

Optimasi Sistem Rekomendasi Pencarian Artikel Ilmiah TIK Menggunakan BERT

Irfan Hilal Yasirrizky¹

¹Universitas Amikom Purwokerto, Purwokerto, Indonesia

¹Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Purwokerto, Indonesia

irfanhilal77@gmail.com

Info Artikel

Sejarah artikel:

Diterima 23-05-2026

Revisi 26-06-2026

Diterima 10-07-2026

Kata kunci:

Sistem Rekomendasi

BERT

TF/IDF

Information Retrieval

Vector Space Model

ABSTRAK

Sistem rekomendasi artikel ilmiah tradisional menghadapi kendala dalam memahami konteks semantik dari query pencarian pengguna. Penelitian ini bertujuan mengembangkan sistem rekomendasi pencarian artikel ilmiah bidang Teknologi Informasi dan Komunikasi menggunakan model BERT untuk meningkatkan akurasi pencarian berbasis pemahaman semantik. Metode yang digunakan meliputi web scraping Portal Garuda untuk mengumpulkan 2345 artikel unik, preprocessing menggunakan BERT multilingual, pembentukan Vector Space Model, dan perhitungan similarity menggunakan Cosine Similarity. Sistem dievaluasi menggunakan metrik Precision, Recall, dan F-Measure. Hasil pengujian menunjukkan sistem mampu mencapai Precision 100% dengan F-Measure 76.92% untuk query "aplikasi kesehatan" dan F-Measure 90.91% untuk query "AI". Sistem berhasil memahami konteks semantik tanpa bergantung pada pencocokan kata kunci literal, mampu mengidentifikasi artikel relevan dari berbagai subdomain, dan menampilkan hasil yang beragam namun tetap relevan. Penelitian ini membuktikan efektivitas BERT dalam meningkatkan kualitas sistem rekomendasi artikel ilmiah berbasis pemahaman semantik kontekstual.

Penulis yang sesuai:

Irfan Hilal Yasirrizky

Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto

Email: irfanhilal77@gmail.com

1. PENDAHULUAN

Perkembangan publikasi artikel ilmiah di Indonesia mengalami peningkatan signifikan dalam beberapa tahun terakhir, khususnya di bidang Teknologi Informasi dan Komunikasi. Portal Garuda sebagai repositori nasional menyimpan ribuan artikel ilmiah yang dapat diakses secara terbuka. Namun, volume data yang besar ini menimbulkan tantangan baru dalam menemukan artikel yang relevan dengan kebutuhan penelitian pengguna. Sistem pencarian konvensional umumnya mengandalkan metode pencocokan kata kunci literal yang memiliki keterbatasan dalam memahami konteks dan makna semantik dari query pengguna [1]. Pendekatan tradisional seperti TF-IDF dan N-Gram hanya mampu menangkap frekuensi kemunculan kata tanpa

mempertimbangkan hubungan kontekstual antar kata. Akibatnya, artikel yang relevan secara semantik namun tidak mengandung kata kunci yang sama persis sering terlewatkan dalam hasil pencarian.

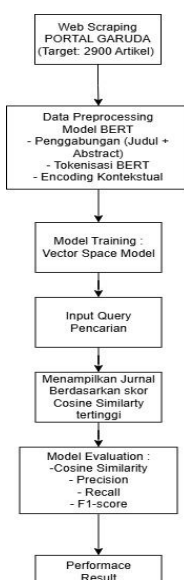
Pada penelitian sebelumnya telah menggunakan model tradisional untuk preprocessing seperti TF-IDF dan N-Gram dalam sistem rekomendasi pencarian artikel TIK [2]. Meskipun metode tersebut cukup efektif untuk pencocokan kata kunci, namun kurang optimal dalam menangkap makna semantik yang mendalam dari teks artikel ilmiah. Bidirectional Encoder Representations from Transformers atau BERT merupakan model pembelajaran mendalam yang dikembangkan untuk memahami konteks bahasa secara bidirectional [3]. Model ini telah menunjukkan keunggulan dalam berbagai tugas pemrosesan bahasa alami termasuk pencarian informasi berbasis semantik [4]. BERT mampu menghasilkan representasi vektor yang menangkap makna kontekstual dari teks, bukan sekadar frekuensi kata [5].

Penelitian terkini menunjukkan bahwa penerapan BERT dalam sistem information retrieval memberikan peningkatan signifikan terhadap pemahaman semantik dibandingkan metode konvensional [6]. Studi oleh Meilinger et al. (2022) membuktikan bahwa penggunaan Sentence Transformer multilingual dapat meningkatkan kualitas pencarian semantik pada dataset dengan metadata tekstual [7]. Sementara itu, Wu (2024) mendemonstrasikan bahwa sistem rekomendasi berbasis BERT dengan k-nearest neighbors mampu menghasilkan embedding semantik yang bermakna untuk data teks tidak terstruktur [8].

Penelitian ini mengusulkan pengembangan sistem rekomendasi pencarian artikel ilmiah dengan mengimplementasikan model BERT untuk menggantikan metode preprocessing tradisional. Pendekatan yang diusulkan meliputi penggabungan judul dan abstrak, tokenisasi menggunakan BERT multilingual, dan encoding menjadi vektor semantik. Sistem kemudian menggunakan Vector Space Model dan Cosine Similarity untuk mengukur kemiripan antara query pengguna dengan artikel dalam database [9]. Nilai inovasi dari penelitian ini terletak pada penerapan BERT embedding untuk meningkatkan pemahaman semantik dalam sistem rekomendasi artikel ilmiah berbahasa Indonesia di bidang TIK, yang berbeda dari pendekatan tradisional berbasis frekuensi kata. Penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan efektivitas pencarian literatur ilmiah dan membantu peneliti menemukan artikel relevan dengan lebih efisien.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode pengembangan sistem berbasis model pembelajaran mendalam.



Gambar 1. Alur Kerja Sistem Rekomendasi

Keseluruhan proses penelitian terdiri dari lima tahapan utama yaitu pengumpulan data, preprocessing data, pembentukan model, implementasi sistem pencarian, dan evaluasi kinerja sistem.

2.1. Pengumpulan Data

Pengumpulan data dilakukan melalui teknik web scraping otomatis menggunakan skrip Python dari Portal Garuda Kementerian Pendidikan dan Kebudayaan. Portal ini dipilih karena merupakan repositori nasional dengan sistem akses terbuka yang fokus pada publikasi karya ilmiah Indonesia. Data yang dikumpulkan difilter untuk memprioritaskan artikel terkait Teknologi Informasi dan Komunikasi mencakup 29 topik seperti Information Retrieval, Kecerdasan Buatan, Pembelajaran Mesin, dan Sistem Informasi. Target pengumpulan data adalah 2900 artikel dengan 100 artikel untuk setiap topik. Proses scraping menghasilkan 2643 artikel mentah yang kemudian melalui tahap pembersihan duplikasi. Setelah eliminasi 298 artikel duplikat, diperoleh dataset bersih sebanyak 2345 artikel unik. Setiap artikel mengandung atribut Judul, Abstrak, dan Link sumber yang kemudian diproses lebih lanjut.

2.2. Preprocessing Data dengan BERT

Tahap preprocessing dalam penelitian ini dirancang khusus untuk memaksimalkan input bagi model BERT, berbeda dari pendekatan tradisional. Proses preprocessing meliputi tiga tahapan utama. Pertama adalah penggabungan teks di mana kolom Judul dan Abstrak digabungkan menjadi satu kesatuan teks lengkap [10]. Penggabungan ini bertujuan memperkaya konteks semantik dokumen sehingga model BERT dapat memahami makna artikel secara lebih komprehensif.

Kedua adalah tokenisasi BERT menggunakan AutoTokenizer dari model paraphrase-multilingual-MiniLM-L12-v2 [11]. Proses ini mengubah teks lengkap menjadi token-token sub-kata sesuai standar model BERT multilingual. Sebelum tokenisasi, dilakukan pembersihan teks ringan untuk menghilangkan karakter atau simbol yang tidak diperlukan namun tetap mempertahankan struktur kalimat. Ketiga adalah vektorisasi atau encoding di mana token-token hasil tokenisasi diproses oleh model BERT menggunakan SentenceTransformer untuk menghasilkan vektor CLS berukuran 768 dimensi [12]. Vektor ini merangkum makna semantik keseluruhan dokumen dan menjadi representasi numerik untuk setiap artikel dalam database.

2.3. Pembentukan Vector Space Model

Vector Space Model dibentuk dengan mengumpulkan seluruh vektor BERT dari 2345 artikel menjadi sebuah matriks vektor berukuran 2345 x 768 [13]. Setiap baris matriks merepresentasikan satu artikel dalam ruang vektor multidimensi. Model ini berfungsi sebagai basis pengetahuan sistem yang akan digunakan dalam proses pencarian dan rekomendasi.

2.4. Proses Pencarian dan Rekomendasi

Ketika pengguna memasukkan query pencarian, sistem melakukan langkah-langkah berikut. Query pengguna diproses melalui tahap preprocessing yang sama seperti dokumen untuk menghasilkan vektor query berukuran 768 dimensi. Sistem kemudian menghitung Cosine Similarity antara vektor query dengan seluruh vektor dokumen dalam database [14]. Cosine Similarity mengukur kedekatan semantik dengan menghasilkan skor antara 0 untuk tidak mirip dan 1 untuk sangat mirip [15]. Artikel diurutkan berdasarkan skor similarity dari tertinggi ke terendah dan sistem menampilkan Top-K artikel teratas sebagai hasil rekomendasi kepada pengguna.

2.5. Metode Evaluasi

Kinerja sistem diukur menggunakan tiga metrik standar Information Retrieval yaitu Precision yang mengukur seberapa banyak artikel yang direkomendasikan benar-benar relevan, Recall yang mengukur kemampuan sistem menemukan semua artikel relevan yang tersedia, dan F-Measure sebagai rata-rata harmonik dari Precision dan Recall yang menunjukkan kinerja keseluruhan sistem. Evaluasi dilakukan dengan metode hybrid yang menggabungkan pencocokan kata kunci dan ambang batas similarity di atas 0.5 untuk menentukan

relevansi artikel. Pengujian dilakukan menggunakan berbagai query untuk mengukur kemampuan sistem dalam memahami konteks semantik yang beragam.

3. HASIL DAN PEMBAHASAN

Pada bagian ini dijelaskan hasil penelitian beserta pembahasan secara menyeluruh mencakup hasil pengumpulan data, pembentukan model, dan evaluasi kinerja sistem rekomendasi.

3.1. Hasil Pengumpulan Data dan Preprocessing

Proses web scraping dari Portal Garuda menghasilkan dataset dengan karakteristik sebagai berikut. Sebagian besar topik seperti Sistem Informasi, Teknik Informatika, Kecerdasan Buatan, Data Mining, Pembelajaran Mesin, dan topik populer lainnya berhasil mencapai target penuh 100 artikel per topik. Hal ini menunjukkan ketersediaan literatur yang melimpah untuk topik-topik inti dalam bidang TIK. Namun beberapa topik mengalami kendala ketersediaan data. Decision Tree Regression hanya menghasilkan 60 artikel, Pemrosesan Bahasa Alami hanya 14 artikel, E-Learning tidak menghasilkan data karena halaman pencarian kosong, dan Komputasi Awan berhenti di 69 artikel. Kendala ini mengindikasikan bahwa tidak semua subdomain TIK memiliki publikasi yang merata di Portal Garuda.

Setelah proses scraping selesai, diperoleh 2643 artikel mentah. Proses eliminasi duplikasi menemukan 298 artikel yang muncul pada lebih dari satu topik pencarian. Setelah pembersihan, dataset akhir terdiri dari 2345 artikel unik yang siap diproses. Setiap artikel kemudian melalui tahap preprocessing BERT menghasilkan representasi vektor 768 dimensi. Proses ini menghasilkan Vector Space Model berupa matriks 2345 x 768 yang menjadi basis pengetahuan sistem rekomendasi.

3.2. Hasil Evaluasi Query "Aplikasi Kesehatan"

Pengujian pertama menggunakan query "aplikasi kesehatan" untuk mengevaluasi kemampuan sistem dalam domain informatika kesehatan. Sistem menampilkan 10 artikel teratas dari 50 hasil pencarian dengan hasil evaluasi sebagai berikut. Sistem mencapai Precision sempurna 100 persen yang berarti seluruh artikel yang direkomendasikan relevan dengan topik aplikasi kesehatan. Recall sebesar 62.5 persen menunjukkan sistem berhasil mengambil sebagian besar artikel relevan yang tersedia dalam database. F-Measure mencapai 76.92 persen yang mengindikasikan keseimbangan baik antara ketepatan dan kelengkapan hasil pencarian.

Analisis kualitatif menunjukkan kemampuan pemahaman semantik yang kuat. Artikel peringkat 4 tentang Smart Medicine Box berbasis IoT untuk lansia direkomendasikan meskipun judulnya tidak memuat kata aplikasi secara eksplisit. Hal ini membuktikan BERT mampu memahami bahwa kotak obat pintar merupakan bagian dari ekosistem aplikasi kesehatan. Beberapa artikel bertema Sistem Pakar untuk diagnosis penyakit seperti gangguan endokrin, campak, dan gangguan emosional anak juga masuk dalam rekomendasi. Sistem mengenali bahwa aplikasi diagnosis medis berbasis sistem pakar termasuk dalam kategori aplikasi kesehatan meskipun tidak menggunakan istilah yang sama persis.

Tabel 1. Metrik evaluasi dari query Aplikasi Kesehatan

Metrik evaluasi	Nilai	Penjelasan kinerja sistem
Top-K Retrieval	10	Sistem menampilkan 10 artikel teratas berdasarkan kemiripan semantik dari total 50 hasil pencarian.
Precision	100%	Seluruh artikel yang direkomendasikan relevan dengan topik aplikasi kesehatan
Recall	62.50%	Sistem berhasil mengambil 62,5% dari seluruh artikel relevan yang tersedia.
F-Measure	76.92%	Menunjukkan keseimbangan yang baik antara ketepatan (precision) dan kelengkapan (recall) hasil pencarian dan kemampuan sistem memanggil kembali dokumen yang tersedia.



Artikel tentang Transformasi Digital di Bidang Kesehatan juga dianggap relevan karena membahas penerapan teknologi informasi dalam layanan kesehatan masyarakat. Sistem berhasil menangkap berbagai subdomain informatika kesehatan meliputi aplikasi tele-konseling gizi berbasis AI, sistem pencarian rumah sakit, diagnosis berbasis web dan Android, serta teknologi pendukung seperti IoT dan cloud computing dalam layanan kesehatan.

3.3. Hasil Evaluasi Query "AI"

Pengujian kedua menggunakan query "AI" untuk mengevaluasi pemahaman sistem terhadap konsep Kecerdasan Buatan secara luas. Hasil evaluasi menunjukkan kinerja yang lebih baik dengan Precision 100 persen, Recall 83.33 persen, dan F-Measure 90.91 persen. Nilai Recall yang lebih tinggi mengindikasikan bahwa topik AI memiliki ketersediaan artikel yang lebih baik dalam database dengan estimasi sekitar 12 artikel relevan di mana 10 artikel berhasil ditampilkan pada halaman pertama. Analisis kualitatif menunjukkan variasi domain yang luas. Sistem berhasil mengidentifikasi artikel tentang implementasi AI pada game dan NPC yang menunjukkan pemahaman penerapan AI di bidang rekayasa perangkat lunak. Artikel mengenai aspek hukum dan analisis strategis AI membuktikan sistem memahami AI sebagai konsep multidisipliner tidak terbatas pada aspek teknis.

Tabel 2. Metrik evaluasi dari query AI

Metrik evaluasi	Nilai	Penjelasan kinerja sistem
Top-K Retrieval	10	Sistem menampilkan 10 artikel teratas berdasarkan kemiripan semantik dari total 50 hasil pencarian.
Precision	100%	Seluruh artikel yang direkomendasikan memiliki relevansi tinggi dengan topik Artificial Intelligence (Kecerdasan Buatan).
Recall	83.33%	Sistem berhasil mengambil sebagian besar artikel relevan AI yang tersedia di dalam basis data.
F-Measure	90.91%	Menunjukkan keseimbangan yang sangat baik antara ketepatan dan kelengkapan rekomendasi.

Artikel tentang AI dalam Natural Language Processing menunjukkan sistem mampu mengaitkan AI dengan sub-bidang spesifik dalam ilmu komputer. Sistem pakar dalam pelayanan publik juga diidentifikasi sebagai aplikasi AI yang menegaskan pemahaman hubungan antara AI dan expert system. Sistem berhasil menangkap spektrum domain AI yang beragam mulai dari implementasi teknis seperti pathfinding dan game AI, aspek teoretis seperti definisi dan konsep AI, aplikasi sosial dalam pelayanan publik dan smart city, hingga kajian strategis dan hukum terkait status legal AI.

3.4. Pembahasan Umum

Hasil pengujian membuktikan bahwa sistem rekomendasi berbasis BERT embedding memiliki beberapa keunggulan signifikan. Pertama, sistem tidak bergantung pada pencocokan kata kunci literal melainkan memahami makna semantik kontekstual dari query dan dokumen. Hal ini terlihat dari kemampuan sistem merekomendasikan artikel relevan meskipun tidak mengandung kata kunci yang sama persis dengan query. Kedua, sistem mampu menangkap variasi subdomain dalam satu topik besar. Untuk query kesehatan, sistem mengidentifikasi berbagai jenis aplikasi mulai dari diagnosis, manajemen obat, hingga transformasi digital layanan kesehatan. Untuk query AI, sistem mengenali implementasi teknis, kajian teoretis, dan aplikasi sosial sebagai bagian dari ekosistem kecerdasan buatan. Ketiga, nilai Precision yang konsisten 100 persen pada kedua pengujian menunjukkan bahwa sistem sangat akurat dalam menyaring artikel tidak relevan. Semua artikel yang ditampilkan memiliki keterkaitan erat dengan topik yang dicari. Nilai Recall yang berbeda antara kedua query mengindikasikan ketersediaan data yang tidak merata untuk berbagai topik dalam database. Query AI memiliki Recall lebih tinggi karena topik kecerdasan buatan memiliki publikasi yang lebih banyak dibandingkan aplikasi kesehatan.



F-Measure yang tinggi pada kedua pengujian dengan nilai 76.92 persen dan 90.91 persen membuktikan sistem mencapai keseimbangan optimal antara ketepatan dan kelengkapan hasil pencarian. Sistem tidak hanya akurat tetapi juga komprehensif dalam mengambil artikel relevan yang tersedia. Kemampuan BERT dalam encoding kontekstual terbukti superior dibandingkan metode tradisional berbasis frekuensi kata. Model ini menghasilkan representasi vektor yang menangkap hubungan semantik mendalam antar konsep sehingga dapat mengidentifikasi relevansi artikel berdasarkan makna bukan sekadar kemunculan kata tertentu. Penggunaan Cosine Similarity pada ruang vektor 768 dimensi memberikan metrik kemiripan yang sensitif terhadap nuansa semantik. Skor similarity yang dihasilkan mencerminkan kedekatan makna kontekstual secara akurat sehingga pengurutan artikel berdasarkan relevansi menjadi lebih presisi.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem rekomendasi pencarian artikel ilmiah bidang TIK menggunakan model BERT yang mampu memahami konteks semantik query pengguna. Sistem yang dibangun telah memenuhi tujuan penelitian yaitu meningkatkan akurasi pencarian berbasis pemahaman semantik dibandingkan metode tradisional berbasis kata kunci. Hasil evaluasi menunjukkan kinerja sistem yang sangat baik dengan Precision konsisten 100 persen pada berbagai query pengujian. F-Measure mencapai 76.92 persen untuk query aplikasi kesehatan dan 90.91 persen untuk query AI yang mengindikasikan keseimbangan optimal antara ketepatan dan kelengkapan hasil pencarian. Sistem terbukti mampu mengidentifikasi artikel relevan tanpa bergantung pada pencocokan kata kunci literal melainkan berdasarkan pemahaman makna kontekstual.

Prospek pengembangan penelitian ini meliputi perluasan dataset untuk mencakup lebih banyak topik dan sumber publikasi, implementasi teknik re-ranking untuk meningkatkan Recall pada topik dengan artikel terbatas, penambahan fitur filtering berdasarkan tahun publikasi atau jurnal tertentu, dan pengembangan antarmuka pengguna yang lebih interaktif. Penelitian lanjutan dapat mengeksplorasi kombinasi BERT dengan metode ensemble learning atau implementasi model BERT yang lebih besar untuk meningkatkan performa sistem. Sistem yang dikembangkan dapat diimplementasikan sebagai layanan pencarian literatur akademik yang membantu peneliti menemukan artikel relevan dengan lebih efisien dan akurat berdasarkan pemahaman semantik kontekstual.

REFERENSI

- [1] H. Ko, S. Lee, Y. Park, dan A. Choi, "A survey of recommendation systems: recommendation models, techniques, and application fields," *Electronics*, vol. 11, no. 1, hlm. 141, 2022.
- [2] A. P. Muria, H. Sujaini, H. S. Pratiwi, dan J. H. Hadari Nawawi, "Sistem rekomendasi artikel sebagai acuan studi literatur menggunakan metode N-Gram," Doctoral Dissertation, Universitas Tanjungpura, 2022.
- [3] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," dalam *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019, hlm. 4171-4186.
- [4] S. Wang, S. Zhuang, dan G. Zuccon, "BERT-based dense retrievers require interpolation with BM25 for effective passage retrieval," dalam *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2021, hlm. 317-324.
- [5] T. Gao, X. Yao, dan D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," dalam *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, hlm. 6894-6910.
- [6] N. Reimers dan I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," dalam *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. Tersedia: <https://arxiv.org/abs/1908.10084>
- [7] V. Meilinger, S. Reji, M. Koerwien, dan G. Pojani, "Enhancing data discoverability: A semantic item-based recommendation system for open data catalogues," dalam *European Conference on Information Retrieval*, Springer, Cham, 2022, hlm. 165-177.
- [8] R. Wu, "RecBERT: Semantic recommendation engine with large language model enhanced query segmentation for k-nearest neighbors ranking retrieval," *Intelligent and Converged Networks*, vol. 5, no. 1, hlm. 42-52, 2024. Tersedia: <https://doi.org/10.23919/ICN.2024.0004>
- [9] Q.-H. Lu dan D. Huang, "English sentence similarity algorithm based on vector space model," *Journal of Nonlinear and Convex Analysis*, vol. 25, no. 1, hlm. 19-46, 2024.
- [10] Y. Zhou, K. Ethayarajh, D. Card, dan D. Jurafsky, "Dense passage retrieval for open-domain question answering," dalam *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, hlm. 6769-6781.
- [11] A. Otterbring, S. Verberne, dan L. Azzopardi, "Improving BERT-based query-by-document retrieval with multi-task optimization,"



- dalam *European Conference on Information Retrieval*, Springer, Cham, 2022, hlm. 128-141.
- [12] Ö. Keleş dan B. Kaya, "Optimizing SBERT for long text clustering: two novel approaches with empirical insights," *The Journal of Supercomputing*, 2025, doi: 10.1007/s11227-025-07414-4.
- [13] S. Verma, "Semantic search with S-BERT is all you need: Building in-house semantic search engine from scratch—fast and accurate," *Medium*, Feb. 2024. Tersedia: <https://subirverma.medium.com/semantic-search-with-s-bert-is-all-you-need-951bc710e160>
- [14] G. Chen, X. Shi, M. Chen, dan L. Zhou, "Text similarity semantic calculation based on deep reinforcement learning," *International Journal of Security and Networks*, vol. 15, no. 1, hlm. 59-66, 2020.
- [15] K. You, "Semantics at an angle: When cosine similarity works until it doesn't," *arXiv preprint arXiv:2504.16318*, 2025.