Vol. 11, No. 1, April 2026, Page. 1-14
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Klasifikasi Tingkat Pemrosesan Makanan Berbasis Teks Komposisi Menggunakan Strategi Weighted Ensemble Large Language Models

Khairul Hudha Nasution[*1], Indra Budi[2]

[1,2]Computer Science, Faculty of Computer Science, University of Indonesia, Depok, Indonesia
Email: [1]khairul.hudha@ui.ac.id, [2]indra@cs.ui.ac.id

Abstrak

Klasifikasi tingkat pemrosesan makanan merupakan langkah krusial dalam mitigasi risiko kesehatan global akibat konsumsi makanan *ultra-proses*. Meskipun label komposisi tersedia pada kemasan, penulisan yang kecil, tidak terstruktur, typo serta terminologi kimia yang kompleks memiliki kecenderungan menyulitkan penilaian manual oleh konsumen. Pemanfaatan *Large Language Models* (LLM) menawarkan potensi efisiensi deteksi otomatis, namun mengandalkan satu arsitektur model tunggal memiliki risiko tinggi akibat variabilitas performa. Penelitian ini bertujuan untuk mengevaluasi efektivitas strategi *Weighted Ensemble Learning* dibandingkan model tunggal dalam memprediksi skor pemrosesan makanan (FPro) pada dataset GroceryDB. Eksperimen dilakukan menggunakan lima arsitektur LLM dengan skala parameter kecil hingga sedang (Gemma-3-4B, Llama-3.2-3B, Qwen3-4B, R1-Distill-1.5B, dan Phi-2) melalui pendekatan *Weighted Voting* berbasis kinerja historis. Hasil evaluasi menunjukkan adanya disparitas ekstrem pada kinerja model tunggal, di mana model berkapasitas rendah (Phi-2) mengalami kegagalan penalaran dengan *F1-score* hanya 10%, sementara model dengan kemampuan instruksi tinggi (Gemma) mencapai 68%. Penerapan strategi *Ensemble* berhasil memitigasi kelemahan model individual melalui mekanisme koreksi silang, meningkatkan akurasi *F1-score* menjadi 70% dan menghasilkan prediksi yang lebih stabil serta *robust* dibandingkan jika hanya mengandalkan satu model terbaik sekalipun.

Kata kunci: *ensemble learning, large language models, ultra-processed foods.*

Abstract

Food processing level classification is a crucial step in mitigating global health risks due to ultra-processed food consumption. Although ingredient labels are available on packaging, small text, unstructured formats, typos, and complex chemical terminology often complicate manual assessment by consumers. The utilization of Large Language Models (LLMs) offers potential for automated detection efficiency, yet relying on a single model architecture carries high risks due to performance variability. This study aims to evaluate the effectiveness of a Weighted Ensemble Learning strategy compared to single models i n predicting food processing scores (FPro) on the GroceryDB dataset. Experiments were conducted using five small-to-medium scale LLM architectures (Gemma-3-4B, Llama-3.2-3B, Qwen3-4B, R1-Distill-1.5B, and Phi-2) through a Weighted Voting approach based on historical performance. Evaluation results indicate an extreme disparity in single model performance, where the low-capacity model (Phi-2) experienced reasoning failure with an F1-score of only 10%, while the model with high instruction capability (Gemma) reached 68%. The application of the Ensemble strategy successful in mitigating individual model weaknesses through a cross-correction mechanism, increasing the F1-score to 70% and producing predictions that are more stable and robust compared to relying on even the single best model.

Keywords: *ensemble learning, large language models, ultra-processed foods.*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. PENDAHULUAN

Dalam satu dekade terakhir, paradigma kesehatan global telah menempatkan konsumsi makanan *ultra-proses foods* (UPF) sebagai salah satu faktor risiko utama terhadap peningkatan penyakit tidak menular, seperti obesitas dan diabetes tipe 2. Studi epidemiologi berskala besar, seperti NutriNet-Santé Cohort, menunjukkan adanya korelasi linier yang signifikan antara konsumsi makanan dengan tingkat pemrosesan tinggi dan peningkatan risiko mortalitas serta gangguan kardiovaskular [1]. Meskipun urgensi kesehatan ini nyata, konsumen menghadapi kendala dalam mengidentifikasi produk *ultra-proses* secara mandiri. Label komposisi pada kemasan umumnya memuat terminologi kimia yang kompleks

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

dan teknis, menjadikan penilaian manual sebagai proses yang tidak efisien dan rentan terhadap kesalahan interpretasi.

Sebagai respon terhadap keterbatasan penilaian manual, Barabási Lab mengembangkan metrik *Food Processing* Score (FPro) yang memungkinkan kuantifikasi derajat pemrosesan makanan secara lebih presisi dibandingkan sistem kategori konvensional [2]. Namun, tantangan terbesar dalam mengimplementasikan perhitungan ini secara massal terletak pada analisis data teks komposisi yang tidak terstruktur.

Perkembangan terbaru dalam *Large Language Models* (LLM) menawarkan kemampuan semantik yang jauh lebih unggul dibandingkan metode NLP tradisional dalam memahami konteks bahan pangan tersebut, serta telah terbukti efektif dalam klasifikasi teks berbasis kesehatan [3]. Kendati demikian, penerapan LLM secara tunggal (*Single Model*) memiliki risiko fundamental. Setiap arsitektur model memiliki karakteristik unik dan bias yang berbeda, model tertentu mungkin unggul dalam mendeteksi istilah kimia namun lemah dalam penalaran nutrisi, sementara model lain rentan terhadap halusinasi. Mengandalkan satu arsitektur model saja untuk aplikasi kesehatan dinilai belum memadai untuk menjamin stabilitas prediksi.

Untuk mengatasi variabilitas kinerja model tunggal, penelitian ini mengusulkan strategi *Ensemble Learning* sebagai mekanisme stabilisasi. Pendekatan ini bekerja dengan mengagregasi prediksi dari beberapa arsitektur LLM untuk menutupi kelemahan spesifik satu model dengan kekuatan model lainnya. Penelitian terbaru oleh Kamen et al. memvalidasi bahwa strategi penggabungan output LLM (LLM *Ensemble*) mampu meningkatkan metrik evaluasi secara signifikan, bahkan melampaui kinerja model terbaik tunggal dalam tugas kategorisasi konten [4]. Namun, penelitian tersebut masih terbatas pada domain umum dan tidak terdapat didalaminya kompleksitas terminologi yang digunakan pada nutrisi informasi pada label makanan. Hingga saat ini, belum ada penelitian yang secara spesifik menerapkan dan mengevaluasi strategi Weighted Ensemble LLM untuk tugas klasifikasi tingkat pemrosesan makanan berbasis teks komposisi.

Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi efektivitas strategi *Weighted Ensemble* dalam mengotomatisasi klasifikasi tingkat pemrosesan makanan menggunakan *dataset* GroceryDB. Untuk mencapai tujuan tersebut dan memperjelas arah investigasi, penelitian ini merumuskan dua pertanyaan penelitian utama:

1. RQ1 : Bagaimana perbandingan performa dan karakteristik dari arsitektur *Large Language Models* (LLM) secara tunggal dengan ukuran model kecil dan menengah dalam mengklasifikasikan tingkat pemrosesan makanan serta menangani kompleksitas terminologi nutrisi pada teks produk?
2. RQ2 : Bagaimana efektivitas implementasi strategi *Weighted Ensemble Learning* dengan mekanisme *Weighted Voting* terbukti mampu menghasilkan prediksi klasifikasi dan estimasi skor FPro yang lebih stabil, akurat, dan robust dibandingkan dengan ketergantungan pada model tunggal terbaik?

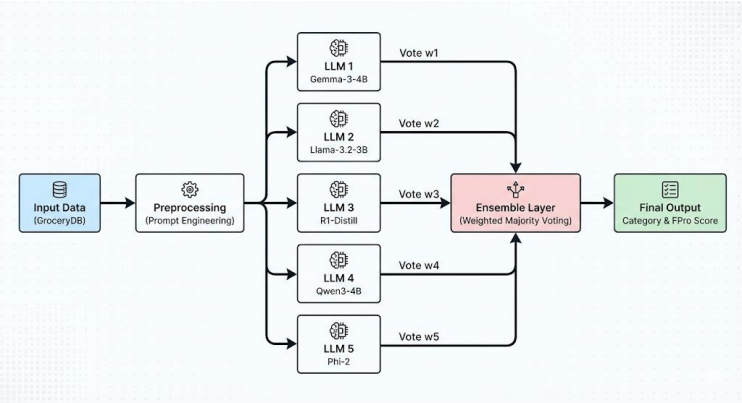
Sebagai batasan ruang lingkup untuk menjaga fokus penelitian, penelitian ini dilakukan menggunakan dataset GroceryDB yang merepresentasikan produk pasar ritel Amerika Serikat dengan label informasi tekstual dalam bahasa Inggris. Selain itu, eksplorasi dibatasi pada lima arsitektur model LLM dengan ukuran parameter kecil hingga menengah antara 1 hingga 6 miliar (1B-6B) yaitu, Gemma-3-4B-IT, Llama-3.2-3B-Instruct, Qwen3-4B, R1-Distill-Qwen-1.5B, dan Phi-2. Penelitian ini berupaya membuktikan bahwa pendekatan gabungan beberapa model mampu menghasilkan prediksi yang lebih akurat dan robust dibandingkan ketergantungan pada model tunggal. Penentuan skala model ini (kecil dan sedang) dilakukan untuk memastikan efisiensi sumber daya komputasi serta memungkinkan replikasi.

2. METODE PENELITIAN

Penelitian ini menerapkan pendekatan eksperimental kuantitatif untuk membangun sistem klasifikasi tingkat pemrosesan makanan yang robust. Kerangka kerja penelitian dirancang secara sistematis, dimulai dari tahap akuisisi dan pra-pemrosesan data teks dari *dataset* GroceryDB, dilanjutkan dengan tahap inferensi menggunakan lima arsitektur *Large Language Models* (LLM) yang berbeda secara paralel. Output prediksi dari masing-masing model kemudian diagregasi melalui mekanisme

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Weighted Ensemble untuk menghasilkan keputusan klasifikasi final dan estimasi skor FPro. Alur kerja keseluruhan dari sistem yang diusulkan diilustrasikan pada Gambar 1.



Gambar 1. Metodelogi Penelitian

2.1. *Pre-Processing* Data

Penelitian ini memanfaatkan *dataset* GroceryDB[5] sebagai sumber data utama. *Dataset* ini merepresentasikan puluhan ribu produk makanan dan minuman dari pasar ritel Amerika Serikat yang dilengkapi dengan informasi tekstual (nama produk dan daftar bahan) serta label kuantitatif berupa *Food Processing Score* (FPro). Tahapan pra-pemrosesan data dilakukan melalui langkah-langkah berikut:

1. Agregasi Data: Data dari tabel *produk.csv* dan *ingredient.json* digabungkan menggunakan pengenal unik produk (id) untuk menyatukan atribut nama dan daftar bahan.
2. Konstruksi Input: Atribut nama produk dan daftar bahan digabungkan menjadi satu *string* teks tunggal yang berfungsi sebagai input konteks bagi model LLM. Pra-pemrosesan teks dilakukan secara minimal (seperti normalisasi spasi dan karakter) untuk menjaga keaslian terminologi bahan kimia yang terkandung di dalamnya.

Setelah melalui proses pembersihan (*data cleaning*) untuk menghapus entri dengan data bahan yang kosong atau label ambigu, *dataset* final dibagi dengan rasio 80:20. Sebesar 80% data digunakan sebagai set pelatihan dan 20% sisanya digunakan sebagai set untuk evaluasi performa.

2.2. Definisi *Task* Penelitian

Setelah tahap pra-pemrosesan data, perlu rasanya menjelaskan bahwa penelitian ini membagi eksperimen ke dalam dua tugas (*task*) utama untuk mengevaluasi kemampuan model secara komprehensif menggunakan input tekstual yang sama:

2.2.1. *Task* 1: Klasifikasi Tingkat Pemrosesan (*Classification*)

Tugas ini bertujuan memprediksi kategori tingkat pemrosesan makanan secara klasifikasi. Model diminta untuk mengklasifikasikan produk ke dalam tiga label kelas berdasarkan ambang batas skor FPro:

- *Low* (Rendah): Produk dengan FPro<0.3, umumnya makanan minim olahan.
- *Medium* (Sedang): Produk dengan 0.3<= FPro <=0.6.
- *High* (Tinggi): Produk dengan FPro >0.6, mencirikan makanan *ultra-proses.*

2.2.2. *Task* 2: Identifikasi Bahan Utama

Tugas ini dirancang untuk menguji kemampuan penalaran model. Bukan hanya memberikan label klasifikasi akhir, model diminta untuk menganalisis daftar komposisi dan mengidentifikasi satu bahan spesifik yang paling berkontribusi terhadap status pemrosesan produk tersebut.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Sebagai contoh, jika sebuah produk diklasifikasikan sebagai *High*, model harus mampu menunjuk bahan aditif tertentu (misalnya: "*High Fructose Corn Syrup*" atau "Sodium Benzoate") sebagai penyebab utamanya.

2.3. Arsitektur Model dan Strategi *Prompting*

Penelitian ini menggunakan pendekatan *fine-tuning* untuk mengevaluasi kemampuan model dalam memahami instruksi dan melakukan penalaran. Landasan utama dari seluruh model yang digunakan adalah arsitektur *Transformer* [6] yang telah berevolusi menjadi *Large Language Models* (LLM) moderen [7], yang memungkinkan pemrosesan ketergantungan antar kata secara *paralel* melalui mekanisme *Self-Attention*.

2.3.1. Arsitektur Model

Pemilihan model dilakukan melalui seleksi berbasis komunitas *open-source* (Hugging Face) dengan batasan ukuran parameter antara 1 hingga 6 miliar (1B-6B). Kriteria ini diterapkan untuk memastikan efisiensi sumber daya komputasi (*resource efficiency*) dan memungkinkan replikasi pada perangkat keras standar. Lima arsitektur model yang dipilih untuk eksperimen ini adalah:

1. Gemma-3-4B-IT (Google) [8]
2. Llama-3.2-3B-Instruct (Meta) [9]
3. Qwen3-4B (Alibaba Cloud) [10]
4. R1-Distill-Qwen-1.5B (DeepSeek/Distilled) [11]
5. Phi-2 (Microsoft) [12]

2.3.2. Strategi *Prompting*

Untuk memastikan konsistensi *output* antar model yang berbeda arsitektur, strategi *prompting* dirancang menggunakan teknik *Role-Playing*. Setiap model diinstruksikan untuk berperan sebagai "Ahli Ilmu Pangan dan Nutrisi" (*Expert Food Scientist*) untuk memicu kemampuan penalaran bertahap (*chain-of-thought*) [13]. Instruksi ini dikombinasikan dengan batasan format yang ketat (*format constraints*) untuk memaksa model menghasilkan keluaran terstruktur berupa objek JSON, guna memfasilitasi proses parsing dan evaluasi kuantitatif secara otomatis. Sehingga secara detail promp yang digunakan dapat dilihat pada Gambar 2.

```
You are a food scientist.

Product: {name}
Category: {category}
Fpro: {Fpro}
Ingredients: {ing_str}
Max_Ingred_Fpro :{ Max_Ingred_Fpro}

Task:
1. Estimate Food Processing Score (FPro) (0.0 to 1.0).
2. Identify main contributor ingredient.

Output strictly valid JSON only.
```

Gambar 2. Detail *Promp*

Dimana:

- Name : Nama produk
- Category : Katagori produk
- Fpro : Nilai pemrosesan makanan (*ground-truth*)
- Ingredients : Bahan-bahan produk, format data {name dan Fpro}
- Max_Ingred_Fpro : Nama bahan dengan Fpro tertinggi (*ground-truth*)

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Dimana jika dalam proses validasi menggunakan *promp* yang sama tapi label Fpro dan Max_Ingred_Fpro akan dihilangkan, serta label Ingredients dari yang awalnya berformat { nama dan Fpro} menjadi { nama }.

2.4. Strategi *Ensemble* (*Weighted Voting*)

Sebagai kontribusi metode utama, penelitian ini menerapkan strategi *Weighted Voting*. Berbeda dengan pendekatan *Simple Majority* yang memberikan hak suara setara (satu model satu suara), metode ini memberikan bobot dinamis (w_i) pada setiap model (i) berdasarkan kinerja historisnya (*F1-Score*). *F1-Score* yang digunakan disini adalah nilai *Macro F1-Score* yang didapat dari hasil masing-masing model pada tahapan pengujian menggunakan *dataset* validasi. Pendekatan ini dipilih untuk mengatasi kelemahan model tunggal serta memitigasi risiko halusinasi yang kerap muncul pada model *generatif* [14], sejalan dengan prinsip *self-consistency* yang membuktikan bahwa agregasi jawaban dapat meningkatkan akurasi penalaran [15].

Mekanisme penentuan keputusan akhir dilakukan dengan menjumlahkan bobot dari seluruh model yang memprediksi kategori yang sama. Kategori yang memiliki total akumulasi bobot tertinggi ditetapkan sebagai hasil prediksi akhir konsensus. Secara sederhana, logika penentuan keputusan dapat dituliskan sebagai berikut:

$$Skor(katagori_i) = \sum Bobot\ Model\ yang\ Memilih\ i \qquad (1)$$

$$Hasil_{Akhir} = Kategori\ dengan\ Skor\ Tertinggi \qquad (2)$$

Dengan mekanisme ini, model yang memiliki performa rendah akan memiliki bobot suara yang kecil atau mendekati nol, sehingga kesalahan prediksinya tidak mendominasi keputusan akhir sistem. Strategi ini selaras dengan temuan Kamen et al. yang menunjukkan bahwa agregasi berbasis kompetensi model mampu meningkatkan stabilitas prediksi secara signifikan dibandingkan model tunggal [4].

2.5. Evaluasi dan Metrik

Evaluasi performa model dilakukan pada *dataset* validasi (20% data) yang tidak digunakan dalam proses *training*. Mengingat tugas utama adalah klasifikasi *multi-kelas* (*Low*, *Medium*, *High*), metrik evaluasi difokuskan pada *F1-Score*. Metrik ini dipilih karena kemampuannya memberikan gambaran performa yang lebih seimbang[16]. F1-Score juga lebih baik jika dibandingkan antara *Precision*, *Recall* dan akurasi semata, terutama jika terdapat ketidakseimbangan distribusi kelas dalam *dataset*.

- *Precision*: persentase rasio dari *true positive* terhadap jumlah dari *true positive* dan *false positive*.
- *Recall*: persentase rasio dari banyaknya *true positive* terhadap jumlah *true positive* dan *false negative*.
- *F1-Score*: rata-rata dari *precision* dan *recall*.
- *Accuracy*: persentase rasio dari data yang diklasifikasi dengan benar terhadap total data.

dimana *True positive* (TP) dan *True Negative* (TN) menunjukkan nilai yang terklasifikasi dengan benar, dan *False Positive* (FP) and *False Negative* (FN) menunjukkan nilai yang terklasifikasi dengan salah. Matrik-matrik tersebut juga dapat ditunjukkan melalui persamaan berikut ini.

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

$$F1_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (5)$$

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \qquad (6)$$

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan evaluasi terhadap efektivitas kerangka kerja *Ensemble* LLM yang diusulkan. Analisis disusun secara bertahap untuk menyoroti kesenjangan kinerja yang antar model tunggal dan membuktikan peran strategi *Weighted Voting* sebagai mekanisme stabilisasi. Pembahasan diawali dengan ringkasan hasil pra-pemrosesan data, dilanjutkan dengan komparasi kinerja baseline pada lima arsitektur model, dan diakhiri dengan analisis mendalam mengenai peningkatan *F1-score* dan stabilitas yang dicapai melalui pendekatan Ensemble.

3.1. Pengaturan Eksperimen

Penelitian ini menggunakan sistem yang memiliki spesifikasi CPU AMD EPYC dengan GPU RTX 5090 dengan VRAM 32 GB dan RAM 64 GB. Dilakukan di bulan Oktober 2025 hingga Desember 2025.

3.2. Preprocessing Data

Sebelum dilakukan inferensi model, *dataset* mentah dari GroceryDB yang terdiri dari dua data utama (untuk mempermudah kita sebut *foods.csv* dan *ingredients.json*) digabungkan menggunakan kunci relasi *Product ID*. Himpunan data awal mencakup 48.625 produk. Untuk menjamin validitas evaluasi, dilakukan proses pembersihan data (*data cleaning*) guna mengeliminasi data yang bersifat ambigu.

Data dikategorikan sebagai ambigu apabila skor FPro ditemukan kosong atau berada di luar rentang validitas 0,0 hingga 1,0. Selain itu, produk yang tidak mencantumkan nama atau daftar bahan juga dieliminasi karena ketiadaan komponen tekstual utama yang digunakan model untuk memahami konteks produk. Proses seleksi ini mereduksi jumlah data secara signifikan menjadi 22.195 produk yang bersih dan teranotasi lengkap.

Dataset final ini kemudian dipartisi dengan rasio 80:20, di mana data uji (*test set*) dipisahkan sepenuhnya untuk memastikan objektivitas evaluasi. Secara spesifik, setelah dilakukan pembersihan dataset final terbagi menjadi 17.756 data Train dengan pembagian sebagai berikut 1.388 Low, 3.115 Medium dan 13.253 High. Serta 4.439 data Val dengan pembagian sebagai berikut 322 Low, 764 Medium dan 3.353 High.

Untuk memberikan gambaran konkret mengenai format data yang digunakan dalam eksperimen, Tabel 1 menyajikan contoh representasi input teks beserta label targetnya. Kolom Input Teks merupakan hasil konkatenasi antara nama produk dan daftar bahan, yang menjadi konteks utama bagi model LLM.

Tabel 1. Sampel Data Input

Field Name	Data
id	wf_virgils-cream-soda-4pk-b004qyhf3g
Name	Cream Soda (4-pk)
category	drink-soft-energy-mixes
Fpro	0.756404762
Max_Ingred_Fpro	<natural> caramel color
ingredients	[{ "Name": "<sparkling> <filtered> water", "Fpro": 0.769330303 }, { "Name": "cane sugar", "Fpro": 0.83363591 }, { "Name": "<natural> caramel color", "Fpro": 0.836743214 }, { "Name": "natural flavor", "Fpro": 0.793640937 }, { "Name": "malic acid for flavor", "Fpro": 0.764555415 }]

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

3.3. Evaluasi Kinerja Model Tunggal (*Baseline*)

Evaluasi tahap pertama dilakukan pada model yang terlebih dahulu dilatih menggunakan data train lalu dilakukan evaluasi dengan menggunakan *dataset* eval dengan perhitungan *F1-score* pada dua task untuk mengukur kapabilitas setiap arsitektur model.

Tabel 2. Hasil Model LLM Secara Tunggal (Baseline)

No.	Model	Baseline	
		T1	T2
1	R1-Distill-Qwen-1.5B	50%	13%
2	gemma-3-4b-it	68%	27%
3	Llama-3.2-3B-Instruct	53%	21%
4	phi-2	10%	7%
5	Qwen3-4B	19%	13%

Berdasarkan hasil eksperimen pada Tabel 2, ditemukan adanya variabilitas performa yang sangat ekstrem antar model, meskipun seluruh model yang diuji berada dalam kategori ukuran parameter menengah (*Small-to-Medium sized models*). Di antara seluruh model tunggal yang diuji, berikut beberapa hal menarik diantaranya:

1. Gemma-3-4B menunjukkan kinerja yang paling baik. Model ini mencatatkan *F1-score* klasifikasi sebesar 68%. Tingginya *F1-score* ini mengindikasikan bahwa arsitekturnya memiliki keunggulan dalam kepatuhan terhadap instruksi dan pemahaman semantik.
2. Hal lain yang menarik adalah model Llama-3.2-3B dan R1-Distill-Qwen menempati posisi menengah dengan *F1-score* masing-masing 53% dan 50%. Meskipun mampu menangkap sebagian besar pola klasifikasi umum, kedua model ini masih menunjukkan tingkat kesalahan yang cukup banyak pada kasus-kasus ambiguitas tinggi.
3. Temuan yang paling menonjol adalah kegagalan sistematis pada model Phi-2 dan Qwen3-4B, yang masing-masing hanya mencapai *F1-score* 10% dan 19%. Rendahnya angka ini (terutama Phi-2 yang hanya 10%) mengindikasikan terjadinya *reasoning collapse*. Tanpa bantuan pengetahuan eksternal, model-model ini gagal memetakan teks komposisi bahan ke label kategori yang benar. Kegagalan ini kemungkinan besar disebabkan oleh keterbatasan kapasitas parameter dalam menyimpan pengetahuan domain spesifik (*internal knowledge*) mengenai istilah-istilah kimia pangan, sehingga model cenderung melakukan halusinasi atau tebakan acak.

3.3.1. *Error Analysis* Kinerja Model Tunggal (*Baseline*)

Untuk memvalidasi temuan kuantitatif di atas, dilakukan analisis lanjutan terhadap *log* prediksi guna memahami pola kegagalan spesifik pada setiap arsitektur model. Analisis difokuskan pada kasus-kasus di mana model mengalami nilai kecil/kegagalan atau ketidakpatuhan terhadap format instruksi. Sebagai studi kasus representatif, Tabel 3 menyajikan perbandingan *output* model terhadap produk "*Honey & Nut Granola*". Produk ini dipilih karena memiliki *profil* bahan yang menipu: mengandung bahan dasar sehat (*rolled oats*) namun juga bahan pemrosesan tinggi (*canola oil, molasses*) yang seharusnya menghasilkan skor FPro tinggi. Data Sebagai berikut:

- Produk Target: *Honey & Nut Granola*
- Komposisi: *Rolled oats, Honey, Canola oil, Coconut, Walnut, Molasses...*
- Label Sebenarnya (*Ground Truth*): *High* (0.79)
- Main Ingredient (*Ground Truth*): *Molasses*

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Tabel 3. Hasil Error Analisis Pada Model LLM Secara Tunggal (baseline)

Model	Status	Raw Output	Analisis Kegagalan
Gemma-3-4B	BERHASIL	{"FPro": 0.988, "Main_Ingredient": "molasse"}	Model berhasil mengabaikan "oats" dan mengidentifikasi "molasse" sebagai penanda ultra-processed food. Format JSON valid dan skor mendekati ground truth
Phi-2	GAGAL (Logic)	{"FPro": 0.0, "Main_Ingredient": "oat"}	Meskipun format output menyerupai JSON, model gagal melakukan penalaran. Ia menetapkan skor 0.0 karena bias terhadap kata kunci "oat", mengabaikan keberadaan minyak dan gula. Ini membuktikan model kecil rentan terhadap keyword bias.
Qwen3-4B	GAGAL	""	Model gagal menghasilkan output apapun yang dapat dibaca (blank response). Ini mengindikasikan ketidakmampuan model dalam mematuhi batasan instruksi sistem (instruction following), menyebabkan crash pada pipeline evaluasi.

Temuan Kunci dari Analisis Log:

1. Fenomena *Reasoning Collapse* dan *Capacity Threshold*: Kegagalan pada Phi-2 bukan sekadar kesalahan prediksi biasa, melainkan indikasi adanya ambang batas kapasitas (*capacity threshold*). Dengan parameter hanya 2.7B, model memiliki ruang representasi yang terbatas untuk memahami hubungan kompleks antara bahan-bahan penyusun. Hal ini memicu *reasoning collapse*, di mana model gagal melakukan penalaran dan melakukan bias. Model cenderung melakukan penyederhanaan (*oversimplification*) dengan hanya mendeteksi kata yang familiar dan positif (seperti "*oats*") serta mengabaikan bahan *ultra-proses* lainnya, sehingga menghasilkan skor 0.0 yang tidak realistis.
2. Penalaran dan Kepatuhan Instruksi: Pada Qwen3-4B menunjukkan bahwa pada model dengan kapasitas rendah, terdapat beban kognitif yang besar saat harus menjalankan dua tugas sekaligus: melakukan penalaran nutrisi dan mempertahankan format JSO. Kegagalan respons (*blank response*) mengindikasikan bahwa komputasi model habis untuk memproses batasan format instruksi sehingga gagal menghasilkan konten jawaban. Sebaliknya, Gemma-3-4B menunjukkan stabilitas arsitektur yang mampu menyeimbangkan kepatuhan format dengan akurasi penalaran, yang menjadi pembeda utama dalam kualitas *baseline*. Kepatuhan Instruksi: Qwen3-4B menunjukkan tingkat kegagalan format yang tertinggi, di mana model tidak merespons atau menghasilkan teks yang tidak dapat di-parse. Sebaliknya, Gemma-3-4B konsisten menghasilkan struktur JSON yang valid, yang menjadi faktor utama tingginya F1-score model ini.

Kesimpulan lainnya hasil ini membuktikan bahwa tidak semua LLM memiliki kapabilitas yang setara untuk tugas spesifik domain. Terdapat kesenjangan performa yang besar (selisih antara Phi-2 dan Gemma) yang menegaskan risiko tinggi jika sistem hanya bergantung pada satu jenis model secara acak tanpa validasi kinerja terlebih dahulu.

3.4. Evaluasi Strategi *Ensemble*

Untuk mengatasi disparitas kinerja yang ditemukan pada model tunggal, strategi *Ensemble* diterapkan sebagai mekanisme "jaring pengaman". Penelitian ini membandingkan kinerja model tunggal

terbaik (Gemma-3-4B) dengan hasil konsensus sistem *Ensemble* yang menggunakan metode *Weighted Voting*.

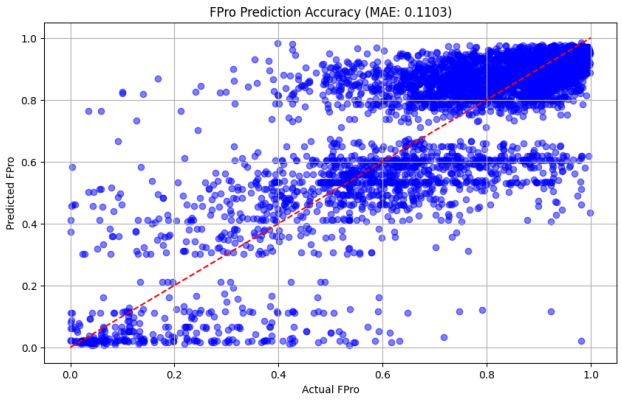
Bobot suara untuk setiap model ditentukan secara dinamis berdasarkan persentase *F1-score* masing masing model pada evaluasi tugas sebelumnya. Hal ini memastikan bahwa model yang berhasil lebih andal memiliki pengaruh lebih besar dalam pengambilan keputusan dibandingkan model yang lebih lemah.

Tabel 4. Hasil LLM Secara Tunggal dengan LLM Secara Ensemble

No.	Model	Baseline	
		T1	T2
1	R1-Distill-Qwen-1.5B	50%	13%
2	gemma-3-4b-it	68%	27%
3	Llama-3.2-3B-Instruct	53%	21%
4	phi-2	10%	7%
5	Qwen3-4B	19%	13%
6	Ensemble	70%	26%

Hasil eksperimen, Tabel 4 mengonfirmasi bahwa pendekatan Ensemble secara konsisten mengungguli kinerja model tunggal terbaik.

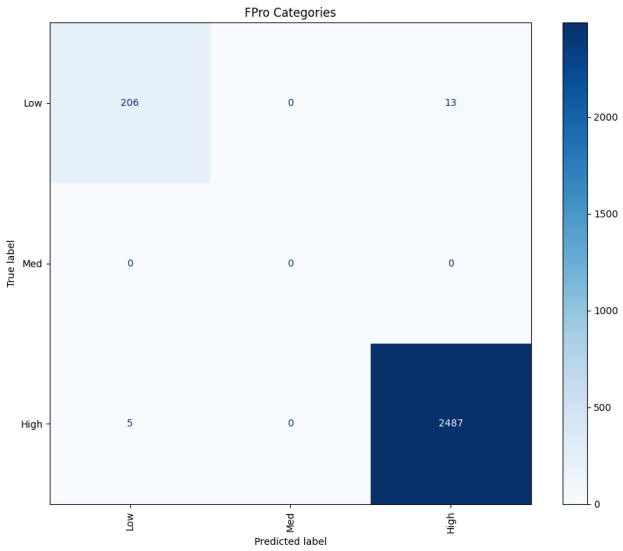
Untuk mengevaluasi performa sistem dalam tugas *regresi*, Gambar 3 menggambarkan perbandingkan antara nilai FPro aktual dengan prediksi sistem dengan model *Ensemble*. Dengan capaian *Mean Absolute Error* (MAE) sebesar 0,1103, distribusi titik pada grafik menunjukkan korelasi linier yang kuat, di mana sebagian besar prediksi terkonsentrasi di sepanjang garis diagonal. Artinya, model tidak hanya pintar mengelompokkan kategori, tetapi juga jago menebak angka skor FPro dengan tepat. Model bekerja sangat baik dalam mengenali produk di kedua ujung skala, yaitu produk yang sangat minim olahan (0,0) maupun yang sangat tinggi tingkat pemrosesannya (1,0).



Gambar 3. Hasil Penghitungan Menggunakan MAE Pada *Ensemble*

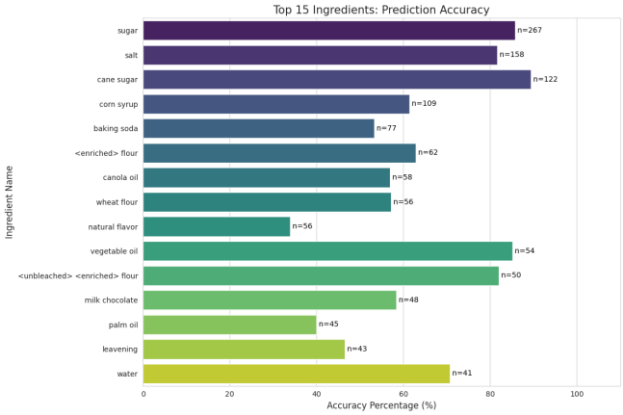
Untuk menjelaskan efektivitas klasifikasi menggunakan *Ensemble* dipetakan secara lebih detail menggunakan *Confusion Matrix* pada Gambar 4. Temuan yang paling menonjol adalah tingginya angka *True Positive* pada kelas *High*, di mana sistem berhasil mengidentifikasi secara tepat sebanyak 2.487 produk *ultra-proses*. Tingginya angka pada kelas ini menunjukkan kehandalan model dalam mendeteksi produk *ultra-proses*, namun perlu diakui bahwa hasil tersebut juga dipengaruhi oleh distribusi dataset yang didominasi oleh sampel kategori *High* dibandingkan kategori lainnya. Meskipun terdapat ketimpangan data tersebut, sistem berhasil sangat jarang membuat kesalahan logika yang ekstrem, seperti salah mengira produk dengan pemrosesan tinggi (*High*) sebagai produk alami (*Low*). Hal ini membuktikan bahwa strategi *Weighted Ensemble* efektif dalam menjaga stabilitas prediksi dan

meminimalkan kesalahan identifikasi yang fatal, yang juga didukung oleh kemampuan pengenalan bahan aditif utama secara akurat sebagaimana ditunjukkan pada Gambar 5.



Gambar 4. *Confusion Matrix* Pada *Ensemble*

Gambar 5 berfungsi untuk menjelaskan lebih lanjut dan menunjukkan bahwa keputusan klasifikasi sistem didasarkan pada penalaran terhadap komposisi bahan, bukan sekadar tebakan label kategori. Grafik ini memetakan 15 bahan dengan frekuensi kemunculan tertinggi dan mengukur keberhasilan sistem dalam menjalankan *Task* 2, yaitu mengidentifikasi bahan spesifik yang paling berkontribusi terhadap status pemrosesan produk tersebut. Capaian rata-rata akurasi yang tinggi pada 15 bahan dominan ini membuktikan bahwa strategi *Weighted Ensemble* berhasil menjaga konsistensi penalaran sesuai instruksi yang dirancang.



Gambar 5. 15 Bahan-bahan Berdasarkan Prediksi Akurasi Pada Ensemble

3.4.1 Koreksi Silang

Keunggulan utama dari *Weighted Ensemble* adalah kemampuannya melakukan koreksi terhadap kesalahan individu sehingga keputusan diambil dengan ikut menilai kemampuan individual model dalam kasus serupa. Analisis mendalam terhadap *log* EnsembleBaseline_Detailed_Analysis.json pada Tabel 5 memperlihatkan bagaimana sistem ini bekerja secara nyata. Pada kasus produk *Cream Soda*, model Llama-3.2 memberikan prediksi yang salah (*Medium*) karena sifatnya yang terlalu konservatif (*underestimate*). Namun, karena Gemma dan R1-Distill (yang memiliki bobot suara gabungan lebih

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

besar) memprediksi dengan benar (*High*), kesalahan Llama berhasil dianulir oleh sistem. Detail target sebagai berikut:

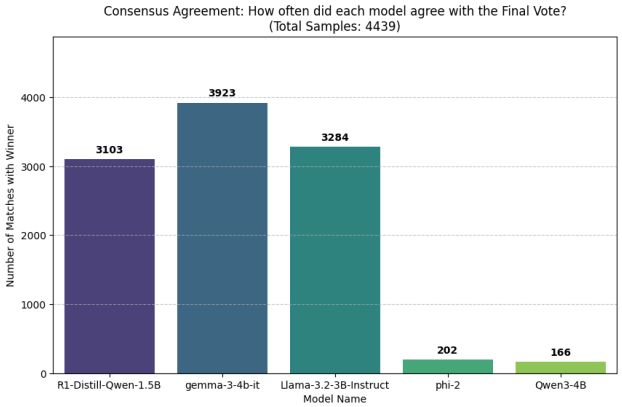
- Produk: *Cream Soda*
- *Ground Truth*: *High* (0.75)

Tabel 5. Studi Kasus Implementasi Weighted Ensemble

Model	Bobot Suara (Weight)	Prediksi Model	Analisis Peran
Gemma-3-4B	0.68	High (Benar)	bobot tertinggi.
Llama-3.2-3B	0.53	Medium (Salah)	*Error:* Terlalu konservatif, gagal mendeteksi bahaya pewarna.
R1-Distill	0.50	High (Benar)	-
Qwen3-4B	0.19	High (Benar)	-
Phi-2	0.10	Low (Salah)	*Model Gagal Memprediksi*
HASIL ENSEMBLE	Total Score	HIGH (WIN)	Sukses: Suara High (1.37) mengalahkan Medium (0.53) dan Low (0.10).

3.4.2 Analisis Kontribusi Mode

Pada Gambar 6 memvisualisasikan distribusi kontribusi setiap model terhadap keputusan *final* yang dimenangkan oleh sistem *Ensemble* (*Winning Vote Contribution*). Grafik ini menghitung frekuensi di mana prediksi individu model selaras dengan hasil akhir konsensus.



Gambar 6. Kontribusi Model

Analisis Dinamika Voting:

1. Terlihat bahwa Gemma-3-4B memiliki tingkat keselarasan tertinggi dengan keputusan *final*. Hal ini menandakan bahwa pengaruh bobot awal sangat mempengaruhi hasil keputusan akhir

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

serta pada kasus ini Gemma berfungsi sebagai "pemimpin opini" (*opinion leader*) dalam sistem *ensemble*.

2. Sebaliknya, Phi-2 menunjukkan tingkat kontribusi yang paling rendah. Fenomena ini mengindikasikan bahwa suara Phi-2 berada di pihak yang kalah (*minority dissent*). Dalam konteks *Weighted Voting*, hal ini adalah indikator positif yang membuktikan bahwa sistem berhasil memveto prediksi salah (biasanya nilai 0.0/Low) yang dihasilkan oleh Phi-2, sehingga kesalahan tersebut tidak merambat ke hasil akhir.
3. Pola grafik menunjukkan terbentuknya "Koalisi Cerdas" antara Gemma, R1-Distill, dan Llama yang voting bersamaan untuk mengalahkan suara *noise* dari model Phi-2. Stabilitas *F1-score* pada sistem *Ensemble* terbentuk dari konsistensi koalisi ini.

3.4.2 Perbandingan dengan Simple *Majority Vote*

Sebagai uji validasi metode, penelitian ini juga melakukan eksperimen komparatif menggunakan skema Simple *Majority Voting*, di mana setiap model memiliki hak suara yang setara (1 model = 1 suara) tanpa memandang kinerja *historis*-nya. Hasil eksperimen menunjukkan bahwa pendekatan tanpa pembobotan ini menghasilkan penurunan kinerja yang lebih jelek dibandingkan metode *Weighted Voting*. Rendahnya *F1-score* pada skema SMV disebabkan oleh fenomena "Dominasi Kegagalan". Sebagaimana dipetakan dalam evaluasi sebelumnya, terdapat disparitas kemampuan yang jelas antara model kelas atas (Gemma, Llama) dan model kelas bawah (Phi-2, Qwen).

Ketika dihadapkan pada produk yang kompleks, model-model kelas bawah mengalami kegagalan teknis (seperti parsing error atau halusinasi) dan secara default memberikan nilai prediksi 0.0 (*Low*). Dalam skema Simple Majority, akumulasi suara "*Low*" dari model-model yang gagal ini dapat mengalahkan suara "*High*" dari model yang benar, menghasilkan keputusan mayoritas yang salah. Detail target sebagai berikut:

- Produk: Pepperidge Farm 100% Whole Wheat Light Style Bread - 16oz
- Ground Truth: *High* (0.93)

Tabel 6. Studi Analisis Kegagalan Simple Majority Vote

Model	Bobot Suara (Weight)	Prediksi Model	Analisis Peran
Gemma-3-4B	1	High (Benar)	-
Llama-3.2-3B	1	Medium (Salah)	-
R1-Distill	1	Low (Salah)	-
Qwen3-4B	1	Low (Salah)	-
Phi-2	1	Low (Salah)	-
HASIL ENSEMBLE	Total Score	Low (WIN) (Salah)	Low:3 Medium:1 High:1

Pada kasus di atas, sistem *Simple Majority* tertipu oleh kuantitas model yang gagal. Tiga model memberikan label "*Low*", tanpa mekanisme pembobotan yang memberikan penalti pada model lemah (seperti memberikan bobot 0.0 pada Phi-2), sistem *Ensemble* justru memperkuat kesalahan kolektif alih-alih mengoreksinya. Hal ini menegaskan bahwa untuk arsitektur multi model dengan variabilitas

kemampuan yang tinggi, *Weighted Voting* adalah syarat untuk memastikan bahwa keputusan akhir didasarkan pada kompetensi model, bukan sekadar jumlah suara.

Secara keseluruhan, penelitian ini memposisikan strategi *Weighted Ensemble* sebagai alternatif terhadap model statistika tradisional dalam domain ilmu pangan. Jika dibandingkan dengan metode pada penelitian Ravandi dkk., penelitian ini membuktikan bahwa klasifikasi derajat pemrosesan (FPro) dapat dilakukan secara efektif melalui penalaran teks komposisi tanpa bergantung pada kelengkapan data nutrisi numerik. Selain itu, hasil riset ini juga memperkuat observasi Kamen dkk., mengenai keterbatasan model bahasa tunggal (rata-rata 55%), di mana mekanisme pembobotan dalam riset ini mampu meningkatkan akurasi hingga 70%. Penggunaan rangkaian model berskala kecil dan menengah (1B-4B parameter) memberikan nilai tambah berupa efisiensi komputasi, membuktikan bahwa stabilitas penalaran untuk tugas teknis yang kompleks dapat dicapai melalui arsitektur gabungan yang optimal tanpa harus selalu bergantung pada sumber daya komputasi yang besar sebagaimana dilakukan pada penelitian lainnya.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Berdasarkan serangkaian eksperimen yang telah dilakukan untuk mengevaluasi kinerja Ensemble *Large Language Models* (LLM) dalam analisis tingkat pemrosesan makanan pada *dataset* GroceryDB, dapat dapat disimpulkan bahwa, evaluasi menunjukkan adanya perbedaan kinerja yang berarti antar arsitektur model. Kapabilitas model tidak semata-mata ditentukan oleh ukuran parameter, melainkan sangat bergantung pada kemampuan *instruction following*. Model Gemma-3-4B menunjukkan performa terbaik sebagai model tunggal (F1-score68%), sementara model yang lebih kecil seperti Phi-2 mengalami *reasoning collapse* (F1-score 10%) akibat kegagalan memahami struktur instruksi yang kompleks. Hal ini membuktikan bahwa penggunaan model tunggal *open*-source berskala kecil memiliki risiko keandalan yang tinggi. Dan juga penerapan strategi *Weighted Majority Voting* berhasil dan efektif dalam memitigasi kelemahan model individual. Sistem *Ensemble* yang diusulkan berhasil mencapai F1-score 70% pada Task-1 dan 26% dalam Task-2, melampaui kinerja model tunggal terbaik. Peningkatan ini terjadi karena mekanisme pembobotan mampu menekan dampak suara dari model yang berkinerja buruk dan memperkuat konsensus dari model-model yang kompeten.

4.2. Saran

Berdasarkan temuan pada penelitian ini beberapa rekomendasi untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Salah satu tantangan utama dalam implementasi LLM adalah risiko halusinasi pada model generatif dalam menjawab pertanyaan user. Oleh karena itu Penelitian selanjutnya sangat disarankan untuk mengintegrasikan mekanisme RAG. Dengan menghubungkan model ke basis pengetahuan eksternal yang terverifikasi dengan harapan dapat meningkatkan F1-score dan meminimalkan kesalahan penalaran.
2. Seperti yang dijelaskan diatas dimana ensemble mampu meningkatkan F1-Score namun dirasa perlu adanya eksperimen lanjutan untuk membandingkan kinerja sistem Ensemble model kecil-sedang (seperti yang diusulkan penelitian ini) melawan satu arsitektur model berskala besar (Large-Scale Model, misal Chat-GPT 120B parameter). Studi komparatif ini krusial untuk menjawab pertanyaan efisiensi sumber daya: apakah agregasi "kecerdasan kolektif" dari beberapa model ringan mampu menyaingi atau bahkan melampaui kemampuan penalaran satu model raksasa (Single Strong Learner)

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

5. DAFTAR PUSTAKA

[1] B. Srour et al., "Ultra-processed food intake and risk of cardiovascular disease: Prospective cohort study (NutriNet-Santé)," The BMJ, vol. 365, p. l1451, 2019.

[2] C. A. Monteiro et al., "Ultra-processed foods: what they are and how to identify them," Public Health Nutrition, vol. 22, no. 5, pp. 936-941, 2019.

[3] Y. Guo, A. Ovadje, M. A. Al-Garadi, and A. Sarker, "Evaluating large language models for health-related text classification tasks with public social media data," Journal of the American Medical Informatics Association, vol. 31, no. 10, pp. 2181–2189, 2024.

[4] A. Kamen and Y. Kamen, "Majority rules: LLM ensemble is a winning approach for content categorization," arXiv preprint arXiv:2511.15714, 2025.

[5] B. Ravandi, G. Ispirova, M. Sebek, P. Mehler, A.-L. Barabási, and G. Menichetti, "Prevalence of processed foods in major US grocery stores," Nature Food, vol. 5, pp. 1-9, 2024.

[6] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[7] W. X. Zhao et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[8] Gemma Team, "Gemma: Open models based on Gemini research and technology," Google DeepMind, London, UK, Tech. Rep., 2024.

[9] A. Dubey et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[10] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv preprint arXiv:2501.12948, 2025.

[11] J. Bai et al., "Qwen Technical Report," arXiv preprint arXiv:2309.16609, 2023.

[12] M. Javaheripi, S. Bubeck, et al., "Phi-2: The surprising power of small language models," Microsoft Research Blog, Dec. 2023. [Online]. Available: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

[13] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824-24837.

[14] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.

[15] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in International Conference on Learning Representations (ICLR), 2023.

[16] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2021.