ANALISIS PERBANDINGAN KUALITAS JAWABAN PADA QA BERBASIS RAG : KOMBINASI ZERO-SHOT INSTRUCTION PROMPTING DAN SELF-VERIFICATION

Nevitya Elmaira Nurjannah[1)], Cendra Devayana Putra[2)]
[1)] Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: nevitya.22138@mhs.unesa.ac.id
[2)] Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: putracendra@unesa.ac.id

Abstract
Retrieval-Augmented Generation (RAG) reduces hallucinations by grounding language model (LM) responses in the context of retrieved results, but does not automatically guarantee evidence-based (faithful) responses. This study examines the effect of combining zero-shot instruction prompting and self-verification on the faithfulness of responses to answerable questions, across four RAG pipeline configurations: (1) RAG, (2) RAG + prompting, (3) RAG + self-verification, and (4) RAG + prompting + self-verification, which were tested on three scales of the Qwen3 language model (0.6B, 4B, 8B) using the SQuAD v2.0 dataset. A total of 100 queries were selected via stratified random sampling from the validation split to avoid topic bias. Faithfulness was measured at the claim level using an NLI model (DeBERTa-v3-large). The results show that basic RAG (configuration 1) achieved the highest average faithfulness score, while configurations 2–4 (with prompting and/or self-verification) showed relatively similar and lower scores. In SLM, adding self-verification lowered faithfulness the most, while in LLM the score remained relatively stable across configurations. A retrieval-quality control analysis indicated that this decline was linked to the generator's capacity rather than retrieval quality. These findings suggest that the combination of prompting and self-verification does not automatically improve the quality of evidence-based answers, and its benefits depend on the adequacy of the language model's capacity.
Keywords: Retrieval-Augmented Generation, Self-Verification, Zero-shot Instruction Prompting, Faithfulness, Large Language Model.

Abstrak
Retrieval-Augmented Generation (RAG) menurunkan halusinasi dengan mendasarkan jawaban language model (LM) pada konteks hasil retrieval, namun tidak otomatis menjamin jawaban berbasis bukti (faithful). Penelitian ini menguji pengaruh kombinasi zero-shot instruction prompting dan self-verification terhadap faithfulness jawaban pada pertanyaan answerable, di antara 4 konfigurasi pipeline RAG: (1) RAG, (2) RAG + prompting, (3) RAG + self-verification, dan (4) RAG + prompting + self-verification, yang diuji pada 3 skala language model Qwen3 (0.6B, 4B, 8B) menggunakan dataset SQuAD v2.0. Sebanyak 100 query diambil dengan metode stratified random sampling dari split validation untuk menghindari bias topik. Faithfulness diukur pada level claim menggunakan model NLI (DeBERTa-v3-large). Hasil menunjukkan bahwa RAG dasar (konfigurasi 1) memperoleh rata-rata skor faithfulness tertinggi, sedangkan konfigurasi 2–4 (dengan prompting dan/atau self-verification) menunjukkan skor yang relatif lebih rendah dan cenderung serupa satu sama lain. Pada SLM, penambahan self-verification menurunkan faithfulness paling besar, sedangkan pada LLM skor relatif stabil di seluruh konfigurasi. Analisis kontrol kualitas retrieval mengindikasikan bahwa penurunan tersebut berkaitan dengan kapasitas generator, bukan kualitas retrieval. Temuan ini menunjukkan bahwa kombinasi prompting dan self-verification tidak otomatis meningkatkan kualitas jawaban berbasis bukti, serta manfaatnya bergantung pada kecukupan kapasitas language model.
Kata Kunci: Retrieval-Augmented Generation, Self-Verification, Zero-shot Instruction Prompting, Kesetiaan, Model Bahasa Besar.

PENDAHULUAN

Perkembangan language model (LM) generatif mendorong pemanfaatan sistem question answering (QA) yang mampu menghasilkan jawaban dengan cepat (Najork, 2025). Namun, model dituntut memberikan jawaban berdasarkan bukti (evidence-based) (Ammann et al., 2025), karena model masih dapat menghasilkan respons yang tampak meyakinkan tetapi mengandung kesalahan faktual atau hallucination (Y. Chen et al., 2023). Skala permasalahan ini tampak pada benchmark TruthfulQA, di mana model terbaik hanya menghasilkan jawaban truthful sekitar 58%, sementara performa manusia mencapai sekitar 94% (Lin et al., 2022).

Retrieval-Augmented Generation (RAG) menjadi salah satu pendekatan yang banyak digunakan untuk mengatasi masalah tersebut, dengan menggabungkan proses pencarian informasi relevan (retriever), penyusunan konteks (augmentation), dan pembentukan jawaban (generator) (Gupta et al., 2024; Klesel & Wittmann, 2025). Meskipun berpotensi menurunkan halusinasi (Y. Zhang, 2025), RAG tidak otomatis menjamin jawaban selalu benar; bahkan perangkat riset hukum berbasis RAG dilaporkan masih berhalusinasi hingga 33% (Magesh et al., 2024).

Penelitian terdahulu mengeksplorasi penguatan RAG melalui strategi prompting dan verifikasi. Integrasi chain-of-thought dengan RAG membuat penalaran lebih terstruktur dan selaras dengan konteks hasil retrieval (Li et al., 2025), sedangkan pendekatan self-verification multi-tahap mendorong model mengidentifikasi klaim, memeriksa dukungan bukti, lalu memperbaiki jawaban (García et al., 2025). Namun, kedua teknik memiliki keterbatasan yang berbeda: pendekatan zero-shot prompting mengarahkan model tanpa mekanisme verifikasi, sedangkan self-verification memeriksa jawaban tanpa instruksi grounding yang jelas pada tahap awal, sehingga belum jelas apakah kombinasi keduanya konsisten unggul dibanding penerapan masing-masing secara terpisah.

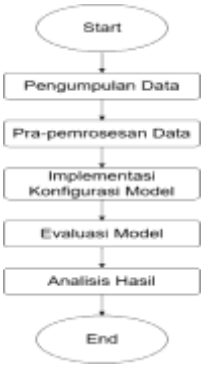
Beberapa penelitian mengombinasikan ketiga elemen tersebut sekaligus. Chen et al. (2025) mengombinasikan RAG, chain-of-thought, dan self-verification untuk tugas relation extraction pada literatur medis berbahasa Mandarin, dan melaporkan peningkatan F1 dibanding baseline RAG-only, dengan skenario single-step prompting yang dikombinasikan chain-of-thought dan self-verification memberikan hasil terbaik. Kumar et al. (2025) menunjukkan bahwa kombinasi RAG, chain-of-thought, dan self-verification meningkatkan reliabilitas jawaban (menekan halusinasi) pada benchmark HaluEval, FEVER, dan TruthfulQA, dengan self-verification secara konsisten menjadi komponen paling berpengaruh di antara ketiganya.

Selain itu, L. Chen et al. (2024) mengusulkan RC-RAG dengan teknik counterfactual prompting untuk menilai kualitas retrieval dan pemakaian evidence, lalu menentukan judgement keep vs discard atas jawaban (gating); pendekatan ini melaporkan peningkatan faithfulness namun disertai trade-off penurunan coverage jawaban. He et al. (2024) memperkenalkan Chain-of-Verification (CoV-RAG) dengan tahapan scoring, judgement, rewriting, dan re-retrieval menggunakan revised query untuk memperbaiki hasil retrieval dan jawaban model secara end-to-end. Sementara itu, Mansurova et al. (2024) menekankan pentingnya evaluasi sistem QA-RAG pada knowledge gap detection (unanswerable), dan menunjukkan bahwa integrasi external truth pada sistem RAG masih menjadi tantangan terbuka.

Berdasarkan penelitian-penelitian tersebut, pendekatan RAG yang dikombinasikan dengan prompting dan mekanisme verifikasi diri terbukti berpotensi meningkatkan konsistensi faktual jawaban model, namun sebagian besar studi mengevaluasi teknik tersebut secara terpisah dan hanya pada satu skala model, sehingga belum jelas apakah manfaatnya konsisten lintas skala language model yang berbeda. Pengujian lintas skala model tersebut penting dilakukan, karena kemampuan mengikuti instruksi berlapis dibatasi oleh kapasitas model, sebuah batasan yang dikenal sebagai capacity wall. Konsep ini merujuk pada fenomena kemunculan kemampuan (emergent abilities) tertentu, seperti reasoning bertahap dan mengikuti instruksi kompleks, yang baru muncul secara konsisten setelah model melewati ambang skala parameter tertentu dan tidak dapat diprediksi hanya dengan mengekstrapolasi performa dari model berskala lebih kecil (Wei et al., 2022). Sejalan dengan itu, (T. Zhang et al., 2024) menunjukkan bahwa language model berskala kecil memiliki keterbatasan dalam memproses instruksi optimasi berlapis, sehingga pada skala kecil, manfaat teknik lanjutan seperti zero-shot instruction prompting dan self-verification belum tentu konsisten. Berdasarkan uraian tersebut, penelitian ini menguji pengaruh kombinasi zero-shot instruction prompting dan self-verification, pada 4 konfigurasi pipeline RAG, terhadap faithfulness jawaban pertanyaan answerable, lintas 3 skala language model keluarga Qwen3.

METODE PENELITIAN

Pada bab ini membahas mengenai kerangka penelitian yang digunakan sebagai dasar pelaksanaan penelitian. Penelitian ini menggunakan pipeline Retrieval-Augmented Generation (RAG) dengan tahapan penelitian ditunjukkan pada gambar berikut:



Gambar 1. Kerangka penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan dataset publik SQuAD v2.0, split validation, diperoleh dari repositori rajpurkar/squad_v2.0 (Rajpurkar et al., 2018). Sebanyak

100 query (kombinasi pertanyaan answerable dan unanswerable) diambil dengan metode stratified random sampling.

3.2 Pra-Pemrosesan dan Pembentukan Indeks

Konteks dataset diproses menjadi embedding menggunakan model Qwen3-Embedding yang skalanya disetarakan dengan generator pada tiap eksperimen, kemudian diindeks menggunakan FAISS untuk mendukung pencarian top-k pada tahap retrieval.

3.3 Implementasi Konfigurasi Pipeline RAG

Pipeline RAG terdiri dari 3 tahap, yaitu retrieval (pencarian top-k), augmentation (penyusunan prompt dari konteks hasil retrieval), dan generation (do_sample=False, temperature=0 agar tidak ada variansi antar run) dan eksperimen dijalankan pada 3 skala language model keluarga Qwen3 berikut:

Tabel 1. Konfigurasi model tiap skala

Skala	Language Model	Embedding Model
SLM	Qwen3-0.6B	Qwen3-Embedding-0.6B
MLM	Qwen3-4B	Qwen3-Embedding-4B
LLM	Qwen3-8B	Qwen3-Embedding-8B

Kemudian, setiap skala tersebut akan diuji pada 4 konfigurasi berikut:

Tabel 2. Konfigurasi pipeline

Konfigurasi	Komponen	Keterangan
1	RAG	Retrieval + generation tanpa instruksi tambahan
2	RAG + Prompt	+ instruksi zero-shot untuk menjawab berbasis konteks & abstain jika tidak memadai
3	RAG + SV	+ tahap self-verification dua-langkah pasca-generation
4	RAG + Prompt + SV	Kombinasi konfigurasi 2 dan 3

Instruksi zero-shot yang digunakan pada konfigurasi 2 dan 4 mengikuti dua aturan, yaitu grounding (model harus menjawab berdasarkan konteks hasil retrieval) dan abstain (model harus menolak menjawab ketika konteks tidak memadai) (Sun et al., 2023), sebagaimana ditampilkan pada tabel dibawah ini.

Tabel 3. Instruksi zero-shot instruction prompting

Instruksi
Answer the question ONLY based on the information in the CONTEXT. Do not guess or add facts not present in the CONTEXT. If the CONTEXT does not contain the information needed to answer the question, politely and specifically acknowledge the limitation and offer one relevant form of further assistance.
OUTPUT FORMAT (REQUIRED): Answer: [your answer based on the CONTEXT, or a polite and contextual abstain statement] Evidence: [the part of the CONTEXT that supports the answer, or "none" if abstaining]

Aturan grounding merujuk pada prinsip faithfulness, yaitu sejauh mana jawaban didukung oleh konteks yang di-retrieve (Es et al., 2024), sedangkan format jawaban yang menyertakan evidence mengikuti prinsip penyertaan bukti pendukung pada jawaban (Gao et al., 2023). Aturan abstain merujuk pada karakteristik unanswerable questions pada SQuAD v2.0 (Rajpurkar et al., 2018).

Sementara itu, self-verification pada konfigurasi 3 dan 4 diterapkan melalui tahap verify pasca-generation untuk memeriksa apakah jawaban yang dihasilkan model didukung oleh konteks, mengikuti prinsip pemeriksaan diri model (Weng et al., 2023) dan konsep Chain-of-Verification (Dhuliawala et al., 2023). Instruksi verifikasi yang digunakan ditampilkan pada Tabel 4 dibawah ini.

Tabel 4. Instruksi self-verification

Instruksi
You are a verifier for a RAG system. Compare the CANDIDATE_ANSWER with the CONTEXT and determine the label: SUPPORTED (all key claims are supported by the CONTEXT) or UNSUPPORTED (there are claims not supported, or the CONTEXT is insufficient).
If SUPPORTED: refine the CANDIDATE_ANSWER without adding new facts. If UNSUPPORTED: produce a polite and specific abstain statement. OUTPUT FORMAT (REQUIRED): Label: SUPPORTED or UNSUPPORTED / Answer: [final answer] / Evidence: [supporting part of CONTEXT, or "none"]
Jika label SUPPORTED, jawaban dikeluarkan; jika UNSUPPORTED, model menghasilkan pernyataan abstain sebagai pengganti jawaban akhir.

3.4 Evaluasi Model

Selanjutnya, hasil penelitian akan diukur menggunakan metrik faithfulness, apakah jawaban model

 ISSN 3063-864X (E)

didukung bukti/tidak, dan dihitung pada level claim dengan model Natural Language Inference (NLI) DeBERTa-v3-large, dengan sebuah claim dianggap didukung konteks apabila skor entailment > 0,5 dan lebih besar dari skor contradiction (Honovich et al., 2022):

$$Faithfulness = \frac{S}{K}$$

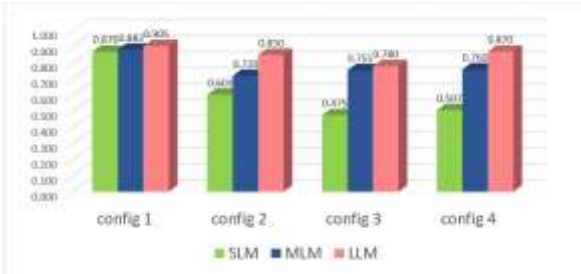
dengan S =jumlah claim yang didukung konteks, K= jumlah seluruh claim hasil pemecahan jawaban model. Metrik ini hanya dihitung pada subset pertanyaan answerable dari 100 sampel tersebut.

3.5 Analisis Hasil

Hasil evaluasi faithfulness dianalisis secara deskriptif dengan membandingkan rata-rata skor antar 4 konfigurasi pipeline dan 3 skala language model, yang divisualisasikan dalam bentuk grafik. Sebagai pelengkap, dilakukan analisis kontrol pada skor retrieval untuk mengonfirmasi sumber perbedaan performa antar skala model.

HASIL DAN PEMBAHASAN

4.1 Hasil Evaluasi Faithfulness



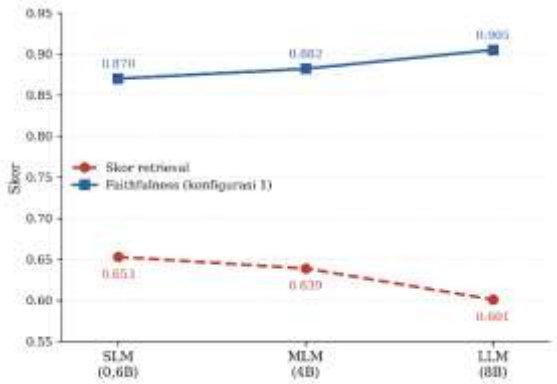
Gambar 2. bar chart faithfulness

Grafik rata-rata skor faithfulness per konfigurasi dan skala model menunjukkan LLM mencatat skor tertinggi dan paling stabil (0,780–0,905) di seluruh konfigurasi. SLM mengalami penurunan saat self-verification ditambahkan, dari 0,870 (konfigurasi 1) menjadi 0,475 dan 0,507 (konfigurasi 3 dan 4), sedangkan MLM berada di posisi menengah dengan pola lebih stabil. Secara umum, konfigurasi 1 (RAG dasar) menunjukkan skor faithfulness lebih tinggi dibanding tiga konfigurasi lainnya di seluruh skala model, sedangkan skor antar konfigurasi 2–4 relatif berdekatan satu sama lain. Pada perbandingan antar skala, perbedaan paling menonjol terlihat pada SLM, sementara MLM dan LLM menunjukkan skor yang relatif berdekatan.

Pola penurunan faithfulness pada SLM ketika self-verification ditambahkan sejalan dengan konsep capacity wall, yaitu keterbatasan model skala kecil dalam mengikuti instruksi berlapis (multi-step) secara konsisten (Wei et al., 2022). Pada tahap verifikasi, model SLM diduga kesulitan membedakan klaim yang benar-benar didukung konteks dengan klaim yang tampak mirip namun sebenarnya tidak relevan, sehingga jawaban yang semula benar justru direvisi menjadi tidak didukung bukti atau diubah menjadi abstain yang tidak diperlukan. Pola ini berbeda dengan temuan (Kumar et al., 2025), yang melaporkan self-verification sebagai komponen paling berpengaruh dalam menekan halusinasi pada model berskala lebih besar. Perbandingan ini memperkuat indikasi bahwa manfaat self-verification bergantung pada kapasitas model, bukan berlaku universal di seluruh skala language model.

4.2 Analisis Kontrol Kualitas Retrieval

Penurunan skor faithfulness pada SLM tersebut bukan disebabkan oleh kualitas retrieval, mengingat skala model embedding dan generator disetarakan dalam desain penelitian ini, maka dari itu dilakukan analisis kontrol sebagaimana ditampilkan pada Gambar 3 dibawah ini.



Gambar 3. Skor Retrieval vs Faithfulness per Skala

Hasil tersebut menunjukkan skor retrieval menurun seiring bertambahnya skala, berlawanan arah dengan skor faithfulness yang meningkat, mengonfirmasi bahwa penurunan faithfulness pada SLM berasal dari kapasitas generator, bukan kualitas retrieval.

4.3 Error Analysis

Berikut contoh kasus kegagalan/analisis error faithfulness pada pertanyaan answerable, dari model SLM konfigurasi 2 (RAG+Prompt):

Tabel 5. contoh kasus error faithfulness

Pertanyaan	**"What party rules in Melbourne's inner regions?"**
Ground truth	The Greens
Konteks ter-retrieve	The centre-left Australian Labor Party (ALP), the centre-right Liberal Party of Australia, the rural-based National Party of Australia, and the environmentalist Australian Greens are Victoria's main political parties...

Jawaban model	"The party that rules in Melbourne's inner regions is the Liberal Party of Australia."
faithfulness: 0,00 ; retrieval score: 0,733	

Contoh kasus tersebut, konteks sudah memuat informasi yang benar (Australian Greens) dengan skor retrieval yang tinggi, tetapi model menjawab “Liberal Party of Australia”. Kesalahan ini terjadi pada tahap generation, umumnya disebabkan oleh keterbatasan pemahaman konteks pada model skala kecil.

KESIMPULAN

Penelitian ini menunjukkan bahwa konfigurasi pipeline dan skala model memengaruhi faithfulness jawaban pada pertanyaan answerable di antara 4 konfigurasi pipeline RAG. Bertentangan dengan dugaan awal, kombinasi zero-shot instruction prompting dan self-verification tidak otomatis meningkatkan kualitas jawaban berbasis bukti; RAG dasar (konfigurasi 1) justru menghasilkan faithfulness lebih tinggi dibanding ketiga konfigurasi lainnya, terutama pada SLM, dan efektivitas self-verification bergantung pada kecukupan kapasitas generator. Bagi peneliti NLP selanjutnya, disarankan mengeksplorasi mekanisme verifikasi alternatif yang berpotensi lebih sesuai untuk model skala kecil. Bagi pengembang sistem QA berbasis RAG, pemilihan konfigurasi sebaiknya disesuaikan dengan skala model yang tersedia.

Ucapan Terima Kasih

Terima kasih kepada bapak Cendra atas bimbingan yang diberikan selama proses penelitian ini, serta kepada Program Studi S1 Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya atas dukungan fasilitas dan lingkungan akademik yang mendukung penyelesaian penelitian.

DAFTAR PUSTAKA

Ammann, P. J. L., Golde, J. & Akbik, A. (2025). Question Decomposition for Retrieval-Augmented Generation.

Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z. & Xiao, Y. (2023). Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 245–255. https://doi.org/10.1145/3583780.3614905

Chen, Z., Hao, J., Sun, H., Zhao, L., Li, J., Qian, Q., Peng, Q., Wang, X., Cong, S., Shen, L., Guo, Z., Pu, S. & Lin, Y. (2025). MedScaleRE-PF: a prompt-based framework with retrieval-augmented generation, chain-of-thought, and self-verification for scale-specific relation extraction in Chinese medical literature. Information Processing & Management, 62(6), 104278. https://doi.org/10.1016/j.ipm.2025.104278

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A. & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models.

Es, S., James, J., Espinosa Anke, L. & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations,150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Gao, T., Yen, H., Yu, J. & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations.

García, F. G., Shi, Q. & Feng, Z. (2025). Enhancing Factual Accuracy and Citation Generation in LLMs via Multi-Stage Self-Verification.

Gupta, S., Ranjan, R. & Singh, S. N. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A. & Matias, Y. (2022). TRUE: Re-evaluating Factual Consistency Evaluation. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics.

Klesel, M. & Wittmann, H. F. (2025). Retrieval-Augmented Generation (RAG). Business & Information Systems Engineering, 67(4), 551–561. https://doi.org/10.1007/s12599-025-00945-3

Kumar, A., Kim, H., Nathani, J. S. & Roy, N. (2025). Improving the Reliability of LLMs: Combining CoT, RAG, Self-Consistency, and Self-Verification.

Li, F., Fang, P., Shi, Z., Khan, A., Wang, F., Wang, W., Zhang, X. & Cui, Y. (2025). CoT-RAG: Integrating Chain of Thought and Retrieval-Augmented Generation to Enhance Reasoning in Large Language Models.

Lin, S., Hilton, J. & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D. & Ho, D. E. (2024). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools.

Najork, M. (2025). Generative Information Retrieval. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, 5987–5987. https://doi.org/10.1145/3711896.3736806

Rajpurkar, P., Jia, R. & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD.

Sun, J., Shaib, C. & Wallace, B. C. (2023). Evaluating the Zero-shot Robustness of Instruction-tuned Language Models.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. & Fedus, W. (2022). Emergent Abilities of Large Language Models.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K. & Zhao, J. (2023). Large Language Models are Better Reasoners with Self-Verification.

Zhang, T., Yuan, J. & Avestimehr, S. (2024). Revisiting OPRO: The Limitations of Small-Scale LLMs as Optimizers.

Zhang, Y. (2025). A Retrieval-augmented Generation Framework with Retriever and Generator Modules for Enhancing Factual Consistency. Applied and Computational Engineering, 166(1), 149–155. https://doi.org/10.54254/2755-2721/2025.TJ24496