
Coordinator
Writer
Entire document
Reviewer
Next round
Test3
Isolated parallel proposals
Draft A
Draft B
Leader
Synthesizer
Synthesis
Critic
Test4
Cross-review
Subsequent rounds
Preceding round
Draft A
Draft B
First round: independent generation
Current round
A revises
B revises
Leader
Consensus Agent
Grounded consensus
Critic

Experimental design

10 × **5** × **5** = **250**

In-house input documents | Generation structures | Successful outputs | Documents

---

Generation model — **Gemma-4-31B**

Common template — Same draft content
all five generation structures

Unit of analysis — Input draft
Mean of the five repetitions

Independent evaluation

Final documents + Input draft

OpenAI GPT-5.1

Item-level binary judgments

YES / NO

→

Evaluation code

Four dimensions + Overall score
0–100

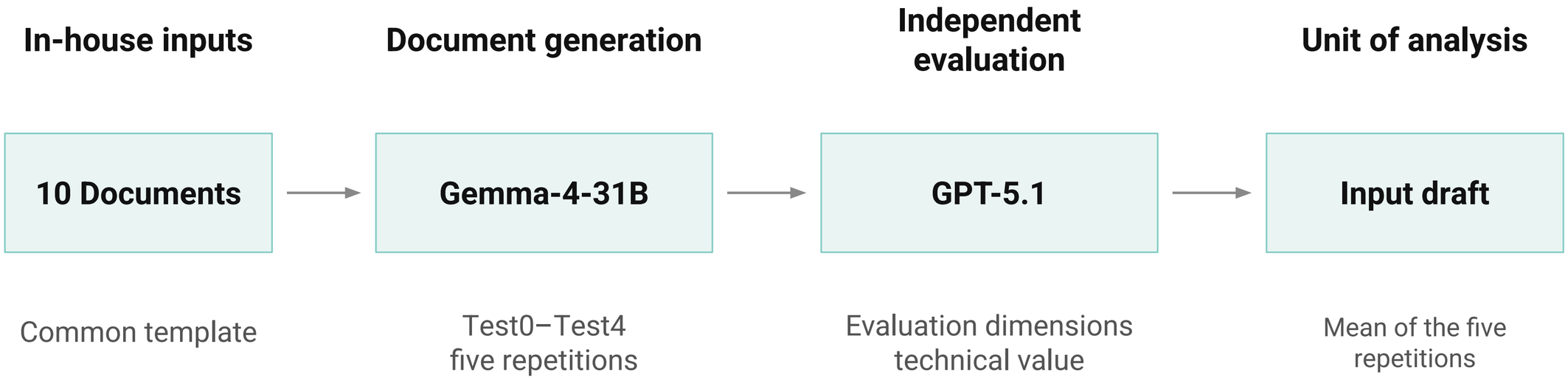
Logical coherence

Structural integrity

Faithfulness to the draft

Technical specificity

Comparative framework



Internal evaluation and review

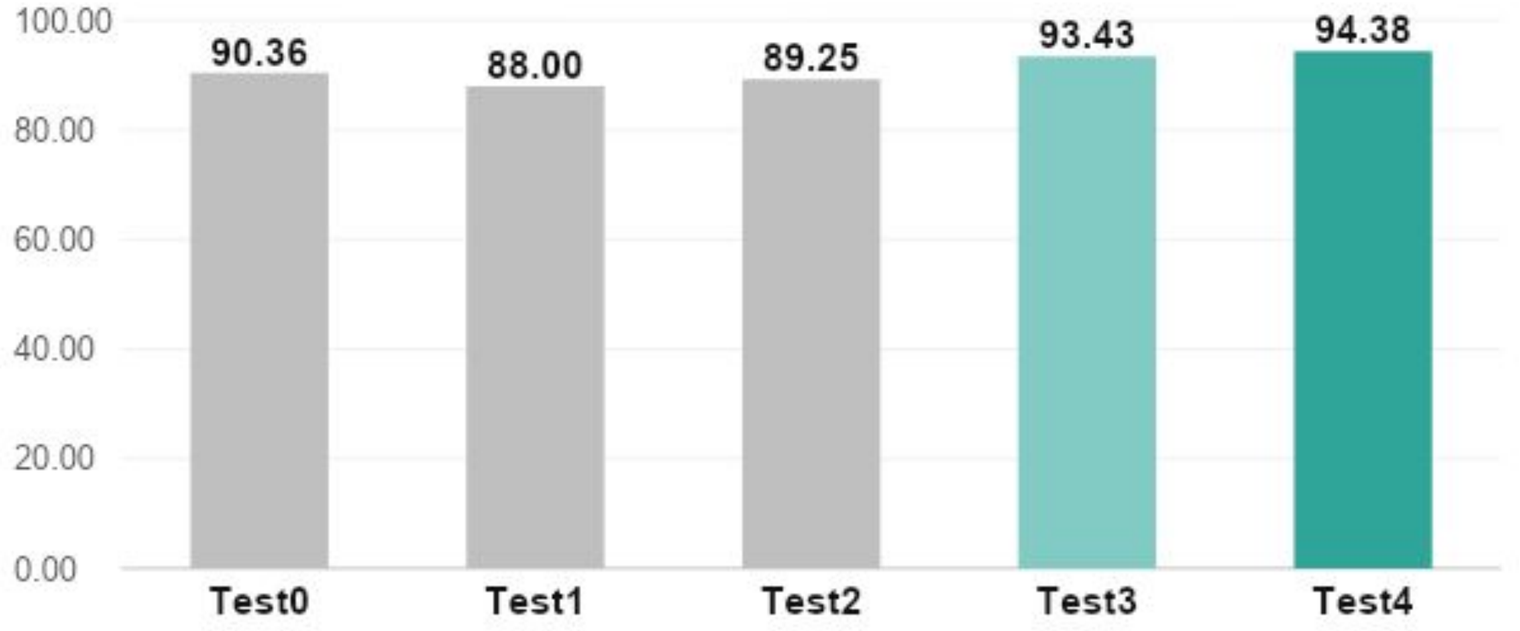
During the generation process

Independent LLM judge

Final documents

Document quality

Mean Judge scores (0–100)



Highest mean score

94.38

Test4

Generation structures

p = .011

Post-hoc comparison

Test4 > Test1
Adjusted p = .039

Test3 / Test4

No significant difference

Analysis by evaluation dimension

Mean evaluation-dimension scores (0–100)

	Logical Coherence	Structural Integrity	Faithfulness to the Draft	Technical Specificity
Test0	99.21	95.33	69.00	98.00
Test1	98.50	69.50	84.83	98.67
Test2	98.00	90.67	75.50	92.83
Test3	98.21	87.78	92.00	95.33
Test4	98.00	91.83	91.00	96.67

Color intervals: 65–74 | 75–79 | 80–89 | 90–94 | 95–97 | 98–100

Structural Integrity

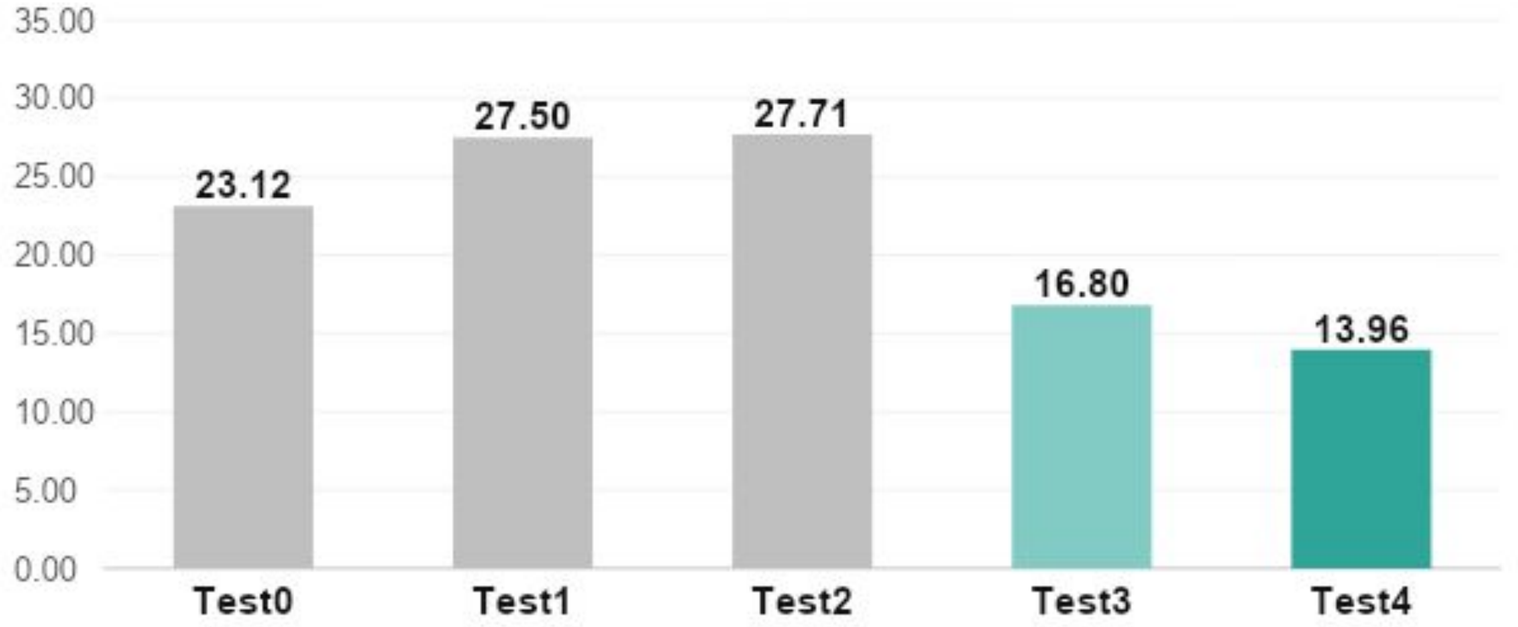
Generation structures: p < .001
Test1: lowest mean score

Faithfulness to the Draft

Test3 / Test4: higher
Not statistically significant

Repeatability robustness

Mean Repetition Range



Smallest mean repetition range

13.96

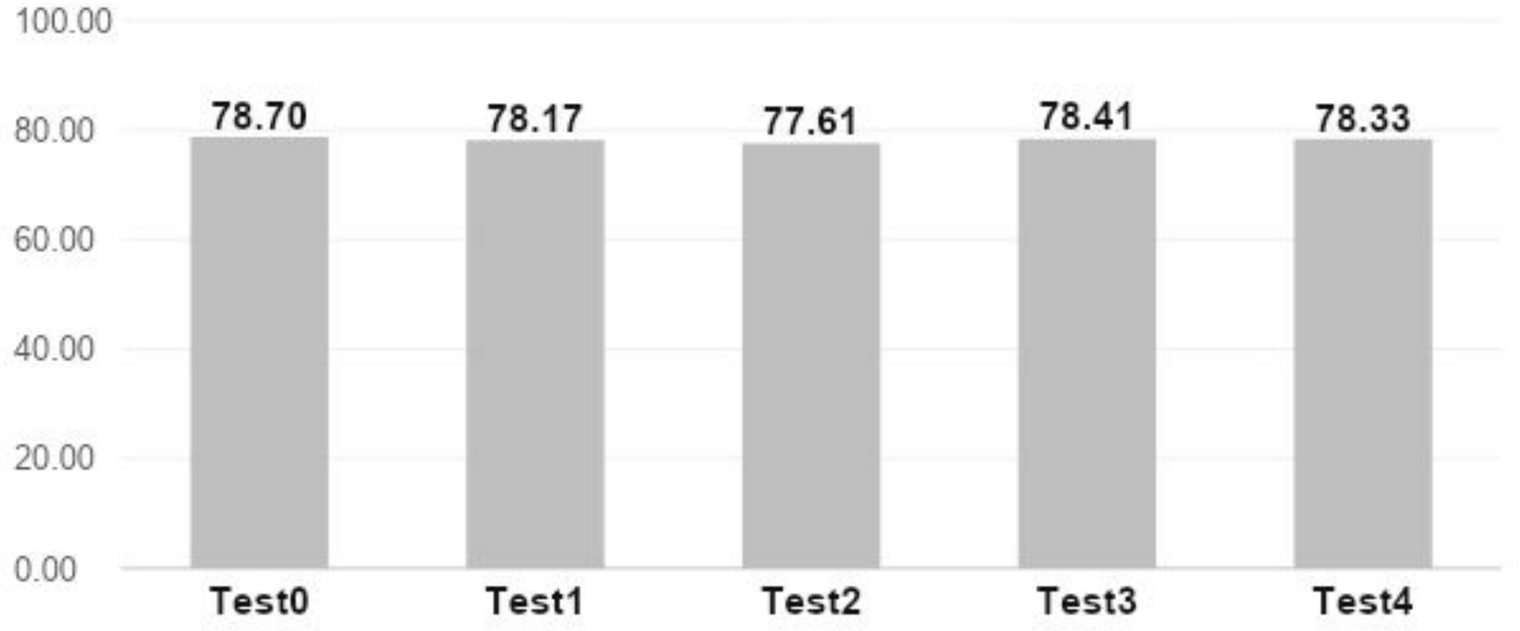
Test4
Maximum − Minimum

Overall difference
p = .009

No significant pairwise differences

Technical value evaluation

Document-based technical value scores (0–100)



Generation structures

p = .997

No significant differences

Technology and intellectual property criteria

55 points → 100-point scale

Excluded from the final score

Technology substitutability
Intensity of technological competition
Marketability

Discussion

Document quality

Test1

Low structural integrity

Test3 and Test4

High faithfulness to the input draft

Test4

Balanced results
without a major weakness

Faithfulness to the input draft: not statistically significant

Limitations

Ten in-house documents

Limited preparation method
and technical domains

Variability

Generation model + LLM judge

Further validation

Real-world technical documents
Evaluations by domain experts

Conclusions

01 Test4: highest mean document-quality score

94.38 / Smallest repetition range: 13.96

02 Significant difference: only Test4 / Test1

After adjustment for multiple comparisons
Test3 / Test4: no significant difference

03 Document-based technical value scores

No significant differences

04 Further validation

Diverse real-world technical documents
Expert evaluation

Thank you for your attention

Presenter

Jinkyung Kwon

Author

Jinkyung Kwon [a], Hogeon Seo [a,b]*

[a]KAERI · [b]UST

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.