

IMPLEMENTASI RETRIEVAL AUGMENTED GENERATION UNTUK EVALUASI PROPOSAL TUGAS AKHIR MAHASISWA

Ikhsan Fanani¹

¹Program Studi Teknologi Komputer, Politeknik Pajajaran ICB Bandung, Indonesia
ikhsan.fanani@poljan.ac.id¹

Abstract

This research aims to develop a thesis title evaluation system based on Retrieval Augmented Generation (RAG), utilizing the thesis repository from the past year as a knowledge source. The system is developed by integrating a retrieval component that employs semantic embedding techniques to identify similar titles from the repository and a generative component that provides evaluation and improvement recommendations. The process includes preprocessing data from the thesis repository, implementing a sentence-transformers model to create a vector database, and integrating it with a Large Language Model (LLM). The test results on 20 new titles showed that the RAG system achieved an answer correctness score of 80%. The implementation also succeeded in automating and improving the objectivity of the evaluation process.

Keywords: *Retrieval Augmented Generation; Large Language Model; Thesis title evaluation; Sentence Embedding; Natural Language Processing;*

Abstrak

Penelitian ini bertujuan mengembangkan sistem evaluasi judul tugas akhir berbasis *Retrieval Augmented Generation* (RAG) yang memanfaatkan *repository* tugas akhir dari satu tahun terakhir sebagai sumber pengetahuan. Sistem dikembangkan dengan mengintegrasikan komponen *retrieval* yang menggunakan teknik *embedding* semantik untuk mengidentifikasi judul-judul serupa dari repositori, dan komponen generative yang memberikan evaluasi serta rekomendasi perbaikan. Proses yang dilakukan mencakup pre-processing data dari repository tugas akhir, implementasi model *sentence-transformers* untuk pembuatan basis data vektor, dan integrasi dengan *Large Language Model* (LLM). Hasil pengujian terhadap 20 judul baru menunjukkan sistem RAG mencapai nilai *answer correctness* 80%. Implementasi juga berhasil mengotomasi dan meningkatkan objektivitas proses evaluasi.

Kata kunci : *Retrieval Augmented Generation; evaluasi judul tugas akhir; large language model; sentence embedding; natural language processing;*

Corresponding author : ikhsan.fanani@poljan.ac.id

PENDAHULUAN

Evaluasi proposal tugas akhir merupakan tahap krusial dalam proses akademik mahasiswa, karena menjadi penentu kualitas penelitian dan kemampuan analitis yang akan dikembangkan (Vasanthakumari, 2021). Namun, pada praktiknya, evaluasi proposal sering kali menghadapi tantangan seperti keterbatasan waktu dosen, subjektivitas penilaian (Jukola, 2017), serta kesulitan mengakses referensi terkini yang relevan dengan topik proposal. Metode konvensional yang mengandalkan pemeriksaan manual oleh dosen pembimbing atau tim *reviewer* berisiko menghasilkan inkonsistensi, terutama ketika jumlah proposal meningkat signifikan. Di sisi lain, perkembangan teknologi kecerdasan buatan, khususnya dalam bidang *Natural Language Processing* (NLP), menawarkan peluang untuk mengoptimalkan proses ini melalui integrasi sistem yang mampu menggabungkan kemampuan pencarian informasi (*retrieval*) dan memberikan respons yang kontekstual (*generation*). Salah satu pendekatan inovatif dalam hal ini adalah *Retrieval Augmented Generation* (RAG), yang memadukan basis pengetahuan eksternal dengan *Large Language Model* (LLM) untuk meningkatkan akurasi dan relevansi *output* (Lewis et al., 2021).

Penelitian ini bertujuan mengimplementasikan RAG sebagai alat evaluasi proposal tugas akhir mahasiswa guna mengatasi masalah ketidakefisienan dan subjektivitas dalam metode evaluasi konvensional. Dengan memanfaatkan basis data tugas akhir sebelumnya, sistem RAG diharapkan mampu memberikan umpan balik objektif, serta memberikan rekomendasi perbaikan berbasis

data. Urgensi penelitian ini terletak pada potensinya untuk meningkatkan kualitas pengawasan akademik di perguruan tinggi, khususnya dalam konteks sumber daya manusia yang terbatas dan tuntutan publikasi penelitian yang semakin kompetitif. Selain itu, implementasi RAG dalam evaluasi proposal tugas akhir menjadi langkah awal yang untuk mengintegrasikan teknologi AI secara dalam ekosistem pendidikan tanpa menggantikan peran dosen, melainkan memperkuat kapasitas evaluasi mereka.

Kontribusi hasil penelitian ini mencakup dua aspek utama: pengembangan keilmuan dan solusi praktis. Secara akademik, studi ini memperluas wawasan tentang aplikasi RAG di luar domain umum seperti chatbot atau pencarian informasi. Mengeksplorasi potensinya dalam konteks evaluasi formatif pendidikan tinggi. Pada tataran praktis, sistem yang dihasilkan dapat diadopsi oleh institusi pendidikan sebagai alat pendamping evaluasi proposal, mengurangi beban administratif, sekaligus memastikan konsistensi dan transparansi penilaian. Dengan demikian, penelitian ini tidak hanya menjawab kebutuhan teknis dalam proses akademik tetapi juga mendorong transformasi digital yang berkelanjutan di lingkungan perguruan tinggi, menjembatani kesenjangan antara inovasi teknologi dan praktik pendidikan yang efektif.

KAJIAN PUSTAKA

Language Model

Language Model atau model bahasa merupakan salah satu jenis kecerdasan buatan yang dirancang untuk mempelajari pola dan struktur bahasa manusia secara statistik dari sejumlah besar data teks (Wei et al., 2024).

Model bahasa pada dasarnya mencoba memecahkan masalah mendasar untuk memprediksi kemungkinan urutan kata dan menghasilkan teks yang koheren.

Kemampuan model bahasa modern saat ini telah jauh melampaui prediksi kata berikutnya yang sederhana. Model ini juga mampu menghasilkan teks yang koheren dan panjang, menjawab pertanyaan, meringkas dokumen, menerjemahkan antar bahasa, menulis konten kreatif, serta memahami konteks dan nuansa kalimat. Hal ini dimungkinkan dengan berkembangnya volume data latih serta arsitektur yang digunakan untuk membangun model bahasa tersebut.

Transformer

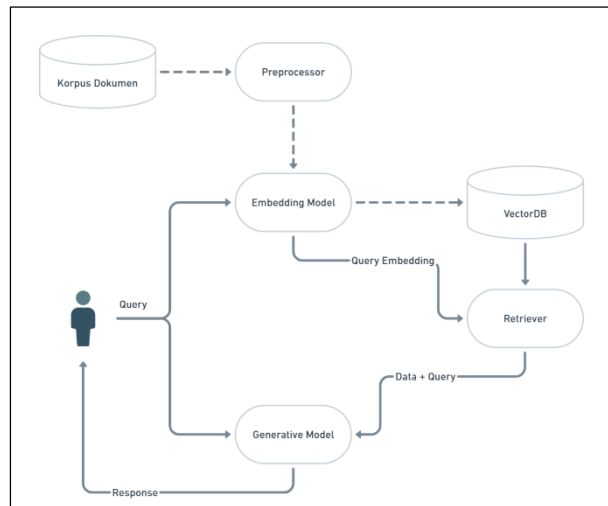
Arsitektur Transformer, yang diperkenalkan pada makalah (Vaswani, 2023) merevolusi pemodelan bahasa. Terobosan *transformer* adalah *attention mechanism* yang memungkinkan model untuk fokus pada bagian-bagian berbeda dari teks input saat membuat prediksi. Tidak seperti model *sequential* seperti RNN dan LSTM, *transformer* dapat melihat seluruh urutan sekaligus dan menentukan bagian mana yang paling relevan untuk memahami setiap kata.

Transformer mengubah pelatihan model bahasa melalui beberapa keunggulan. Kemampuan pemrosesan paralel seluruh input yang mempercepat proses pelatihan secara drastis (berbeda dengan RNN yang berurutan), *attention mechanism* yang unggul dalam menangani hubungan antar kata yang berjauhan dalam teks, serta skalabilitas yang mampu menghasilkan performa lebih baik dengan melakukan peningkatan parameter (Vaswani, 2023). Hal ini memungkinkan pengembangan model bahasa berukuran besar

seperti yang saat ini dikenal sebagai *Large Language Model* (LLM).

Vector Embedding

Vector Embedding merupakan salah satu cara untuk mengubah teks, kalimat, paragraf, gambar, atau jenis data lainnya menjadi vektor numerik yang mengindikasikan semantik dari



Gambar 1. Retrieval Augmented Generation

data tersebut dalam dimensi matematika multidimensional (Mikolov et al., 2013). Vektor-vektor ini memungkinkan komputer untuk menghitung relasi semantik antar berbagai data tidak terstruktur tersebut. Proses konversi dari teks menjadi vektor dapat dilakukan dengan memanfaatkan model bahasa.

Setelah konversi menjadi vektor dilakukan, kedekatan semantik dapat dihitung dengan menghitung jarak/kedekatan antar vektor matematika biasa menggunakan *cosine similarity*, *euclidean distance*, maupun *dot product*.

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) adalah pendekatan inovatif dalam teknologi kecerdasan buatan yang menggabungkan kekuatan *Large Language Model* (LLM) dengan kemampuan pencarian informasi eksternal (Lewis et al., 2021). Sistem RAG bekerja dengan mengambil informasi yang relevan dari basis pengetahuan atau basis data eksternal sebelum menghasilkan respons, sehingga dapat memberikan jawaban yang lebih akurat dan sesuai dengan konteks.

Berbeda dengan LLM tradisional yang hanya dapat mengandalkan pengetahuan yang tertanam dalam data latih-nya, RAG memungkinkan LLM untuk mengakses dan memanfaatkan sumber informasi tambahan tanpa perlu pelatihan ulang. Hal ini sekaligus mengurangi masalah halusinasi dan meningkatkan transparansi melalui kemampuannya untuk mengutip sumber informasi (Lewis et al., 2021).

METODE PENELITIAN

RAG terdiri dari beberapa komponen utama. Pertama, basis pengetahuan atau korpus dokumen yang akan menjadi sumber informasi eksternal. Kedua, komponen *embedding model* yang mengubah dokumen dan *query* dari pengguna menjadi representasi vektor untuk memfasilitasi pencarian semantik. Ketiga, *retriever* yang berperan mengambil dokumen atau potongan informasi yang paling relevan dengan *query* menggunakan metode *similarity search*. Keempat, komponen *context builder* menyusun informasi yang diperoleh menjadi konteks terstruktur. Kelima, *generative model* berupa LLM yang menerima kombinasi *query*

pengguna dan konteks hasil *retrieval* untuk menghasilkan respons yang informatif. Terakhir, banyak sistem RAG modern juga dilengkapi dengan modul evaluasi yang mengukur kualitas dan akurasi respons, serta komponen pengelolaan data untuk memperbarui dan memelihara basis pengetahuan.

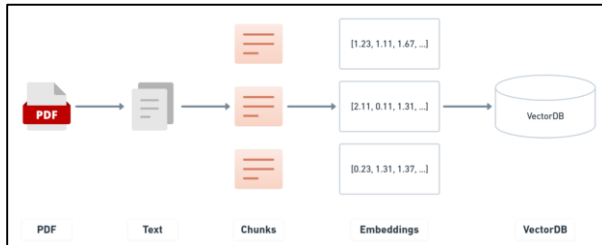
Data Eksternal

Data eksternal dalam arsitektur RAG berfungsi sebagai basis pengetahuan tambahan di luar parameter LLM. Data eksternal dapat berupa dokumen terstruktur maupun tidak terstruktur yang terkait dengan sistem yang akan dibangun. RAG mengandalkan data eksternal untuk membatasi proses *retrieval* dan mengurangi risiko halusinasi yang umum terjadi pada model generatif konvensional. Dengan mengintegrasikan data eksternal, sistem RAG dapat menghasilkan evaluasi proposal yang konsisten, objektif, dan transparan.

Pada penelitian ini, data eksternal yang digunakan sebagai korpus dokumen berupa 106 dokumen tugas akhir mahasiswa program studi teknologi komputer dengan format pdf.

Proses *Embedding*

Proses pengolahan data eksternal melibatkan tahapan *pre-processing* seperti *chunking* (pemecahan dokumen menjadi segmen yang lengkap secara semantik), *embedding*, dan *pengindeksan* untuk memfasilitasi proses pencarian yang efisien.



Gambar 2. Proses Embedding

Proses *embedding* merupakan proses untuk mentransformasi representasi tekstual menjadi vektor dalam representasi numerik multidimensional (Huyen, 2025). Tahap ini dilakukan menggunakan sebuah *embedding model*, yaitu model bahasa yang dilatih secara khusus untuk menilai relasi semantik dari teks maupun dokumen. Keluaran dari model bahasa ini adalah vektor dalam format numerik multidimensi. Dengan menggunakan *embedding model* ini, kata-kata yang memiliki kedekatan secara makna akan ditempatkan pada ruang vektor yang berdekatan.

Saat pengguna melakukan *query*, teks *query* juga diproses melalui *embedding model* yang sama untuk menghasilkan representasi vektor dari *query*. Kesesuaian representasi *embedding* antara *chunk* data eksternal dan *query* menjadi faktor penentu dalam efektivitas sistem *retrieval*. Selanjutnya, sistem menghitung kedekatan semantik antara vektor *query* dan vektor *chunk* dokumen eksternal pada basis data vektor menggunakan metrik seperti *cosine similarity* untuk mengambil beberapa *chunk* yang paling relevan (Riyani et al., 2019).

Kualitas dan kemampuan *embedding model* yang digunakan sangat memengaruhi performa keseluruhan sistem RAG, sehingga pemilihan *embedding model* yang tepat dan *fine-tuning* pada domain akademik secara spesifik menjadi pertimbangan penting dalam

pengembangan sistem evaluasi proposal tugas akhir yang efektif dan akurat. Karena teks pada dokumen tugas akhir umumnya menggunakan bahasa Indonesia dan bahasa Inggris, maka pada penelitian ini perlu digunakan *embedding model* yang multi-bahasa.

Proses *Retrieval* dan *Augmentasi*

Setelah melakukan proses *embedding* pada *query* dari pengguna, sistem *retriever* akan melakukan pengurutan dan mengambil beberapa *chunk* dokumen yang memiliki tingkat relevansi paling tinggi dengan vektor *query* dari pengguna.

Data yang diperoleh dari hasil *retrieval* kemudian disusun dengan *query* dari pengguna beserta instruksi tambahan untuk menghasilkan prompt yang akan diberikan pada model generatif. Model generatif diperlukan untuk menghasilkan respons yang ringkas, akurat dan informatif

Pengujian

Pengujian sistem RAG dapat dilakukan dengan menggunakan *framework* RAGAS (*Retrieval Augmented Generation Assessment*). Pengujian RAGAS dilakukan untuk memastikan sistem berfungsi dengan baik. Ini mencakup pengujian unit untuk setiap bagian dari sistem, pengujian integrasi untuk memastikan semua komponen bekerja bersama dengan baik, dan pengujian sistem dengan pengguna akhir untuk mengevaluasi kegunaan dan kinerja keseluruhan sistem.

Dalam RAGAS, terdapat beberapa metrik yang digunakan untuk menilai kinerja sistem (Es et al., 2023). Metrik pertama adalah *context precision*, yang mengukur seberapa tepat konteks yang diambil oleh sistem dari sumber data. *Answer Correctness* menilai kebenaran

jawaban yang diberikan oleh sistem. *Context Recall* mengukur seberapa lengkap konteks yang diambil oleh sistem dari konteks yang seharusnya. *Faithfulness* mengukur seberapa akurat informasi yang diberikan oleh sistem dibandingkan dengan sumber data aslinya. *Answer Relevancy* menilai sejauh mana jawaban yang diberikan oleh sistem relevan dengan pertanyaan yang diajukan, sedangkan *Answer Similarity* mengukur tingkat kemiripan jawaban yang diberikan oleh sistem dengan jawaban yang diharapkan.

Pada penelitian ini, yang diharapkan adalah untuk menentukan jika tugas akhir yang diajukan sudah pernah dikerjakan sebelumnya. Karena itu kriteria yang akan dievaluasi adalah *Answer Correctness*.

HASIL DAN PEMBAHASAN

Implementasi dilakukan dengan menggunakan komponen dan spesifikasi seperti yang ditunjukkan pada Tabel 1.

Korpus dokumen yang digunakan adalah 106 dokumen tugas akhir mahasiswa pada Program Studi Teknologi Komputer di Politeknik Pajajaran. *Pre-processing* dilakukan terlebih dahulu agar setiap dokumen memiliki format yang seragam.

Tabel 1. Komponen Implementasi

Komponen	Keterangan
Vector DB	Chroma
Korpus Dokumen	Dokumen 106 tugas akhir mahasiswa program studi teknologi komputer
<i>Embedding Model</i>	<ul style="list-style-type: none"> intfloat/multilingual-e5-small (118M <i>params</i>, 384 <i>dimensions</i>) intfloat/multilingual-e5-large-instruct (560M <i>params</i>, 1024 <i>dimensions</i>)

	<ul style="list-style-type: none"> jinaai/jina-embeddings-v3 (572M <i>params</i>, 32-1024 <i>dimensions</i>) Alibaba-NLP/gte-multilingual-base (305M <i>params</i>, 768 <i>dimensions</i>)
<i>Generative Model</i>	<ul style="list-style-type: none"> llmware/bling-phi-3-gguf (3.82B <i>params</i>)
Data Tes	20 usulan judul proposal tugas akhir

Proses *embedding* cukup penting pada sistem RAG yang akan dibangun. Karena itu akan diuji dengan 4 *embedding model* dengan jumlah parameter dan dimensi vektor yang berbeda untuk melihat pengaruhnya pada tingkat akurasi proses *retrieval*. *Embedding model* yang digunakan dipilih dari benchmark MTEB *Leaderboard* (Muennighoff et al., 2023) peringkat 20 besar dengan penggunaan parameter relatif kecil. Vektor yang dihasilkan disimpan pada *database* Chroma. Model generatif yang digunakan adalah llmware/bling-phi-3-gguf, yaitu hasil *fine-tuning* dari LLM Microsoft phi-3.

Untuk menguji sistem, data tes yang digunakan adalah 20 usulan judul proposal tugas akhir dengan distribusi data yang (1) sama dengan data pada korpus dokumen, (2) modifikasi dari korpus namun memiliki semantik yang sama, (3) mirip dengan judul pada korpus namun memiliki fokus yang berbeda, dan (4) berbeda sama sekali dengan data dari korpus. Hasil pengujian dapat dilihat pada Tabel 2.

Tabel 2. Hasil Evaluasi

Model	Benar	Salah	<i>Answer Correctness</i>
intfloat/multilingual-e5-small	15	5	75%

intfloat/multilingual-e5-large-instruct	14	6	70%
jinaai/jina-embeddings-v3	16	4	80%
Alibaba-NLP/gte-multilingual-base	15	5	75%

Hasil pengujian menunjukkan nilai *answer correctness* 70-80% untuk masing-masing *embedding model*. Menunjukkan bahwa perbedaan *embedding model* tidak memberikan perbedaan kinerja sistem yang signifikan.

KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa penggunaan RAG dapat membantu melakukan proses evaluasi proposal tugas akhir mahasiswa. RAG memungkinkan proses evaluasi yang lebih cepat dan objektif dibandingkan dengan metode konvensional. Hasil pengujian menggunakan menunjukkan skor *answer correctness* 70%-80%,

Untuk meningkatkan efektivitas dan efisiensi layanan informasi, beberapa saran yang dapat dipertimbangkan adalah: (1) mengembangkan korpus dokumen, (2) menggunakan *embedding model* yang sudah dilakukan *fine-tuning* untuk teks-teks akademik, (3) menggunakan variasi *prompt* untuk model generatif, (4) mengembangkan fitur untuk memberikan rekomendasi judul berdasarkan kesimpulan tugas akhir sebelumnya dengan topik yang mirip, serta (5) menggunakan beberapa variasi kombinasi model generatif dan *embedding model*.

DAFTAR PUSTAKA

Anam, M. R., Akbar, A. S., & Saputro, H. (2025). QnA Chatbot with Mistral 7B and RAG method: Traffic Law Case Study.

Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 15(03), 207. <https://doi.org/10.24843/LKJITI.2024.v15.i03.p06>

Cowley, S. J. (2015). How peer-review constrains cognition: On the frontline in the knowledge sector. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01706>

Elysia, S. & Herianto. (2024). Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah. *Journal TIFDA (Technology Information and Data Analytic)*, 1(2), 52–58. <https://doi.org/10.70491/tifda.v1i2.52>

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation (No. arXiv:2309.15217). *arXiv*. <https://doi.org/10.48550/arXiv.2309.15217>

Huyen, C. (2025). *AI Engineering* (First Edition). United States of America: O'Reilly Media, Inc.

Iskandar, D., & Kurniawati, A. (2025). Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 12(1), 133–144. <https://doi.org/10.25126/jtiik.20251219143>

Jukola, S. (2017). A social epistemological inquiry into biases in journal peer review. *Perspectives on Science*, 25(1), 124–148. https://doi.org/10.1162/POSC_a_00237

Kurniawan, D., & Triloka, J. (2025). Penerapan Teknologi Langchain dan LLM pada Sistem

- Question Answering Berbasis Chatbot Telegram: Literature Review. *Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2025*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (No. arXiv:2005.11401). *arXiv*. <http://arxiv.org/abs/2005.11401>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). *arXiv*. <https://doi.org/10.48550/arXiv.1301.3781>
- Milasanti, D. (n.d.). (2024). *Sistem Chatbot Berbasis LLM dan RAG pada Artikel Ilmiah Garuda Kemdikbud* (Skripsi). Universitas Islam Sultan Agung.
- Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive Text Embedding Benchmark (No. arXiv:2210.07316). *arXiv*. <https://doi.org/10.48550/arXiv.2210.07316>
- Muhammad Irfan Syah, Nazruddin Safaat Harahap, Novriyanto, & Suwanto Sanjaya. (2024). Penerapan Retrieval Augmented Generation Menggunakan Langchain Dalam Pengembangan Sistem Tanya Jawab Hadis Berbasis Web. *ZONasi: Jurnal Sistem Informasi*, 6(2), 370–379. <https://doi.org/10.31849/zn.v6i2.19940>
- Prasetyo, E. A. (n.d.). (2024). *Chatbot Untuk Informasi Pembangunan Wilayah Kota Semarang Menggunakan Metode Retrieval Augmented Generation (RAG)* (Skripsi). Universitas Islam Sultan Agung.
- Riyani, A., Naf'an, M. Z., & Burhanuddin, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *JLK (Vol. 2, Issue 1)*.
- S .Vasanthakumari. (2021). Writing research proposal. *World Journal of Advanced Research and Reviews*, 10(1), 184–190. <https://doi.org/10.30574/wjar.2021.10.1.0138>
- Tribber, Y., & Asfi, M. (2024). Implementasi Retrieval Augmented Generation untuk Layanan Informasi Kampus dengan Chatbot Berbasis AI. *Prosiding Seminar Nasional Sistem Informasi dan Teknologi (SISFOTEK) ke 8 Tahun 2024*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). *arXiv*. <https://doi.org/10.48550/arXiv.1706.03762>
- Wei, C., Wang, Y.-C., Wang, B., & Kuo, C.-C. J. (2024). An Overview on Language Models: Recent Developments and Outlook. *APSIPA Transactions on Signal and Information Processing*, 13(2). <https://doi.org/10.1561/116.00000010>