



Rancang Bangun Virtual Boiler Assistant Berbasis RAG Menggunakan Large Language Model untuk Taruna Politeknik Pelayaran

Dirhamsyah^{1*}, Tika Septia¹

¹Program Studi Teknologi Teknik Mesin, Politeknik Pelayaran Surabaya, Jawa Timur, Indonesia

*Corresponding Author's e-mail: dirhamsyah@poltekpel-sby.ac.id

Article History:

Received: July 12, 2026

Revised: July 22, 2026

Accepted: July 31, 2026

Keywords:

RAG; Virtual Assistant;
Boiler; NLP; Maritime
Education

Abstract: This study presents the design and development of a Retrieval-Augmented Generation (RAG)-based Virtual Boiler Assistant (VBA) integrated with a Large Language Model (LLM) to support boiler-system learning for cadets at Politeknik Pelayaran Surabaya. The system uses FastAPI as the backend framework, ChromaDB as a local vector database, Gemini 2.5 Flash as the generative LLM, and React as the frontend interface. A total of 192 text chunks were extracted from four official training documents used in the steam plant laboratory and Computer-Based Training (CBT) simulator. For each query, the RAG pipeline retrieves the four most semantically relevant chunks using cosine similarity over Gemini embeddings, then injects that context into the LLM prompt to produce accurate responses with source citations. The system was deployed on a VPS with Server-Sent Events (SSE) streaming, achieving an average response time of 15 seconds and an average semantic-similarity score of 0.73. Accuracy was measured using 20 validated ground-truth questions covering four domains: startup procedures, fuel systems, safety mechanisms, and monitoring software. Each response was assessed with a three-point rubric: 0 for incorrect answers, 1 for partially correct answers, and 2 for fully correct answers, all compared against reference answers derived directly from institutional teaching documents. The system achieved 30 of 40 points, or 75%, exceeding the minimum threshold of 70% set at the start of the study. These results indicate that a RAG-based virtual assistant, at least at this testing scale, can serve as a practical supplementary learning tool for cadets in maritime technical training.

Copyright © 2026, The Author(s).

This is an open access article under the CC-BY-SA license



How to cite: Dirhamsyah, D., & Septia, T. (2026). Rancang Bangun Virtual Boiler Assistant Berbasis RAG Menggunakan Large Language Model untuk Taruna Politeknik Pelayaran. *SENTRI: Jurnal Riset Ilmiah*, 5(7), 3588–3598. <https://doi.org/10.55681/sentri.v5i7.6992>

PENDAHULUAN

Bagi institusi pendidikan perwira pelaut di Indonesia, Standards of Training, Certification and Watchkeeping for Seafarers (STCW) merupakan standar kompetensi yang wajib dipenuhi dalam proses pendidikan dan pelatihan. International Maritime Organization (IMO) menetapkan konvensi ini sebagai acuan global bagi penyelenggaraan pendidikan pelaut profesional sehingga setiap institusi pendidikan pelayaran harus memastikan lulusannya memiliki kompetensi sesuai standar internasional. Dalam konteks tersebut, Politeknik Pelayaran Surabaya sebagai institusi di bawah Kementerian

Perhubungan memiliki tanggung jawab untuk menghasilkan perwira pelaut yang mampu menguasai kompetensi teknis, termasuk pengoperasian sistem boiler kapal. Penguasaan boiler menjadi kompetensi penting karena sistem ini menyediakan uap bertekanan yang mendukung berbagai operasi utama kapal, seperti pemanasan bahan bakar, sistem bantu mesin, dan penggerak turbin uap (Woodruff et al., 2017).

Meskipun demikian, proses pembelajaran boiler masih menghadapi berbagai keterbatasan. Praktikum laboratorium dilaksanakan dengan jadwal yang padat, sementara jumlah taruna relatif lebih besar dibandingkan kapasitas peralatan yang tersedia. Di sisi lain, buku ajar belum mampu memberikan pengalaman belajar prosedural secara kontekstual, terutama untuk menjelaskan tahapan operasional boiler yang kompleks. Akibatnya, taruna memerlukan media pembelajaran yang mampu memberikan penjelasan interaktif, adaptif, dan mudah diakses di luar jam praktikum sehingga proses belajar dapat berlangsung secara mandiri (Primo Nautic, 2024).

Perkembangan Natural Language Processing (NLP), khususnya Large Language Models (LLMs) berbasis arsitektur Transformer, membuka peluang baru dalam pengembangan media pembelajaran interaktif. Salah satu pendekatan yang banyak digunakan adalah Retrieval-Augmented Generation (RAG) yang menggabungkan kemampuan model bahasa dalam menghasilkan jawaban dengan proses pencarian informasi dari basis pengetahuan eksternal. Pendekatan ini memungkinkan jawaban yang dihasilkan tidak hanya bersifat generatif, tetapi juga didasarkan pada dokumen sumber yang relevan sehingga meningkatkan akurasi dan keterlacakan informasi (Lewis et al., 2020; Vaswani et al., 2017).

Berbagai penelitian menunjukkan bahwa penerapan RAG mampu meningkatkan faktualitas jawaban serta mengurangi risiko *hallucination* yang masih sering ditemukan pada LLM konvensional. Selain itu, penggunaan dokumen sebagai sumber pengetahuan eksternal menjadikan sistem lebih dapat dipertanggungjawabkan, terutama pada domain yang memerlukan ketepatan informasi seperti pendidikan teknis dan vokasi (Borgeaud et al., 2022; OpenAI et al., 2023). Namun demikian, keberhasilan sistem RAG tidak hanya bergantung pada model bahasa yang digunakan, tetapi juga ditentukan oleh kualitas *embedding*, mekanisme *retrieval*, serta relevansi dokumen yang menjadi sumber pengetahuan sehingga konteks yang diberikan kepada model benar-benar sesuai dengan kebutuhan pengguna (Ji et al., 2023).

Pemanfaatan *virtual assistant* berbasis NLP telah banyak diterapkan dalam bidang pendidikan dan menunjukkan hasil yang positif dalam meningkatkan akses terhadap informasi serta pengalaman belajar mahasiswa. Akan tetapi, sebagian besar penelitian masih berfokus pada domain umum, sedangkan implementasi pada bidang pendidikan kemaritiman, khususnya pembelajaran sistem boiler kapal menggunakan Bahasa Indonesia, masih sangat terbatas. Kondisi tersebut menunjukkan adanya kebutuhan terhadap pengembangan sistem yang mampu memberikan informasi teknis secara akurat, kontekstual, dan sesuai dengan karakteristik pembelajaran taruna pelayaran.

Berdasarkan kesenjangan tersebut, penelitian ini merancang dan membangun Virtual Boiler Assistant (VBA) berbasis Retrieval-Augmented Generation (RAG) dengan memanfaatkan Large Language Model (LLM) yang dikalibrasi menggunakan dokumen operasional resmi laboratorium *steam plant* dan simulator CBT Politeknik Pelayaran Surabaya. Kebaruan penelitian ini terletak pada pengembangan sistem RAG berbahasa Indonesia yang secara khusus dirancang untuk domain boiler kapal sebagai media pembelajaran interaktif bagi taruna politeknik pelayaran

METODE PENELITIAN

Skema Research and Development (R&D) dipakai sebagai tulang punggung metodologi. Pilihan ini diambil agar iterasi perbaikan dapat berjalan secara luwes. Praktiknya sendiri terbagi ke dalam empat langkah kunci—mencakup analisis kebutuhan, konstruksi knowledge base, tahapan implementasi, lalu ditutup dengan evaluasi.

Pada tahapan analisis kebutuhan mulai dirancang Virtual Boiler Assistant. Virtual Boiler Assistant dibangun dengan arsitektur RAG yang terdiri dari empat komponen utama. Pada sisi backend, sistem dikembangkan dengan FastAPI (Python 3.12), yang menyediakan endpoint REST API untuk menerima pertanyaan dan mengirimkan respons. Penyimpanan vektor memakai ChromaDB yang berjalan secara lokal, dengan fungsi utama menyimpan embedding dari knowledge base. LLM generatif menggunakan Gemini 2.5 Flash (Google AI) melalui API. Frontend dikembangkan menggunakan React 18 dengan TypeScript dan Tailwind CSS, menampilkan antarmuka bertema Marine Control Panel.

Tahapan *knowledge base* dibangun dari empat dokumen operasional resmi Politeknik Pelayaran Surabaya: (1) Petunjuk Pengoperasian Monitoring Boiler dari laboratorium Steam Plant (PT Tecno Multi Utama), (2) Modul CBT Pengoperasian Mesin Kapal, (3) Modul Pengoperasian Boiler Fuel System, dan (4) Modul CBT Engine Cargo Ship. Dokumen-dokumen ini dikonversi ke format PDF kemudian dipecah menjadi 192 potongan teks (chunks) menggunakan teknik chunking dengan mempertimbangkan koherensi konteks. Setiap chunk di-embedding menggunakan model Gemini Embedding-001 dan disimpan di ChromaDB.

Tahap ketiga terdiri dari *Pipeline RAG* dan *deployment*. Ketika pengguna mengajukan pertanyaan, sistem melakukan pencarian semantik terhadap 192 chunks menggunakan cosine similarity antara vektor pertanyaan dan vektor-vektor chunks. Empat chunks dengan skor kemiripan tertinggi ($k=4$) dipilih sebagai konteks, kemudian disuntikkan ke dalam prompt LLM bersama pertanyaan pengguna. LLM menghasilkan respons berdasarkan konteks tersebut dengan instruksi untuk menyertakan sitasi sumber. Respons dikirim ke frontend menggunakan mekanisme streaming Server-Sent Events (SSE) sehingga teks muncul secara bertahap. Pada *Deployment*, sistem dideploy pada Virtual Private Server (VPS) Ubuntu 24.04 menggunakan PM2 sebagai process manager. Frontend di-build sebagai static files dan disajikan melalui backend FastAPI pada port 8001.

Tahapan terakhir yaitu evaluasi Akurasi. Evaluasi dilakukan menggunakan 20 pertanyaan ground truth. Butir-butir tersebut dikonstruksi berdasarkan analisis konten dari keempat dokumen sumber yang dijadikan basis knowledge base. Distribusi pertanyaan mencakup enam kategori: prosedur startup 3 butir, prosedur shutdown 1 butir, sistem bahan bakar 4 butir, mekanisme keselamatan 4 butir, komponen dan fungsi 4 butir, serta monitoring dan CBT 4 butir. Jawaban acuan untuk setiap butir diambil langsung dari dokumen ajar dan disertai nomor halaman agar sumbernya mudah ditelusuri. Masing-masing respons kemudian diberi skor dengan rubrik tiga tingkat: 0 bila jawaban keliru, 1 bila hanya benar sebagian, dan 2 bila jawabannya tepat sepenuhnya. Skor maksimal adalah 40 dengan target penelitian skor persentase minimal 70%.

HASIL DAN PEMBAHASAN

Hasil Implementasi

Sistem Virtual Boiler Assistant berhasil diimplementasikan dan dideploy secara live. Knowledge base terdiri dari 192 chunks yang diekstraksi dari 4 dokumen sumber (termasuk 1 file dummy awal yang dipertahankan sebagai referensi dasar). Sistem menggunakan Gemini Embedding-001 untuk embedding dan Gemini 2.5 Flash untuk generasi respons. Tabel 1 menyajikan spesifikasi teknis sistem yang berhasil diimplementasikan.

Tabel 1. Spesifikasi Teknis Virtual Boiler Assistant

Komponen	Spesifikasi
Backend Framework	FastAPI (Python 3.12)
LLM	Gemini 2.5 Flash
Embedding Model	Gemini Embedding-001
Vector Database	ChromaDB (lokal)
Frontend	React 18 + TypeScript + Tailwind CSS
Total Chunks	192
Chunks per Query (k)	4
Response Delivery	Streaming SSE
Server	VPS Ubuntu 24.04, PM2

Kinerja Sistem.

Setelah optimasi yang mencakup pengurangan jumlah chunks yang di-retrieve dari default ke $k=4$, aktivasi streaming SSE, dan perubahan transport Gemini ke gRPC, sistem mencapai rata-rata waktu respons 15 detik (dari sebelumnya lebih dari 90 detik yang menyebabkan timeout). Skor kesesuaian semantik rata-rata (match score) antara pertanyaan dan chunks yang di-retrieve adalah 0,73.

Hasil Evaluasi Akurasi.

Evaluasi dilakukan dengan mengirimkan 20 pertanyaan ground truth ke sistem secara programatis melalui API endpoint. Hasil evaluasi disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Akurasi Virtual Boiler Assistant

No	Kategori Pertanyaan	Jumlah	Skor	Skor Max	Persentase
1	Prosedur Startup	3	4	6	67%
2	Prosedur Shutdown	1	2	2	100%
3	Sistem Bahan Bakar	4	6	8	75%
4	Mekanisme Keselamatan	4	6	8	75%
5	Komponen dan Fungsi	4	6	8	75%
6	Monitoring dan CBT	4	6	8	75%
	Total	20	30	40	75%

Setelah diuji dengan 20 pertanyaan, sistem berhasil mengumpulkan 30 dari 40 poin—yang berarti menyentuh angka 75%. Mengingat batas bawah yang disepakati sejak awal

adalah 70%, capaian ini jelas telah melampaui ekspektasi minimum. Bila dibedah lebih jauh, ada 12 pertanyaan yang dijawab dengan presisi penuh (skor 2) dan 6 pertanyaan yang bernilai parsial (skor 1). Adapun 2 pertanyaan sisa terpaksa mendapat skor 0 murni akibat terputusnya koneksi (timeout) pada API Gemini, bukan karena keterbatasan sistem dalam menemukan jawaban. Andai kedua anomali timeout tersebut diabaikan dari kalkulasi, maka akurasi riil sistem sebenarnya melonjak ke angka 30/36 atau sekitar 83,3%. Distribusi skor per pertanyaan dapat dilihat pada Tabel 3.

Tabel 3. Distribusi Skor Evaluasi Per Pertanyaan

No	Topik Pertanyaan	Skor
Q1	Prosedur start-up boiler	0*
Q2	Fungsi gelas penduga	2
Q3	Pemeriksaan sebelum menyalakan boiler	2
Q4	Prosedur shutdown boiler	2
Q5	Penyebab ignition failure	2
Q6	Pemanasan bertahap boiler	2
Q7	Fungsi safety valve	2
Q8	Tekanan steam operasi CBT	1
Q9	Drum blowdown	1
Q10	Penanganan overpressure	2
Q11	Komponen utama boiler	1
Q12	Fungsi Boiler Fuel System	1
Q13	Fungsi feed water pump	2
Q14	Mode MANUAL vs AUTO	2
Q15	Prosedur purging furnace	2
Q16	Perbedaan DO dan HFO	2
Q17	Fungsi preheater	0*
Q18	Fungsi gas detector	2
Q19	Prosedur CBT lengkap	1
Q20	Software monitoring boiler	1
Total		30/40

Catatan: * = skor 0 disebabkan timeout API Gemini, bukan ketidakmampuan sistem.

Pembahasan

Angka 75% yang dicapai VBA melampaui target minimum penelitian (70%) dan menunjukkan bahwa sistem mampu menjawab pertanyaan teknis boiler secara memadai. Kalibrasi knowledge base terhadap dokumen operasional resmi Politeknik Pelayaran Surabaya tampak bekerja dengan baik. Melalui pendekatan RAG, sistem dapat mengambil potongan teks yang tepat lalu menggunakannya sebagai dasar respons yang faktual dan

relevan. Ada beberapa pola menarik dalam distribusi skor yang layak disoroti. Sistem tampil paling kuat pada pertanyaan prosedural yang spesifik: prosedur pemanasan bertahap (Q6, skor 2), prosedur purging (Q15, skor 2), dan perbedaan bahan bakar DO versus HFO (Q16, skor 2) semuanya mendapat skor penuh. Ini sejalan dengan kekuatan pendekatan RAG: ketika dokumen sumber memuat prosedur yang tersusun jelas, sistem dapat mengambil potongan yang relevan dan menyajikannya dengan akurat. Temuan lain yang juga penting ialah konsistensi sistem dalam menyertakan sitasi sumber. Dokumen dan halaman yang dirujuk terbukti tepat, sehingga kepercayaan pengguna terhadap respons ikut menguat. Di sisi lain, sejumlah keterbatasan juga teridentifikasi. Pertanyaan Q8 tentang tekanan steam operasi (skor 1) menunjukkan bahwa sistem tidak selalu berhasil mengambil informasi numerik spesifik meskipun data tersebut ada di knowledge base. Ini mengindikasikan bahwa semantic search berbasis embedding kurang optimal untuk retrieval angka-angka spesifik dibandingkan teks deskriptif. Pertanyaan Q11 dan Q12 (masing-masing skor 1) menunjukkan bahwa pertanyaan yang membutuhkan respons panjang dan komprehensif cenderung terpotong sebelum selesai, sehingga optimasi pada mekanisme streaming atau pembatasan panjang respons masih diperlukan. Dua pertanyaan yang mendapat skor 0 (Q1 dan Q17) disebabkan oleh timeout pada API Gemini, bukan ketidakmampuan sistem. Analisis log menunjukkan bahwa pertanyaan prosedural panjang memicu retrieval yang berat dan generasi respons yang panjang, melebihi batas waktu yang ditetapkan. Dalam implementasi produksi, masalah ini dapat diatasi dengan meningkatkan timeout, mengoptimalkan prompt untuk membatasi panjang respons, atau menggunakan mekanisme retry otomatis. Dibandingkan dengan penelitian terdahulu, akurasi 75% yang dicapai VBA berada dalam rentang yang kompetitif. Wahyuwono et al. (2024) melaporkan akurasi 97,76% untuk chatbot LSTM, namun domain pengetahuannya berupa FAQ akademik yang jauh lebih sederhana dibandingkan prosedur teknis boiler. Hafizh (2024) mencapai akurasi 0,84 untuk domain akademik dan keuangan. Perbandingan langsung memang tidak sederhana, sebab domain, metode evaluasi, dan kompleksitas pertanyaan berbeda satu sama lain. Meski begitu, skor 75% pada domain teknis yang spesifik dan kompleks tetap menunjukkan kinerja yang baik. Keunikan penelitian ini terletak pada penggunaan dokumen operasional asli dari laboratorium dan simulator CBT Politeknik Pelayaran Surabaya sebagai knowledge base. Berbeda dari chatbot edukasi umum yang bertumpu pada pengetahuan generik LLM, VBA secara eksplisit menjangkar setiap jawaban pada materi pembelajaran resmi institusi, sehingga kesesuaiannya dengan kurikulum dan standar STCW dapat terjaga. Dengan cara itu, taruna menerima informasi yang tetap selaras dengan apa yang mereka pelajari di kelas maupun di laboratorium.

KESIMPULAN

Penelitian ini berhasil merancang dan membangun Virtual Boiler Assistant berbasis RAG menggunakan LLM Gemini 2.5 Flash yang diintegrasikan dengan knowledge base dari dokumen operasional resmi Politeknik Pelayaran Surabaya. Sistem yang dikembangkan memiliki 192 chunks pengetahuan dari 4 dokumen sumber, menggunakan ChromaDB sebagai vector database lokal, dan menyajikan respons melalui streaming SSE dengan rata-rata waktu respons 15 detik. Evaluasi akurasi dengan 20 pertanyaan ground truth tervalidasi menghasilkan skor 30 dari 40 (75%), sehingga target minimum penelitian sebesar 70% terlampaui. Sistem tampak paling kuat pada pertanyaan prosedural yang spesifik dan juga berhasil memberikan sitasi sumber yang akurat. Keterbatasan utama

teridentifikasi pada retrieval informasi numerik spesifik dan penanganan respons panjang yang memerlukan optimasi lebih lanjut. Hasil penelitian ini menunjukkan bahwa teknologi RAG dengan LLM modern dapat secara efektif diaplikasikan sebagai alat bantu pembelajaran domain teknis spesifik dalam pendidikan maritim. Pengembangan berikutnya sebaiknya memperluas knowledge base dengan manual boiler kapal niaga, menguji usability melalui SUS bersama taruna, dan menambahkan evaluasi otomatis berbasis RAGAS untuk metrik faithfulness serta answer relevancy.

DAFTAR REFERENSI

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264-75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186.
- Gao, Y., Xiong, Y., Gong, X., Jia, W., Wang, J., Bi, Y., ... & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Hafizh, M. (2024). Pengembangan chatbot menggunakan metode SEMMA dan model Transformer untuk layanan akademik. *Jurnal Teknik Informatika*, 12(3), 45-56.
- Hamzani, A., et al. (2024). Implementasi chatbot berbasis AI dan NLP untuk proses penerimaan mahasiswa baru. *Jurnal Pendidikan Teknologi Informasi*, 8(2), 112-125.
- International Maritime Organization. (2017). *International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW)*. IMO Publishing.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP*, 6769-6781.
- Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Marine Insight Academy. (2025). *Maritime AI Assistant for COC preparation*. <https://www.marineinsight.com>
- Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427-469. <https://doi.org/10.1007/s40593-014-0029-5>
- Primo Nautic. (2024). *Maritime GPT: AI model for maritime education*. *Maritime Technology Review*, 15(1), 22-31.
- PT Tecno Multi Utama. (n.d.). *Petunjuk pengoperasian monitoring boiler: Laboratorium steam plant*. Politeknik Pelayaran Surabaya.
- Puspitasari, D., et al. (2024). Pengembangan chatbot website menggunakan NLP dengan algoritma Neural Network. *Jurnal Informatika*, 11(1), 78-89.

- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982-3992.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- Wahyuwono, A., Oktavia, R., & Kartikasari, D. (2024). Virtual assistant berbasis web menggunakan NLP dan LSTM untuk layanan informasi akademik. *Jurnal Teknologi Pendidikan*, 9(2), 201-215.
- Woodruff, E. B., Lammers, H. B., & Lammers, T. F. (2017). *Steam plant operation* (10th ed.). McGraw-Hill Education.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3), Article ep429. <https://doi.org/10.30935/cedtech/13152>
- Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, 3, Article 60. <https://doi.org/10.1007/s44217-024-00138-2>
- Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in teaching and learning: A systematic review. *Education Sciences*, 14(6), Article 643. <https://doi.org/10.3390/educsci14060643>
- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations*. <https://openreview.net/forum?id=hSyW5go0v8>
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., et al. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206-2240). PMLR.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. <https://arxiv.org/abs/2303.12712>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, Article 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20, Article 22. <https://doi.org/10.1186/s41239-023-00392-8>
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. <https://arxiv.org/abs/2407.21783>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI. *International Journal of Information Management*, 71, Article 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>

- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158). <https://aclanthology.org/2024.eacl-demo.16/>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, 66, 111-126. <https://doi.org/10.1007/s12599-023-00834-7>
- Google DeepMind, Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., et al. (2024). Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. <https://arxiv.org/abs/2312.11805>
- Guizani, S., Mazhar, T., Shahzad, T., Ahmad, W., Bibi, A., & Hamam, H. (2025). A systematic literature review to implement large language model in higher education: Issues and solutions. Discover Education, 4, Article 35. <https://doi.org/10.1007/s44217-025-00424-7>
- Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929-3938). PMLR.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 42. <https://doi.org/10.1145/3703155>
- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 874-880). <https://doi.org/10.18653/v1/2021.eacl-main.74>
- Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., et al. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1-43.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Li, J., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6449-6464). <https://doi.org/10.18653/v1/2023.emnlp-main.397>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), Article 410. <https://doi.org/10.3390/educsci13040410>
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 9802-9822). <https://doi.org/10.18653/v1/2023.acl-long.546>
- Mialon, G., Dessi, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., et al. (2023). Augmented language models: A survey. Transactions on Machine Learning Research. <https://openreview.net/forum?id=jh7wH2AzKK>

- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076-12100). <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*. <https://arxiv.org/abs/2112.09332>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*. <https://arxiv.org/abs/2303.08774>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
- Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11, 1316-1331. https://doi.org/10.1162/tacl_a_00605
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784-3803). <https://doi.org/10.18653/v1/2021.findings-emnlp.320>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, Article 15. <https://doi.org/10.1186/s40561-023-00237-x>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. <https://arxiv.org/abs/2302.13971>
- UNESCO. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21, Article 15. <https://doi.org/10.1186/s41239-024-00448-3>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90-112. <https://doi.org/10.1111/bjet.13370>
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., et al. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*. <https://arxiv.org/abs/2309.01219>

