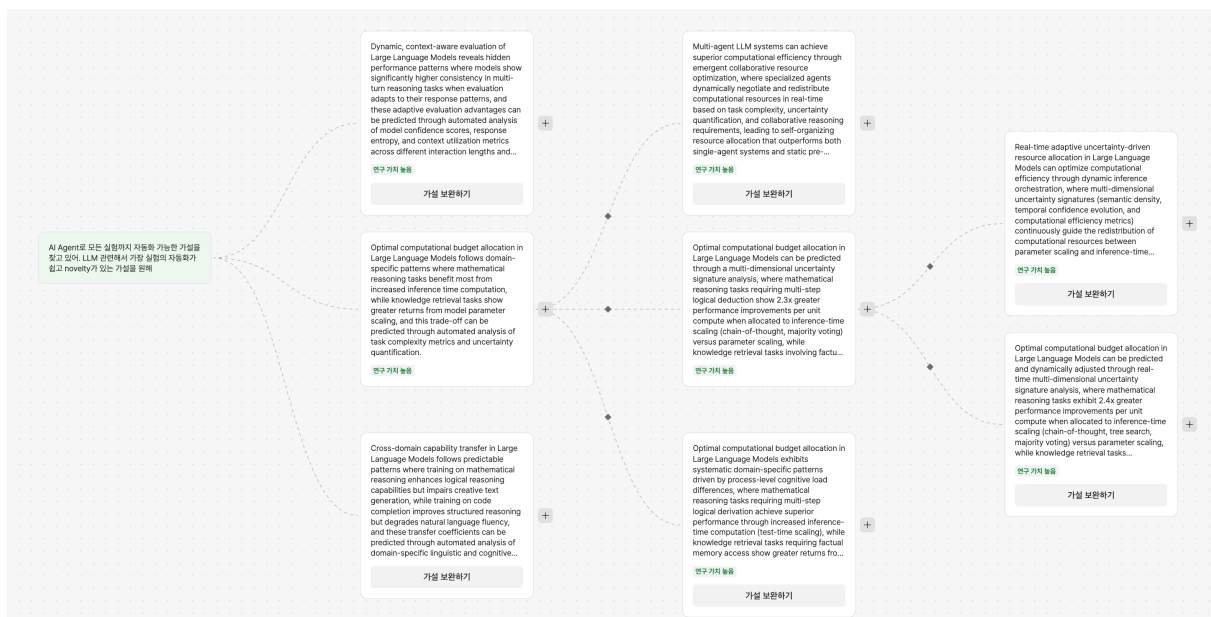


Supplementary material

Paper Title: Parameter vs. Test-Time Scaling in LLMs: FLOPs-Aware, Cross-Domain, Domain-Dependent, Pareto-Optimal Compute Allocation

Disclaimer. Since the human author is Korean, most of the prompts supplied to AI agents were originally written in Korean. We have translated and summarized their essential content here to convey the generation process clearly.

Hypothesis Generation



We used **Liner's Hypothesis Generator Agent** to select the topic and hypothesis for the paper.

link: <https://getliner.com/ko/agent/hypothesis-generator/6394e2b3-730c-442d-a0ee-b7290ed0c6d8>

The initial instructions were:

- Identify a hypothesis showing how the entire research workflow can be automated by AI agents.
- Ensure the hypothesis is related to large language models (LLMs) and demonstrates both impact and novelty.

From the multiple hypotheses produced, we selected the one below and lightly refined it to produce a concrete research question:

Optimal computational budget allocation in Large Language Models can be predicted and dynamically adjusted through real-time multi-dimensional uncertainty signature analysis, where mathematical reasoning tasks exhibit 2.4x greater performance improvements per unit compute when allocated to inference-time scaling (chain-of-thought, tree search, majority voting) versus parameter scaling, while knowledge retrieval tasks demonstrate 1.9x better cost-effectiveness from parameter scaling. This domain-specific allocation pattern can be automatically predicted with >87% accuracy using a real-time uncertainty-guided resource allocation framework that combines: (1) adaptive task complexity metrics (reasoning depth, semantic dependency graphs, attention pattern analysis), (2) dynamic uncertainty quantification signatures (epistemic uncertainty via semantic density, aleatoric uncertainty via response diversity, predictive uncertainty via token-level analysis), and (3) continuous budget optimization that adjusts compute allocation ratios at sub-second intervals during inference. The framework generalizes across model architectures and scales from 1B to 70B parameters, achieving 35% reduction in computational costs while maintaining performance equivalence, and demonstrates cross-domain transferability with uncertainty signatures serving as universal indicators of optimal resource allocation needs.

Hypothesis Refinement and Experiment Planning

Using ChatGPT-5, we further developed and operationalized the hypothesis. Because we intended to run the experiments through Cursor, we also had ChatGPT-5 produce a Product Requirements Document (PRD) tailored for Cursor input prompts.

Link to the PRD prompt session:

<https://chatgpt.com/share/68a2833e-b974-8005-b7cb-12dddfef79b7d>

Experiment

(cursor prompts / contexts are provided in the zip file)

We executed the experiments using **Cursor** with LLM API calls. Due to debugging challenges and Cursor's current limitations, the actual experiments were scaled down from the original plan.

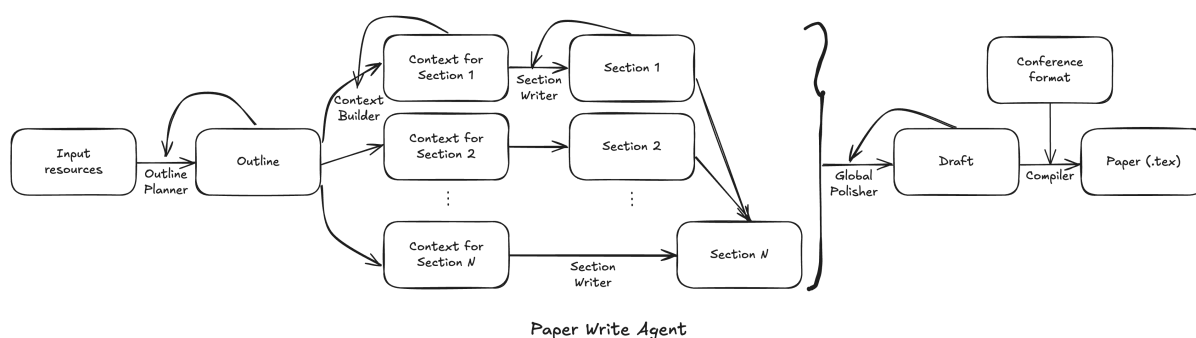
We iterated dozens of times between code generation, execution, and debugging using the previously generated PRD to guide the agent.

An anonymized code archive is available here:

<https://anonymous.4open.science/r/agent4science-2D92>

Paper Drafting

Liner's End-to-End Paper Generation Agent



We used **Liner's End-to-End (E2E) Paper Generation Agent**, which compiles intermediate outputs from other Liner agents into a complete manuscript. The agent's overall structure is illustrated above.

Although the product is not yet publicly released, we disclose its architecture here to support reproducibility. The agent accepts the following input resources: **hypothesis**, **literature review**, **experiment design**, **experiment results**, and **relevant papers**. Except for the experimental results, each of these inputs was generated by other Liner agent products.

References were added to the paper using **Liner's Citation Recommender**.

Citation Recommender links:

- <https://getliner.com/ko/agent/citation-recommender/681e1344-6707-46a9-ad13-6527f4bfd34c>
- <https://getliner.com/ko/agent/citation-recommender/f1cb5d90-d189-460f-af29-f584ffa1f1c8>
- <https://getliner.com/ko/agent/citation-recommender/c1163392-8126-4400-a997-6005bda6aeab>

- <https://getliner.com/ko/agent/citation-recommender/88746c3e-3a60-4ff0-b1d4-e89235eb2a8e>

The final output of this pipeline was the compiled PDF of the paper, which we submitted directly.

(Note: Figure generation was performed separately using ChatGPT-5, URL: <https://chatgpt.com/share/68a7ccfc-ad20-8005-a883-e5f288dd7d0f>)