Universiteit Leiden
The Netherlands

Evaluating LLM-Generated Definitions for KBBI

Muridan, Galih Pradipta

Evaluating LLM-Generated Definitions for KBBI

Galih Pradipta Muridan

MA Thesis

Supervisor
Prof. Dr. Carole Tiberius

Second Reader
Prof. Dr. Stephan Raaijmakers

Leiden University Centre for Linguistics
The Netherlands
July 2025

Abstract

Explaining what words mean is one of the most difficult tasks in lexicography. For Kamus Besar Bahasa Indonesia (KBBI), this challenge is particularly complex due to Indonesia's linguistic diversity, with more than 700 regional languages that contribute vocabulary to the Indonesian language. Given this complexity, KBBI must maintain their linguistic standards while also accommodating regional variations. Recent progress in Large Language Models (LLMs) shows potential to support lexicographers in this task, but their effectiveness in Indonesian lexicography is still under investigation. This work evaluated LLM-generated definitions using IndoSBERT and human evaluation to assess whether LLMs can be used to generate KBBI-style definitions, while also examining how word frequency in an existing Indonesian corpus (idTenTen24) affects performance. Three models were tested: GPT-4o, Gemini 2.0 Flash, and Llama 3.2, using 1,244 new entries from the 31 October 2023 KBBI update with Chain-of-Thought prompting guided by KBBI's five definition writing principles. GPT-4o achieved the highest semantic accuracy (mean IndoSBERT score: 0.51), followed by Gemini 2.0 Flash (0.46) and Llama 3.2 (0.35). Human evaluation revealed a strong preference for LLM-generated definitions. The result shows that LLMs perform better for domain-specific entries and perform worse for specific language origin entries. Performance varied systematically between word classes. The prompt achieved moderate success 59.7% of the GPT definitions containing no principle violations compared to 51.4% for the KBBI definitions. Frequency analysis revealed that zero-frequency words, mostly from regional Indonesian languages (63.27%), presented the greatest challenge for LLMs. The findings suggest significant potential for LLM assistance in Indonesian lexicography, particularly for domain-specific terminology, but require careful adaptation for regional language terms.

Contents

Acknowledgements

I would like to thank the Indonesia Endowment Fund for Education (LPDP) from the Ministry of Finance Republic of Indonesia for the scholarship that allowed me to study at Leiden University and conduct this research to contribute to Indonesian language studies.

I am grateful to my supervisor for their guidance, insightful feedback, and continuous support throughout the writing of this thesis. I also sincerely thank my friends in Indonesia and in the Netherlands for their encouragement and helping me stay happy along the way.

Galih Pradipta Muridan
Rotterdam, 23 June 2025

Chapter 1

Introduction

1.1 Background

Explaining what words mean is one of the functions of a monolingual dictionary and one of the most difficult tasks in lexicography(Atkins and Michael. Rundell, 2008). Definition writing is not limited to just description, but also enables users to understand unfamiliar words in context, interpret the words successfully in new contexts, and ideally use them correctly and appropriately in their own communication (Atkins and Michael. Rundell, 2008). Samuel Johnson noted in 1755 (Johnson, 2021) that definition writing was the aspect of the dictionary work that he expected to attract the most criticism, acknowledging that he had not always been able to satisfy himself with his own definitions. This challenge in writing definitions is related to the fundamental difficulty of capturing accurate meanings while ensuring comprehensiveness for various user needs.

The complexity may increase when lexicographers must balance providing the information for user comprehension while avoiding unnecessary detail that could overwhelm the readers (Atkins and Michael. Rundell, 2008). For Kamus Besar Bahasa Indonesia (KBBI, or The Great Indonesian Dictionary), this challenge is especially complex due to Indonesia's linguistic diversity, with more than 700 regional languages that contribute vocabulary to the Indonesian language (Badan Pengembangan dan Pembinaan Bahasa, n.d.), and the responsibility of KBBI as the standard for the national language (Nurjaman, 2024). Consequently, KBBI must maintain their linguistic standards while accommodating regional variations. This dictionary is published and handled directly by the Indonesian Ministry of Education under the organization of

the Language Development and Cultivation Agency (Badan Pengembangan dan Pembinaan Bahasa). KBBI continuously updates its entries every year in April or October to add new words and adjust to the meaning changes within the Indonesian language. For these updates, KBBI has a particularly unique approach by opening public participation through KBBI Daring (the online version), launched in 2016 (Moeljadi, Kamajaya, and Amalia, 2017). Unlike a traditional dictionary, KBBI Daring runs as a transparent and non-anonymous crowdsourcing system where any registered Indonesian speaker can contribute by proposing new entries, definitions, or even corrections (Kamajaya, Moeljadi, and Amalia, 2017). This process involves a structured flow with different user groups, starting from registered users who give proposals, to registered editors who review submissions, and then to validators who make the final decision (Moeljadi, Kamajaya, and Amalia, 2017). These updates and processes are important to ensure that the dictionary remains credible as a reference for Indonesian speakers.

Although KBBI's collaborative model made a significant advancement in their workflow, the challenge of lexicographical work to maintain their linguistic standards while accommodating regional variations still exists. The emergence of computational tools may offer promising ways to improve the way dictionaries are made. In recent years, the progress in large language models (LLMs) has shown their potential to help lexicography in making dictionaries. These models showed the ability to solve many natural language tasks due to the data and linguistic knowledge they learned in pre-training (Jurafsky and Martin, 2025). LLMs existence and capabilities have also significantly impacted the landscape of lexicography. Schryver (2023) has made a comprehensive review of the first ten studies on ChatGPT applications in lexicography, revealing both promising capabilities and significant limitations that warrant further investigation.

Several studies have examined different aspects of AI-generated lexicographic tasks, including definition writing. Michael Rundell (2023) examined ChatGPT's definition handling capabilities, finding that technical terms were well defined but weak with polysemous words. Hallucination was also another problem that was found, where the model generated incorrect information with confidence. McKean and Fitzgerald (2023) found that GPT failed to simplify definitions from adult level to child level. The result was awkward and clunky definitions. Their work also mentioned concerns about how LLMs trained on general English web content may underrepresent minorities in dictionaries and perpetuate stereotypes. Lew (2023) demonstrated that, with prompt engineering, ChatGPT-generated definitions in COBUILD-style entries can be "practically indistinguishable in quality from those written by highly trained hu-

man lexicographers." Recently, Cai et al. (2024) proposed OxfordEval as a metric to evaluate generated sentence examples compared to the Oxford dictionary as a baseline. For explanations and definitions, Rees and Lew (2023) evaluated the generated explanations by measuring the effectiveness for English learners. Pham et al. (2025) evaluated the LLM definitions for English by comparing the similarity to three dictionaries using cosine similarity. In relation to definition writing, GPT performance on automated definition extraction was also explored. Tran et al. (2023) found that GPT could distinguish between definition and non-definition in Slovene text corpora quite well in an unstructured text sequence.

This rapidly evolving field is still dominated by the English language, prompts and data are processed "through English", which causes them to have English linguistic characteristics and structures for other languages (Schryver, 2023). Although LLMs developments demonstrate the potential of LLMs for definition generation, they are still limited and urgently need to be evaluated in language contexts other than English, particularly for the Indonesian language, with 300 million speakers worldwide (Grehenson, 2022). Indonesian language has a significant geographical distribution, with 52 countries offering formal Indonesian language education (Office of Assistant to Deputy Cabinet Secretary for State Documents Translation, 2023), but despite this large number, the Indonesian language is still underrepresented in LLM research compared to its demographic size. Although there are rapid advances in the field of Indonesian NLP, such as comprehensive benchmarks for common Indonesian NLP tasks like IndoNLU (Wilie et al., 2020), IndoLEM (Koto et al., 2020), and even regional Indonesian languages like NusaCrowd (Cahyawijaya, Lovenia, Aji, et al., 2023) and NusaWrites (Cahyawijaya, Lovenia, Koto, et al., 2023), exploration of LLM capabilities for lexicography in Indonesian language remains unexplored. This need becomes notably urgent given that KBBI, as mentioned before, need to keep maintaining their linguistic standards while accommodating regional language variations.

1.2 Research Questions

RQ1: To what extent can large language models (GPT-4o, Gemini 2.0 Flash, and Llama 3.2) generate semantically accurate definitions for new Indonesian words, as measured by IndoSBERT, and how do they compare with human-written KBBI definitions in terms of overall quality and user preference?

RQ2: How do LLM definition generation results vary across word categories (word class, definition type, domain-specific words, and language origin) and what patterns

can we see from corpus frequencies?

RQ3: To what extent do LLM-generated definitions adhere to the KBBI's guidelines (five core principles and appropriate definition types) when guided by Chain-of-Thought prompting, and which principles are frequently violated?

Chapter 2

Methodology

2.1 Dictionary Definition

Word meanings and dictionary definition are two different things. The meaning of a word is constituted by its contextual relations (Cruse, 1986). That means that the meaning of a word can be infinitely rich, but will depend heavily on the context. In contrast, a dictionary definition is what lexicographers make to register meanings in a language. It has practical purpose of fulfilling the communicative needs of dictionary users. After word sense disambiguation, definition writing is considered the most difficult aspect of monolingual lexicographer's job (Atkins and Michael. Rundell, 2008). As noted by Hanks (Durkin, 2016), definition writing requires lexicographers to devote more time, effort, and money than anything else.

The complexity comes from the multiple functions that definitions must provide, users need to decode (understand words in context) or encode (find words to express meaning), requiring definitions that enable understanding in contexts they encounter, interpretation in new contexts, and appropriate usage of the word (Atkins and Michael. Rundell, 2008). There are many approaches on making a good definition, but the most important aspect is that good definitions succeed when they have both the appropriate content and form for their intended users. Noted by Bolinger (1965), effective definitions work as "hints and associations that will relate the unknown to something known." It is also important to note that different dictionaries may define the same concept in different ways, as the selection of facts for definitions is not a scientific enterprise but varies appropriately from one dictionary to another (Atkins and Michael. Rundell, 2008). Explanations that satisfy specialists may not serve common readers,

emphasizing that what matters is not the writer's intention but the reader's interpretation (Johnson, 2021). Thus, the precise configuration of definitions is determined by the needs and skills of each dictionary's target users (Atkins and Michael. Rundell, 2008). Each dictionary establishes its own guidelines to ensure consistency and approproateness for its intended users and context. Similarly, KBBI has their own specific guidelines (Pelindungan Bahasa dan Sastra Badan Pengembangan Bahasa dan Perbukuan Kementerian Pendidikan dan Kebudayaan, 2019) for definition writing that may reflect the needs of Indonesian speakers and the unique characteristics of the Indonesian language. These guidelines, which shape the definitions within KBBI which were also used as a guide for the Large Language Models (LLMs) to generate definitions, is explained in the following section.

2.2 KBBI

Kamus Besar Bahasa Indonesia (KBBI) is published by Badan Pengembangan dan Pembinaan Bahasa (Agency for Language Development and Cultivation) or Badan Bahasa (The Language Agency). It is currently under the Ministry of Primary and Secondary Education, Republic of Indonesia (Badan Pengembangan dan Pembinaan Bahasa, n.d.). The making of the Indonesian dictionary started in 1952 by W.J.S. Poerwadarminta (Endarmoko and Moeliono, 2008), one of the lexicographers of Lembaga Penyelidikan Bahasa dan Kebudayaan (Agency for Language and Culture) under the University of Indonesia. The dictionary was titled Kamus Umum Bahasa Indonesia (General Dictionary of Indonesia). The primary goal of making an Indonesian monolingual dictionary at that time was to make a great dictionary or a standard dictionary for the Indonesian language so that Indonesians could have a gold standard or reference for their language (Peristilahan, 2022). The organization eventually changed its name to Pusat Bahasa (Center for Language) and then Badan Bahasa (The Language Agency), and it was transferred directly under the government of Indonesia. The first edition of KBBI was released in October 1988 by Badan Bahasa, titled *Kamus Besar Bahasa Indonesia* (KBBI), it translates to *The Great Dictionary of the Indonesian Language* with approximately 62.000 lemmas (Tim Penyusun Kamus, Pusat Pembinaan dan Pengembangan Bahasa, 1988). There are a printed version and an online version of KBBI. The printed versions follow sequential numbering (KBBI I, II, III, IV, V, VI), with the last published printed version being KBBI V in 2018 (Sunendar, 2018). KBBI VI was planned for publication in 2024 (Aziz, 2023) but there are no information has been released about it since then. Those printed versions are still

available, but they are not as frequently updated as the online version.

2.2.1 Online Version of KBBI

The online version of KBBI is called KBBI Daring (KBBI Online), launched on 28 October 2016. KBBI Daring can be freely accessed at `kbbi.kemdikbud.org`. KBBI Daring is updated twice a year, so it is more up-to-date than the printed version. KBBI Daring is made for both its lexicographers and its users. There are two main features users can do with the online version, looking up words in the dictionary, and suggesting new words or changes to the entries in the dictionary. The KBBI team allows anyone, simply by registering, to suggest changes or propose new words for the dictionary (Pelindungan Bahasa dan Sastra, 2019).

Figure 2.1: A sample of an entry in KBBI Daring, explaining noun *taman* (park), which contains several components: a headword (1) with entry labels indicating that the headword has multiple entries, senses and the definitions (2), the word class labels (3), example (4), additional labels, in this case *ark*, means it is an archaic word (5), derivative forms (6), and frequent collocations or fixed expressions (7). Each entry has interactive buttons for registered users to propose reviews for the word or access detailed information about the entry (8), shortcut button for the entry's thesaurus (9), and occasionally etymological information (10). Accessed: 20-06-25

Whenever users propose new words or changes, the suggestion will go through several people in the KBBI team before being accepted. It will go through editors,

redactors, and then validaters (Moeljadi, Kamajaya, and Amalia, 2017). The editors are mostly language students, freelance linguists, and lexicographers contracted by the KBBI team to be the first to review user suggestions. In addition to users, editors are also encouraged to be active in proposing new words and making changes (Holy, 2024). After that, the redactors will review the work of the editors and the user suggestions that the editors have accepted. The redactors are the lexicographers of Dinas Bahasa, the core team of KBBI. The redactors are also allowed to propose new words, but their main task is to review the editors' work to give it to the validators. Validators are those who make the final decision for any changes (Holy, 2024).

2.2.2 KBBI-style Guideline

For evaluating and generating them with large language models, understanding KBBI's approach on writing definition is important. Thus, the KBBI guideline will be explained and summarized in this section. This will be the foundation for the prompt that will be explained in Section 2.5. The guideline is called *Petunjuk Teknis Penyusunan Kamus Ekabahasa* (Technical Guideline on Building Monolingual Dictionary) (Pelindungan Bahasa dan Sastra Badan Pengembangan Bahasa dan Perbukuan Kementerian Pendidikan dan Kebudayaan, 2019). This comprehensive guideline provides the theoretical foundation and practical standards for KBBI's editorial team. The guideline is based on other lexicography guidelines and theories. It directly cites Conklin (1975) and Kaye (1999) many times in some of the explanations. After that, they adjust the guideline accordingly with Indonesian cases and examples. The guideline has a systematic approach to writing definitions, which starts by categorizing different types of definition, prescribes specific formatting and content requirements, and then the principles to follow. The following definition types are distinguished in the guidelines.

I. **Analytical definition** This is a classical definition that is used to explain what words are based on their generic characteristics (*genus*) and their distinguishing features (*differentia*). It is important to correctly identify the *genus* and the *differentia* for words with this type of definition.

Word	KBBI Definition	Translation (English)
membelek	**melihat** dengan teliti; mengincar; membidik	**to see** carefully; to observe; to aim at something with intention
mencerling	**melihat** ke sebelah kanan atau kiri; menjeling; mengerling	**to see** sideways; to peek; to look quickly to the side (often subtly or slyly)

Both words have the same generic characteristics *"melihat"*, but they are different in **nuance and usage**.

- **membelek** focus on seeing for the sake of observing carefully.
- **mencerling** emphasize on the method and the direction of the glance.

II. **Encyclopedic definition** This type of definition explains encyclopedic knowledge instead of linguistic knowledge. In practice, the encyclopedic definition overlaps with the linguistic definition. The encyclopedic definition tends to be more detailed in explaining a concept or an object, starting from the form to any nature phenomenon relating to the object. In contrast, general definitions explain a concept based on the their genus and distinguishing features, focusing on linguistic understanding rather than comprehensive factual knowledge. For example, the sun might be defined as a star around which the earth orbits in general definition, while an encyclopedic definition would explain it as a celestial object made of gas, with a vast amount of energy, and include more details on its size and components.

III. **Synonym definition** This type of definition uses another word with the same meaning or almost the same meaning as the word that is being defined. As the guideline stated, the synonym definition was previously used to secure space in traditional dictionaries (Pelindungan Bahasa dan Sastra Badan Pengembangan Bahasa dan Perbukuan Kementerian Pendidikan dan Kebudayaan, 2019). As the risk of circularity is high with synonym definitions, the guideline encourages to complement it with another type of definition.

IV. **Antonym definition** Similar to the synonym defintions, but for words with opposite meanings. Negation words are typically used in antonym definitions. Examples of negation words in Indonesian are *tanpa* ('without'), *bukan* ('not' for nouns), and *tidak* ('not' for adjectives and verbs).

Word	KBBI Definition	Translation (English)
sedikit	**tidak** banyak; **tidak** seberapa	**not** a lot; **not** high in quantity
nirkabel	**tanpa** menggunakan kabel	**without** using cable

V. **Cross-references**[1] This is used to guide dictionary users to more accurate information in other parts of the dictionary. It is usually written with the arrow symbol (→) or *lihat* ('see') that points to another entry. In KBBI, it is generally used for non-standard words, which is then referred to the standard version of it. It is also used to guide the user from the derivative words to their lemma. The other use of it is for dependant words or words that mostly only occur together with another word.

VI. **Ostensive definition** This type of definition refers directly to the object or has a strong association with the object. The guideline explicitly states that this is used only for defining colors.

Word	KBBI Definition	Translation (English)
merah	warna dasar yg serupa dengan darah; mengandung atau memperlihatkan warna yang serupa dengan merah	basic color that is similar to blood; contains color or showing color similar to red
hijau	warna yang serupa dengan warna daun pada umumnya	color similar to a common leaf

After explaining the appropriate definition type for a word, the guideline outlines the principles. There are five principles that KBBI definitions should adhere to in order to make a good definition. These are based on Conklin (1975) and have been adjusted accordingly.

[1] this definition type is mentioned in the guideline but is not used in the evaluation for reasons explained in section 2.3

1. **Self-explanatory** The definition has to be self-explanatory, which means that every word in the definition should also be defined in the dictionary. It is unacceptable to use a word that is not explained in the dictionary because the reader cannot understand what it is while using the dictionary. The definition should explain the word without the need for other references.

2. **Obscurum per obscurius** The words used in the definitions should not be harder to understand than the word being defined.

3. **Substitutionality** The definition can be substituted for the word. This means that the definition should begin with the same word class as the word being defined. The definition of nouns should begin with a noun, the definition of verbs should begin with a verb, and so on. On the other hand, for a grammatical word, the definition can be made with an explanation based on its function and context. In Indonesian, copula words are common when giving explanations, such as 'adalah' and 'merupakan'. In English, these are typically translated as 'is' or 'are'. These copula words in Indonesian should be avoided in the definition because they do not satisfy substitutability and do not help to identify the *genus* of the word being defined.

4. **Avoid Circular Definition** Circular definition is when the word being defined is used directly or derivatively in its own definition, thus failing to provide new semantic information (Atkins and Michael. Rundell, 2008). Such definitions violate the principle of substitutability and do not help in understanding the term. According to the guideline, this generally happens with synonym definitions and when there is a lack of words to define a word. To avoid this, the synonym definition should also be explained in more detail or defined with another type of definition.

5. **Specific but not too specific** The definition must be specific but not too specific. According to the guideline, to ensure that the definition is specific enough, the definition must have the correct *genus* and a minimum of one of its most distinguishing features.

2.3 Dataset

To test whether LLM-generated definitions can follow KBBI-style guidelines, we used data from the KBBI update on 31 October 2023, which is the most recent update

that includes an official word list. Although there were later updates on 4 December 2024 and 30 April 2025, no official word lists were released with those versions. In the official statement of the 2023 update, a total of 1,202 new entries were added. Those data were collected by extracting directly from the latest KBBI phone application (Bahasa, 2017). Then it is filtered, only definitions listed from the official word list were taken and compiled as a single dataset. After checking, there were 1196 new entries with a total of 1,244 new senses. There were six items in the official list that were duplicates or blanks, so they have not been included in the dataset and have been treated as errors.

Table 2.1: Distribution by Word Class and Definition Type

(a) Word Class (N = 1,199)

Word Class	Count	%
Noun	956	79.7
Verb	160	13.3
Adjective	67	5.6
Adverb	10	0.8
Affix	2	0.2
Phrase	2	0.2
Numeral	1	0.1
Particle	1	0.1
Total	**1,199**	**100.0**

(b) Definition Type (Labels = 1,286). There are 87 definitions with multiple labels

Definition Type	Count	%
Analytical	967	75.2
Synonym	213	16.6
Abbrev. Expansion	84	6.5
Encyclopedic	14	1.1
Other	7	0.5
Antonym	1	0.1
Total	**1,286**	**100.0**

Table 2.2: Language Origin Distribution (N = 1,199)

Language Origin	Count	Language Origin	Count	Language Origin	Count
No language origin	893	Wolio	5	Awyu	3
Sundanese	63	Bugis	4	Yalahatan	2
Javanese	32	Muna	4	Nias	2
Batak	27	Kimaam	4	Ambon Malay	2
Sanskrit	21	Min. Tonsawang	4	Bahau	2
Belitung Malay	17	Tolaki	4	English	2
Jambi Malay	13	Chinese	4	Sahu	2
Gorontalo	7	Arabic	4	Alune	2
Dayak	7	Kalimantan Malay	4	Riau Malay	2
Japanese	7	Acehnese	3	*23 languages*	
Kei	6	Benuaq	3	*w/ 1 entry each*	*23*
Kulisusu	6	Balinese	3		
Toraja	6	Sangir	3		
		Kur	3		
Total: 59 different language origins					

All definition types were labeled manually. 87 entries have multiple definition

Table 2.3: Languages with Single Entries

Languages with 1 entry each (23 total)			
Gane	Modole	Papuan	Ciacia
Passer	Boing	Luhu	Bacan
Sentani	Minangkabau	Kenyah	Korean
Madurese	Banjar	Tobelo	Karey
Russian	Lampung	Manado Malay	Minahasa
Medan Malay	Spanish	Minahasa Tonsea	

types. For example, the word *campoang* is defined as "orang yang membersihkan kapal; perawat kapal" ('a person who cleans ships; ship caretaker'), which has analytical definition and synonym definition, respectively. Similarly, *saus cocol* is defined as "saus yang disajikan sebagai cocolan untuk menemani hidangan; saus celup" ('sauce served as a dip to accompany dishes; dipping sauce'), showing analytical definition and synonym definition. In this dataset, 45 cross-references have been ignored because they refer to existing definitions from older updates, leaving us with 1,199 definitions. During the manual labeling process, several definition types that are not explicitly included in the KBBI guidelines were identified. The most notable of these is the abbreviation expansion, where entries are defined only by their expanded form without any additional explanation. This definition type was discovered when manually categorizing the dataset, and these entries follow a distinct pattern from the definition types outlined in the KBBI guidelines. For example, KBBI defines *bakhor* only as "tembakan kehormatan" ('ceremonial gunfire'). These abbreviation expansion definitions may present a challenge for LLM since they were not included in the prompting examples or instructions, as they fall outside the scope of the official KBBI definition writing guidelines.

For the word class, they were manually labeled due to the absence of word class labels in 357 of the 1,199 entries (29.8%). Word class labels primarily consist of part-of-speech. However, there are some labels referring to the grammatical of morphological elements rather than part-of-speech like bound form, prefix, infix, and clitics. Domain and language origin labels are already there as part of the dictionary's focus on capturing local languages, foreign languages, and domain-specific words used in Indonesia (Badan Pengembangan dan Pembinaan Bahasa, n.d.). The number of entries with domain and language origin labels in the data is 40%. There is a label indicating the sense number of a word in the dataset, whether the definition is the primary sense, the second sense, and so on. However, similar to word class labels, not all entries appear

to have this information. Based on the available data, there are 99 entries marked as non-primary senses. This means that 99 entries are new senses for already existing words.

The frequencies of each word in the official word list were also checked. This was done using the idTenTen24 (Jakubiček, Miloš and Kilgarriff, Adam and Kovář, Vojtěch and Baisa, Vít and Suchomel, Vít, 2017) corpus within the Sketch Engine. It is the largest and latest corpus available in the platform for the Indonesian language. Frequency data was used as an indicator of how common and novel the word is within the latest dataset. Given that this analysis used KBBI's updates for a dataset, rare words are expected to appear. Frequency data were used to examine the impact of commonness on definition generation results, such as the challenges posed by less frequent words in the corpus. The frequency analysis was only performed for all single-word entries in the word list (812 entries). Multiword entries (387 entries) were excluded due to the difficulties in checking them within the Sketch Engine and the variety of multiwords in the dataset.

2.4 Large Language Model

Large language models (LLMs) are a significant advancement in natural language processing (NLP) that operate on probabilistic principles by learning statistical patterns and relationships from massive text data (Jurafsky and Martin, 2025). LLMs are built on an architecture called the *transformer*, a neural network design that revolutionized language processing using its innovative attention mechanism (Vaswani et al., 2023). Unlike traditional models where they analyze sentences sequentially, the transformer can analyze all sentences simultaneously using self-attention to determine which words in a sequence are most relevant to understand each other. This attention mechanism allows the model to "focus" on different parts when generating each word, similar to how humans might emphasize certain words when understanding or producing language. We can interact with the models primarily through prompting (Liu et al., 2021), where we provide instructions or examples that guide the model to our desired result. For lexicographic applications, this means that we can design prompts that specify the style and format of dictionary definitions. Basically, we instruct the models to follow the style of our guideline, which will be explained in Section 2.5. It is important to note that LLMs can make mistakes. One of the most common mistakes is called hallucination when LLMs generate convincing texts that sound good but have no factuality in them at all (Marcus and Davis, 2020). This is why constant evalua-

tion and human oversight on generated texts is still very important. In a context of dictionary definition, hallucination could manifest as well-structured definitions that are semantically incorrect.

There are many models that we could use to generate definitions. For this study, three models were chosen mostly due to their accessibility and low computational usage. GPT-4o (OpenAI et al., 2024) is accessible using the OpenAI API key through the OpenAI website (`https://platform.openai.com/`) and is by far one of the most popular LLMs. It is important to note that ChatGPT is different from GPT. ChatGPT is the model fine-tuned for conversational task, and GPT is what ChatGPT is based on and capable of doing many tasks. Gemini 2.0 Flash (Team et al., 2025) is Google's large language model that offers performance with API access through the Google AI Studio (`https://aistudio.google.com/`). This model is chosen for its strong multilingual capabilities, which is particularly relevant for Indonesian language tasks, and its ability to handle complex reasoning tasks effectively. Llama 3.2 (Grattafiori et al., 2024) (3B parameters) is included as a smaller open-source alternative that can be run locally using Ollama (`https://ollama.com/`). Despite its smaller size compared to GPT-4o and Gemini 2.0 Flash, Llama 3.2 is not reliant on cloud computing resources, so we included Llama 3.2 to explore the performance of smaller and more open LLM options.

Model	**Version**	**Parameters**	**API/Platform**
Llama 3.2	3b	3.21B	Ollama
GPT-4o	2024-05-13	200B - 1.8T (estimated)	OpenAI API
Gemini 2.0 Flash	001	>200B (estimated)	Google Studio API

Table 2.4: Large Language Models Used in the Research

For running Llama 3.2 locally using Ollama, the following hardware specifications were used:

Component	**Specification**
Processor	Intel Core Ultra 7 155H
System RAM	16 GB
Graphics Card	NVIDIA GeForce RTX 4060 (Laptop)
VRAM	8 GB GDDR6
Storage	1 TB SSD

Table 2.5: Hardware Specifications for Local Model Inference

2.5 Prompting

For this research, a method called *prompting* is used to guide a language model (LLM) to write dictionary definitions. Prompting means giving the model a carefully designed instruction so that the models can understand what kind of task they need to do. There are many types of prompt, each serving different purposes for different tasks (Fagbohun, Harrison, and Dereventsov, 2024). For this study, a method called Chain-of-Thought (CoT) (Wei et al., 2023) prompting is used. CoT prompting was selected to incorporate KBBI's guideline multi-step reasoning, where they select an appropriate definition type, analyze the word's part of speech, and apply KBBI's five core principles of writing definitions. This approach aligns with CoT's strength in guiding models through sequential decision-making processes. It is important to note that the prompts were designed to provide only the target word without additional information in this study, such as word class labels or examples, so we can see whether LLMs could autonomously follow the KBBI guidelines, including the step of analyzing the word's part of speech.

Prompt Text The full prompt used in this research is written below. It contains role instructions, thinking steps, definition rules, and examples.

```
You are an expert Indonesian linguist specializing in creating
dictionary definitions in the style of Kamus Besar Bahasa
Indonesia (KBBI).

INSTRUCTIONS (Do not include in output):
1. For each word, ANALYZE its part of speech.
2. Choose ONE definition type UNLESS the word naturally requires
    more than one:
   - Analytical Definition (genus + differentia)
   - Encyclopedic Definition (can be detailed, but avoid to be
       too encyclopedic)
   - Synonym Definition (use another word with the closest or
       the same meaning)
   - Antonym Definition (use negation in the beginning of
       definition)
   - Ostensive Definition (define something as if pointing at
       the object)
3. If needed, COMBINE multiple definitions within one sentence
    using a semicolon (;).
```

```
   Example: komputer: alat untuk mengolah data secara
       elektronik; laptop
4. CREATE the definition using KBBI principles:
   - The definition must be self-explanatory
   - Avoid using words more difficult than the word being defined
   - Match the part of speech
   - Do not use copula words like "adalah" or "merupakan"
   - Avoid circular definitions
   - Be specific but not too detailed

OUTPUT FORMAT:
Return ONLY the final definition without any explanation or
    extra text.

REFERENCE EXAMPLES:
Analytical - pohon: tumbuhan yang berbatang keras dan besar...
Encyclopedic - matahari: benda angkasa, titik pusat tata surya...
Synonym - kudus: suci; murni
Antonym - nirkabel: tanpa menggunakan kabel
Ostensive - biru: warna dasar yg serupa dng warna langit...
```

The prompt has three main parts:

1. **Role**: The model is told to act like an expert linguist.
2. **Reasoning Steps**: The word must be analyzed, select a suitable definition type, and follow the KBBI guidelines.
3. **Examples**: Five examples from the guideline are given for the model to follow, one for each type of definition.

This approach helps the model think step-by-step, but only gives the final definition as output. By combining clear instructions and examples, the prompt supports the model in producing consistent and correct definitions that follow the style of KBBI.

2.6 Evaluation

The definitions generated by LLM will be evaluated with semantic similarity scores using IndoSBERT (Diana, 2023), a specialized model for Indonesian texts, inspired by the Sentence Transformers (SBERT) (Reimers and Gurevych, 2019). SBERT is a

Python library that is generally used for creating and utilizing advanced text embeddings. It makes linguistic tasks such as semantic search, comparing sentence similarity, and identifying paraphrases possible by using both Sentence Transformer models for generating embeddings and Cross-Encoder models for calculating similarity scores.

IndoSBERT is a modification of IndoBERT that has been fine-tuned using the SBERT siamese network scheme. This model was fine-tuned with the STS Dataset (2012-2016), which was machine translated into Indonesian language. In this work, we used the definitions from KBBI as a reference for the evaluation. This approach measures the semantic equivalence between generated and reference definitions, focusing on meaning preservation rather than lexical or syntactic similarity. The model produces similarity scores ranging from 0 to 1, where higher scores mean greater semantic overlap between the generated definition and the reference definition.

To evaluate definition quality beyond semantic similarity, we conducted a human evaluation with three tasks: rating the quality of the definitions, their adherence to the principles in the KBBI guideline, and preference assessment between generated and reference definitions. Based on the results of the semantic similarity scores, the GPT-4o definitions consistently performed better than those produced by Gemini 2.0 Flash and Llama 3.2, so we select the definitions generated by GPT-4o to be evaluated in the human evaluation process. For the first and second task, the evaluators examined the compliance of the definitions with the KBBI guidelines, including clarity, precision, and appropriate meaning based on their own understanding. The rating used a 5-point Likert scale (Schuff et al., 2023) ranging from -2 (very bad) to +2 (very good), with 0 representing a neutral assessment. For the second task, the evaluators were asked which principles are violated by that definition if there are any. In the third task, they were presented with pairs of definitions, one from KBBI and another from GPT, and then asked to indicate their preference based on overall quality.

2.7 Reproducibility

Reproducibility is important to ensure that the findings can be verified, especially in computational linguistics, where it is considered a challenge (Arvan, Pina, and Parde, 2022). To support the reproducibility of this study in generating Indonesian definitions, all data, codes, and output are made publicly available at `https://github.com/galihpm/LLMandKBBI`. The parameters used in this study are detailed in Table 2.6. Temperatures of 0.3 was chosen to ensure consistent definition quality while allowing some variation in language expression for different word class and defini-

tion types. Token limit was set to accomodate definition length similar to KBBI's typical length without excessive details. Max retries were given to ensure the production of definition for all words in dataset, considering the occasional system unavailability.

Table 2.6: Model Configuration and Parameter Settings

	Configuration
Models & Access	GPT-4o (OpenAI API) Gemini 2.0 Flash (Google AI Studio API) LLaMA 3.2 3B (Ollama)
Temperature	0.3
Max Tokens	500 per definition
Max Retries	5 attempts

Chapter 3

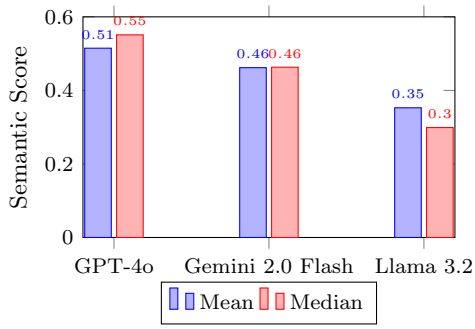
Results

3.1 IndoSBERT Semantic Similarity Analysis

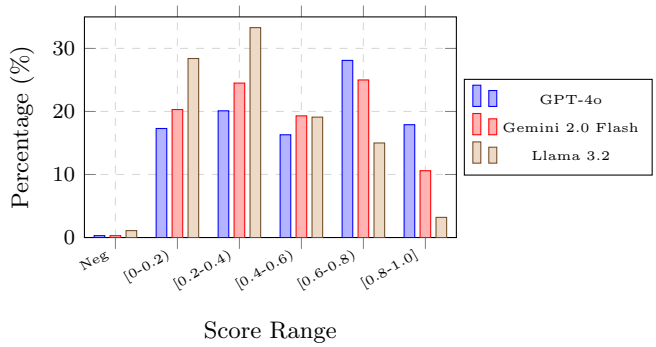
3.1.1 Overall Performance

To analyze the performance of the generated Indonesian definition across the three language models, we begin by examining the overall performance of semantic similarity in IndoSBERT semantic similarity scores before investigating how specific linguistic and lexical factors influence the semantic similarity scores. As shown in Figure 3.1a, GPT-4o achieved the highest mean semantic score of 0.51 and a median of 0.55, followed by Gemini (mean 0.46, median 0.46) and Llama 3.2 (mean 0.35, median 0.30).

Higher median scores for GPT and Gemini compared to their means suggest that there is a positive skewness in their performance distributions. This suggest more instances of high-quality definitions in both GPT-4o and Gemini 2.0 Flash. The score distribution can be seen in Figure 3.1b. The score distribution histogram reveals distinct performance patterns in all three models. GPT-4o shows the best distribution, with 46% of the generated definitions achieving high semantic similarity scores (0.6-1.0 range), with 17.9% reaching the highest performance level (0.8-1.0). In contrast, Llama 3.2 shows a heavily skewed distribution towards lower performance, with 61.7% of the definitions falling in the two lowest score ranges (0-0.4), and only 18.2% achieving scores above 0.6. Gemini is in the middle position, concentrating 44.8% of the results in the lower performance categories (0-0.4).



(a) Mean and Median of Models' Semantic Scores



(b) Score Distribution Comparison Across Models

Figure 3.1: Semantic Score Analysis: (a) Central tendency comparison showing overall performance, (b) Distribution patterns across score ranges

3.1.2 Word Class and Definition Type Influences

After we understand the overall performance of each mode, we now see whether the scores are influenced by word classes and definition types. The semantic similarity scores across the word classes have a consistent performance in all three language models shown in Table 3.1. Adverbs have the highest performance across all models, achieving mean scores of 0.677 (GPT-4o), 0.467 (Llama 3.2), and 0.542 (Gemini 2.0 Flash), despite representing the smallest sample size with only 10 words. Verbs consistently ranked second with scores of 0.53, 0.35, and 0.49 respectively, followed by nouns (0.51, 0.35, 0.46) that made up the largest portion of the dataset with 956 words. Adjectives consistently performed lowest among major word classes, scoring 0.52, 0.33, and 0.43 in all three models.

Table 3.1: Semantic Scores by Word Class (Ranked by Average Score)

Word Class	Count	GPT-4o	Llama 3.2	Gemini 2.0 Flash	Average
adverb	10	0.68	0.47	0.54	**0.56**
verb	160	0.53	0.35	0.49	**0.46**
other	6	0.55	0.33	0.47	**0.45**
noun	956	0.51	0.35	0.46	**0.44**
adjective	67	0.52	0.33	0.43	**0.42**

For the definition types, analytical definitions are the majority in the dataset with 967 definitions and showed consistent performance, ranking first or second in all models with scores of 0.53 (GPT-4o), 0.37 (Llama 3.2), and 0.474 (Gemini 2.0 Flash).

Synonym definitions, the second most frequent type with 213 definitions, showed moderate performance (0.51, 0.34, 0.47). Abbreviation expansion definitions consistently underperformed in all models, particularly with Ollama (0.23), while encyclopedic definitions showed the most variable performance.

Table 3.2: Semantic Scores by Definition Type (Ranked by Average Score)

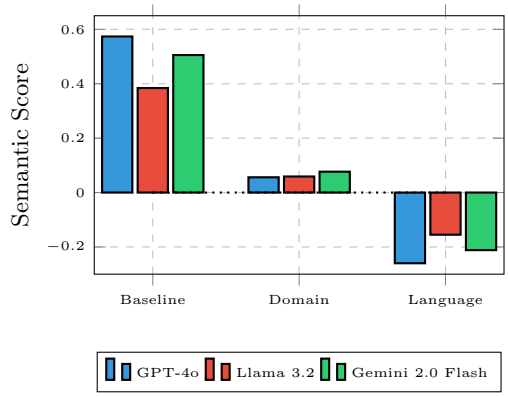
Definition Type	Count	GPT-4o	Llama 3.2	Gemini 2.0 Flash	Average
Antonym	1	0.83	0.42	0.46	**0.57**
Other	7	0.64	0.39	0.65	**0.56**
Analytical	967	0.53	0.37	0.47	**0.46**
Synonym	213	0.51	0.34	0.47	**0.44**
Encyclopedic	14	0.42	0.38	0.36	**0.38**
Abbreviation Expansion	84	0.5	0.23	0.38	**0.37**

Abbreviation expansion low scores may have occurred because they are not mentioned in the guideline. Therefore, no examples or explanations of abbreviation expansion were provided to the models at all. This type of definition was found when manually labeling data, where abbreviations are defined only by their expanded form, without any additional explanation provided. For example, KBBI defines *bakhor* only as *tembakan kehormatan* ('ceremonial gunfire'), or the word *FIK*, is defined only as *Fakultas Ilmu Kesehatan* ('Faculty of Health Sciences'). The LLMs made totally nonsense definitions for those two words. GPT-4o defined *bakhor* as *wewangian yang dibakar untuk menghasilkan aroma harum; dupa* ('a resin that is burned to produce a fragrant aroma; incense'), and Gemini 2.0 Flash defined *FIK* as *khayalan; angan-angan* ('imagination; daydream'). All models were not given both examples and instructions for recognizing this type of definition, therefore, failed to understand these words as shortened forms of longer terms or phrases. This is why they produced inaccurate definitions, resulting in low scores.

3.1.3 Domain and Language Origin Influences

In this section, we examine how domain and language origin labels influence semantic similarity scores. In Figure 3.2, we can see the score differences between words from specific domains and from specific language origins compared to other data without language or domain label. Data from specific domains scored higher throughout all models, which is consistent with the observation of Michael Rundell (2023) that LLMs performed well in technical terms. However, words originating from specific languages achieved lower scores than baseline, with decreases ranging from -0.16 to -0.26. The

effects of domain and language origin labels on the data are statistically significant (Cohen's d > 0.8), with language origin effects showing particularly large effect sizes in all models. These challenges with regional terms also reflected concerns from Schryver (2023) about LLMs underrepresentation of minorities in dictionaries, as these models are typically trained on 'general English web content' that may not capture linguistic diversity.



(a) Performance by Category

Model	Domain Effect	Language Effect	Cohen's *d*
GPT-4o	+0.056	-0.260	-1.097
Llama 3.2	+0.059	-0.155	-0.817
Gemini 2.0 Flash	+0.077	-0.212	-0.928

Sample Sizes:
Baseline: n=727
Domain: n=167
Language: n=306

(b) Effects Summary

Figure 3.2: Domain and Language Origin Effects on Semantic Similarity. Domain-specific words show consistent improvements (+0.056 to +0.077) while words with specific language origins show significant decreases (-0.155 to -0.260) with large effect sizes.

3.1.4 Frequency Analysis from idTenTen24

Corpus-based frequency statistics offer another result for understanding the performance of the models and their challenges, particularly for rare versus common vocabulary items. The frequency analysis (Figure 3.3) of the idTenTen2024 corpus reveals a non-linear relationship between word frequency and semantic similarity scores. Across all models, words with zero frequency in the corpus achieved the lowest semantic scores. This indicates that rare words are a significant challenge to generate definitions for.

It is also important to mention that 98 words checked for their frequency have

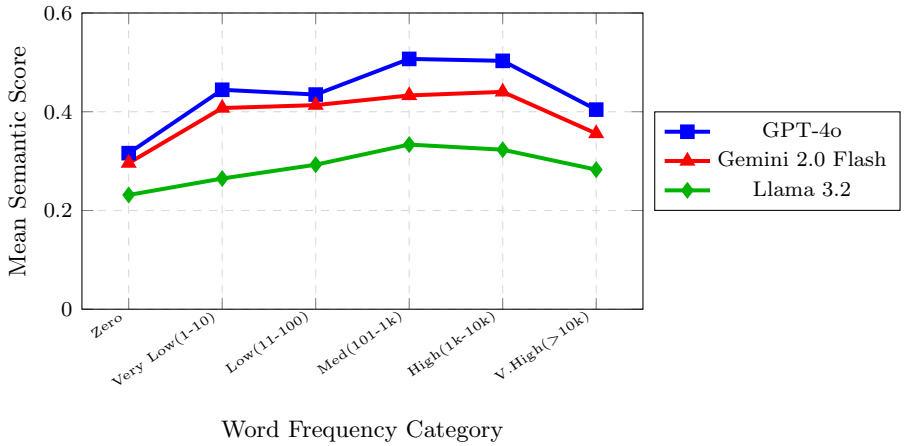


Figure 3.3: Impact of Word Frequency on Semantic Scores Across AI Models, showing non-linear relationship between word frequency and semantic similarity scores. The models perform worst at 0 frequency, best at around medium to high frequency, and decreases again at very high frequency

0 occurrences in the corpus. Furthermore, 62 (63.27%) of the zero-frequency words are identified as regional or domain-specific words. This suggests that idTenTen24 still underrepresents many of the language variants in KBBI. This may also impacted the models since they rely on pre-training data or possibly fine-tuning data during processing instead of retrieving information from the Internet. For that reason, the underrepresentation of such words is what caused the semantic similarity scores to be reduced. Table 3.3 shows five entries that LLM has successfully generated definitions that are close to the KBBI. The high semantic scores for these zero-frequency words are likely due to LLM's capability to understand these terms through their knowledge of other languages. Words like *astrofilia*, *astrofili*, and *eksokanibalisme* closely resemble their English counterparts (*astrophilia*, *astrophile*, and *exocannibalism*), suggesting that the models may infer meanings from morphological or phonetic similarities despite their absence from the idTenTen24 corpus.

In contrast, the five lowest-scoring entries all originate from regional Indonesian languages. This suggests that LLMs struggle to comprehend these words from local language varieties, likely due to limited context and examples. As mentioned earlier, entries originated from regional language are the majority in the zero-frequency word category.

Performance improves as frequency increases from zero to medium ranges, with the highest score occurring at medium frequency (101-1000 occurrences). However, per-

3.1. IndoSBERT Semantic Similarity Analysis

Table 3.3: Top 5 Semantic Scores for Zero Frequency Words

Rank	Word	GPT-4o	Llama 3.2	Gemini 2.0 Flash	Avg Score	Frequency
1	astrofilia	0.825	0.762	0.825	0.804	0
2	eksokanibalisme	0.793	0.526	0.801	0.706	0
3	astrofili	0.737	0.590	0.689	0.672	0
4	morfoleksikon	0.715	0.597	0.607	0.640	0
5	berderup-derup	0.820	0.251	0.792	0.621	0

Table 3.4: Bottom 5 Semantic Scores for Zero Frequency Words

Rank	Word	Origin	GPT-4o	Llama 3.2	Gemini 2.0 Flash	Avg	Freq
1	tapahosut	Minahasa Tonsawang*	-0.02	0.13	0.15	0.08	0
2	hoiltom	Kei	0.09	0.02	0.15	0.09	0
3	asahehe	Luhu	0.07	0.08	0.12	0.09	0
4	tentari	Sangir	0.14	0.12	0.03	0.10	0
5	campoang	Javanese	0.06	0.09	0.15	0.10	0

* This label was unavailable in the October 2023 dataset; this information was added in the official 2024 revision.

formance notably declines in the very high frequency category (>10,000 occurences). This suggests that extremely common words may present different challenges, most likely polysemy and new word senses.

Table 3.5: Semantic Scores and Frequency Data

Rank	Word	GPT-4o	Llama 3.2	Gemini 2.0 Flash	Avg Score	Frequency
1	yang	0.306	0.249	0.394	0.316	206M
2	pasar	0.136	0.083	0.177	0.132	3.3M
3	buku	0.194	0.199	0.218	0.204	2.7M
4	bali	0.004	0.060	0.108	0.057	1.8M
5	usai	0.009	0.079	-0.003	0.028	1.0M

Table 3.5 shows how low the scores are for the words with the highest frequency in the dataset. Most of them achieved scores below 0.2, although their data should be more than enough for the models to understand. The lowest semantic similarity score in this table is *usai*, where while the word commonly functions as a verb meaning "something has ended or finished" (almost equivalent to the English word *over*), the KBBI entry refers to a specific noun from the Alune language, which is *kulit buah pala*

('nutmeg skin'). All three models defaulted to the common interpretation, GPT as *selesai; berakhir* ('finished; ended'), Llama 3.2 as *waktu yang sudah berlalu; kesempatan untuk melakukan sesuatu yang tidak akan ada lagi* ('time that has passed; opportunity that will no longer exist'), and Gemini as *waktu berakhir; selesai* ('time ends; finished'), resulted in low scores.

The frequency analysis reveals that LLMs need more context and data for two distinct categories of words. High-frequency words with low semantic scores due to new sense additions to those words and zero-frequency words originating from regional Indonesian languages, which may fall outside the knowledge of the LLM.

3.2 Human Evaluation

The automated semantic similarity metrics have provided a scalable evaluation across the large dataset, but they did not explore whether these definitions would be acceptable or not in KBBI. For example, the word *hipokinetik* ('hypokinetic') from the medicinal domain scored for GPT (0.73), but GPT defined it as 'related to or caused by minimal movement or physical activity' which clearly violated the principle of using the same part of speech of the word being defined, in this case, noun. Thus, human evaluation is still essential to see how these definitions are perceived by Indonesian speakers. The human evaluation found two important things that automated metrics cannot see, those are adherence to KBBI's principles from the guideline and user preferences. Nine evaluators assessed both GPT-generated and KBBI definitions for 16 words from different word categories. The evaluators consisted of two linguists who have contributed to KBBI, two professionals who frequently use KBBI as a reference in their work, and five general Indonesian speakers.

3.2.1 Definition Ratings and Adherence to Principles

In the first part of the human evaluation results, where we evaluated eight words, GPT received higher ratings in every word category, except for nouns. For the noun *krim dingin* ('cold cream'), GPT defined it as 'cream that gives cold sensations', while KBBI defined it as 'cold cream to chill the skin'. Both definitions were evaluated as violating Principle 5 (not specific enough), indicating that the evaluators found both definitions insufficiently precise. However, the GPT definition was also evaluated as violating Principle 1 (not self-explanatory). An evaluator commented that this definition raises more question, "does the cream itself feel cold? does it produce a

3.2. Human Evaluation

Table 3.6: Evaluation Results: GPT vs KBBI Word Definitions (Scale: -2=Very Bad, -1=Bad, 0=Neutral, 1=Good, 2=Very Good)

Category	Word	GPT Mean	KBBI Mean	GPT Scores	KBBI Scores
Noun 1	krim dingin	0.56	**0.78**	[-1,0,1,1,1,-1,1,1,2]	[2,0,0,1,1,2,0,-1,2]
Noun 2	colet	0.00	**0.89**	[-1,0,0,-1,1,-1,1,-1,2]	[1,-1,2,2,1,1,0,0,2]
Verb 1	membujetkan	**1.67**	0.67	[2,2,1,2,1,2,2,1,2]	[1,2,1,0,-1,1,1,-1,2]
Adjective 1	jenong	**1.11**	0.67	[2,1,1,2,0,1,1,1,1]	[1,1,-1,2,2,0,1,-2,2]
Domain 1	emosi palsu	**0.78**	0.44	[1,0,-1,2,2,-1,2,0,2]	[1,-1,-1,1,0,1,1,0,2]
Domain 2	jurnal kas	**0.78**	0.11	[2,-1,-1,2,1,2,1,-1,2]	[2,1,-1,-1,-1,0,0,0,1]
Lang-spec 1	martarombo	**1.11**	-0.22	[2,1,1,-1,1,2,2,0,2]	[2,-1,-1,-1,-2,0,0,-1,2]
Lang-spec 2	somen	**1.22**	0.78	[2,1,0,1,2,-1,2,2,2]	[1,-1,0,1,1,1,1,1,2]
Overall		**0.903**	**0.514**		

cold sensation? what is it for?" Another evaluator commented that "this is not what it means" and rated the GPT definition lower. This may suggests that the KBBI definition was perceived as more accurate in explaining what *krim dingin* refers to among the evaluators. For the second noun, *colet*, where KBBI received the highest score, defined it as 'a technique to paint on cloths', whereas GPT defined it as 'to scribble or mark with a small stroke; spraying colors or paint on a surface'. The evaluators rated the GPT definition relatively low because it clearly violated Principle 3 (the first word does not match part of speech). KBBI appropriately started with a noun, consistent with the word being defined, while GPT inaccurately started with a verb. Additionally, since *colet* describes an action, defining it as a technique was perceived by the evaluators as more accurate.

Table 3.7: Definition Quality Principle Violations: GPT vs KBBI

Code	Principle Description	Count		Percentage	
		GPT	**KBBI**	**GPT**	**KBBI**
0	No violation (definition is fine)	**43**	37	**59.7%**	51.4%
1	Not self-explanatory	8	9	11.1%	12.5%
2	Uses words more complicated than definiendum	1	**7**	1.4%	**9.7%**
3	First word doesn't match part of speech	**12**	8	**16.7%**	11.1%
4	Circular definition	9	**12**	12.5%	**16.7%**
5	Not specific enough	5	**13**	6.9%	**18.1%**
	Total evaluations	**72**	**72**		

Regarding principle violations, although it was rated higher than KBBI in every other word category, GPT violated Principle 3 (part-of-speech matching) more frequently (16.7%) than KBBI (11.1%). GPT may have failed to follow the instruction in the prompt or made errors when analyzing the word's part-of-speech. On the other hand, KBBI violated the other principles somewhat more frequently than GPT.

KBBI's entries were frequently considered not specific enough, as reflected in the previously mentioned *krim dingin* example and the adjective *jenong*. GPT defined the latter as 'forehead which is wide and frontal; forehead that is wider than normal', which violated Principle 3, but KBBI defined it as 'striking and frontal (about a forehead and others)', which potentially includes other things being frontal besides foreheads. While KBBI did mention foreheads, their explanation still suggested that the word *jenong* could be used in contexts other than to describe foreheads, which was considered insufficiently specific and too broad by the evaluators.

3.2.2 Users' Definition Preferences

For the third part of the human evaluation, the GPT achieved greater overall approval, winning 8 out of 10 words. This aligns with Lew (2023) that ChatGPT-generated definitions could be 'practically indistinguishable in quality from those written by highly trained human lexicographers,' which corresponds with our human evaluation results showing strong user preference for LLM definitions. The only wins KBBI achieved were in the adjective category, most notably in the word *parenteral* with 100% approval versus 22.2% for GPT. KBBI defined *parenteral* as 'not through the digestive system regarding medicine or nutrition intake, usually through the blood vessel' while GPT defined it as 'not through the digestive system, usually through injection or infusion'. The evaluators may have preferred KBBI because it explained what is being inserted into the body (medicine or nutrition), whereas GPT being vague about the infusion of something unspecified into the body not through digestion.

Table 3.8: Survey Results: GPT vs KBBI Definition Preferences by Individual Word Examples

Category	Headword	Vote Counts			Total	Approval (%)	
		GPT	KBBI	Both	Resp.	GPT	KBBI
Noun 1	*bromansa*	4	3	2	9	**66.7**	55.6
Noun 2	*afkiran*	8	1	0	9	**88.9**	11.1
Verb 1	*mengeataskan*	5	3	1	9	**66.7**	44.4
Verb 2	*mengimpersonasi*	6	1	2	9	**88.9**	33.3
Adjective 1	*parenteral*	0	7	2	9	22.2	**100.0**
Adjective 2	*acakadul*	1	2	6	9	77.8	**88.9**
Domain 1	*apelium*	5	2	2	9	**77.8**	44.4
Domain 2	*ahli onkolog*	6	0	3	9	**100.0**	33.3
Language-spec 1	*baluse*	7	1	1	9	**88.9**	22.2
Language-spec 2	*baby boomer*	4	3	2	9	**66.7**	55.6

It is interesting to note that the preference of the evaluators contradicts the rating

from the previous analysis. While the quality ratings showed KBBI performing better for nouns like *krim dingin* and *colet*, the preference survey shows GPT winning for the adjectives like *bromansa* and *afkiran*. This may suggest that evaluators' preferences may not be restricted by the technical principles but could extend beyond that, including factors such as clarity, comprehensibility, or familiarity with the words.

Additionally, GPT's strength with domain-specific words is consistent with the semantic score analysis in Section 3.1. The result shows *ahli onkolog* achieving perfect GPT approval and zero KBBI preferences, and *apelium* achieving 77.8% GPT approval. This further suggests that GPT excels at explaining a specialized or technical vocabulary. GPT's better performance with technical terms is probably caused by its training on diverse specialized data, making it able to understand and generate domain-specific knowledge into comprehensible explanations.

Chapter 4

Conclusions

4.1 Summary

RQ1: *To what extent can large language models (GPT-4o, Gemini 2.0 Flash, and Llama 3.2) generate semantically accurate definitions for new Indonesian words, as measured by IndoSBERT, and how do they compare with human-written KBBI definitions in terms of overall quality and user preference?*

LLMs show moderately good capability to generate semantically accurate definitions using KBBI's style guideline as references. GPT-4o achieved the highest semantic accuracy (mean IndoSBERT score: 0.52), followed by Gemini 2.0 Flash (0.46) and Llama 3.2 (0.35). In particular, 46% of the GPT-4o definitions achieved semantic similarity scores in the range of 0.6-1.0, indicating substantial semantic similarity with the KBBI definitions.

Human evaluation showed a strong preference for LLM-generated definitions over KBBI. GPT-4o in particular received significantly higher quality ratings (0.9 vs 0.51 for KBBI) and won user preference in 8 of 10 direct comparisons. However, this preference was not universal. KBBI definitions were rated higher for certain nouns and preferred for certain adjectives where specificity and clarity were critical. These results contribute to previous work, supporting the assertion Lew (2023) about the potential of LLM in lexicography, while also confirming the observation by Schryver (2023) that effectiveness varies significantly between different types of lexical content. Our findings extend this research to Indonesian, showing that the capabilities and limitations identified in the English language studies are also found in other major languages.

RQ2: *How do LLM definition generation results vary across word categories (word class, definition type, domain-specific words, and language origin) and what patterns can we see from corpus frequencies?*

LLM performance varies across linguistic categories, revealing both strengths and limitations. On average, the effect of the word class is quite consistent across all models, adverbs performed the best (0.56), followed by verbs (0.45), nouns (0.45), and adjectives (0.42). The definition type significantly affected LLMs' performance, especially for abbreviation expansion definitions where they consistently underperformed (particulary with Llama 3.2, which is 0.23). This is likely due to the absence of this definition type in the guideline and in the prompts.

The effects of domain and language origin were statistically significant. Domain-specific words consistently achieved higher scores (+0.056 to +0.077 improvement), demonstrating LLMs' strength with technical vocabulary. In contrast, words from specific language origins showed substantial decreases (-0.155 to -0.260) with large effect sizes (Cohen's $d > 0.8$), indicating a critical limitation for Indonesia's multilingual lexical landscape.

Frequency analysis revealed a non-linear relationship between corpus frequency and definition quality. Zero-frequency words performed worst, with 98 such words identified, of which 63.27% were regional or domain-specific terms underrepresented in idTenTen24. Performance improved through medium frequencies (101-1000 occurrences) but declined for very high-frequency words (>10,000), showing challenges with polysemy and new sense detection.

RQ3: *To what extent do LLM-generated definitions adhere to the KBBI's guidelines (five core principles and appropriate definition types) when guided by Chain-of-Thought prompting, and which principles are frequently violated?*

LLM-generated definitions show mixed adherence to KBBI's five core principles, with both strengths and systematic weaknesses. However, the KBBI definitions themselves also violated many principles from the guideline. Chain-of-Thought prompting demonstrated moderate success, with 59.7% of the GPT definitions containing no principle violations compared to 51.4% for the KBBI definition. LLMs commit specific violations of the principles, they violated Principle 3 (part-of-speech matching) more frequently than KBBI (16.7% vs 11.1%), suggesting difficulties in maintaining consistent grammatical categorization despite explicit instructions. This indicates that prompt engineering alone may be insufficient to ensure linguistic consistency.

The effectiveness of Chain-of-Thought prompting was context-dependent. It suc-

cessfully guided definition type selection and principle consideration to some extent, but failed to address abbreviation expansion definitions (absent from training examples and the KBBI guideline itself) and struggled with specific regional language terms that is likely outside the models' knowledge base.

4.2 Limitation

The availability of Indonesian data fine-tuned LLMs is still severely limited. Although Indonesian-specific models like IndoGPT exist, they are based on outdated models (GPT-2) that significantly underperform than the latest models. For evaluation metrics, even though semantic similarity scores are scalable, it cannot capture crucial aspects of definition quality beyond semantic similarity. IndoSBERT similarity scores cannot evaluate stylistic appropriateness, adherence to KBBI's specific editorial conventions, or cultural sensitivity, which are critical factors for practical lexicographic applications. For stylistic consistency with KBBI, additional evaluation methods are still required. ROUGE (Lin, 2004) were initially considered, but preliminary analysis showed high zero-score rates (33-53%) due to vocabulary differences rather than semantic inconsistency, so we excluded ROUGE from the final evaluation. Our human evaluation, though more comprehensive, was limited to 16 words with 9 evaluators, which may not provide sufficient statistical power for strong conclusions across all word categories and definition types.

The focus of this study, which uses new KBBI entries (31 October 2023 update) as a dataset, provides only a partial view of LLM capabilities for Indonesian lexicography. New entries often represent specialized, technical, or borrowed terms that may not reflect the challenges of defining the core of Indonesian vocabulary itself. The morphologically complex nature of Indonesian words, including productive affixation, reduplication, and compound formation, is still largely unexplored, potentially undermining the difficulties LLMs may face with the full scale of all Indonesian lexical items.

4.3 Future Works

Future work should focus on developing Indonesian-specific models using the latest architectures. Those approaches should include domain-specific fine-tuning and the integration of Indonesian linguistic knowledge such as morphological analyzers and syntactic information. Substantial Indonesian datasets are needed for comprehensively exam-

ining LLMs' potential in Indonesian language and lexicography, particularly through the acquisition of more data samples from Indonesian regional languages. The development of comprehensive evaluation methods and definition generation benchmarks for Indonesian dictionaries would also greatly benefit the field.

Additionally, future research should explore practical integration pathways with KBBI's collaborative editorial workflow so that they can balance AI assistance with human users and experts. The existing KBBI workflow is already transparent and well documented, so it could benefit by incorporating new technologies. This work showed both potentials and critical limitations of current LLMs for Indonesian lexicography, so while these models may support dictionary compilation process, they still need careful adaptation and oversight to adhere to the linguistic and cultural complexities of Indonesia's diverse language landscape.

Appendix

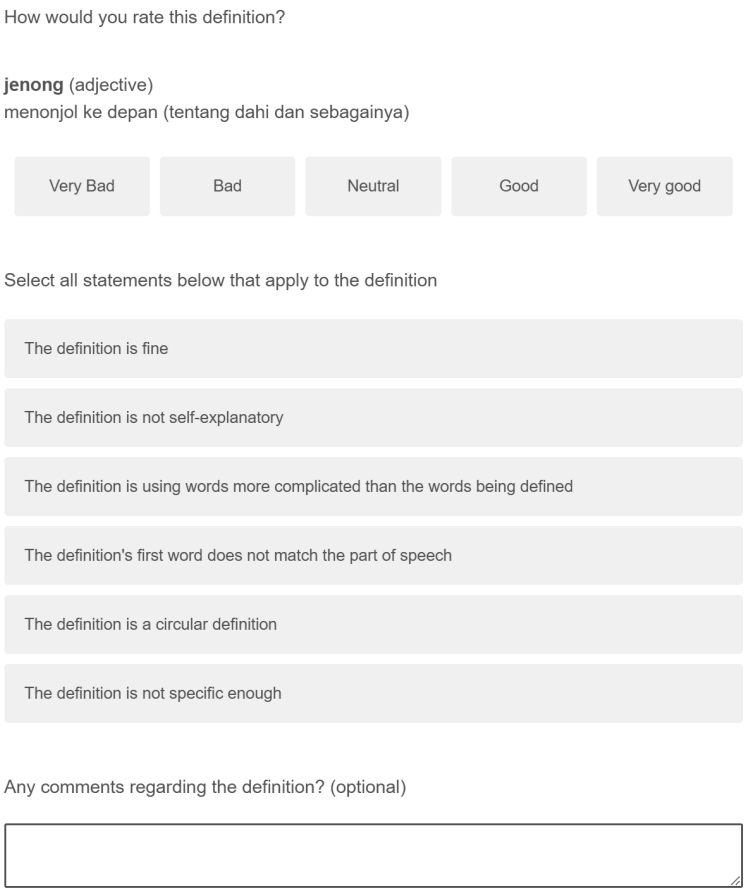


Figure 1: Survey Question Sample First and Second Task, rating the definitions and definition's adherence to principles

.0. Appendix

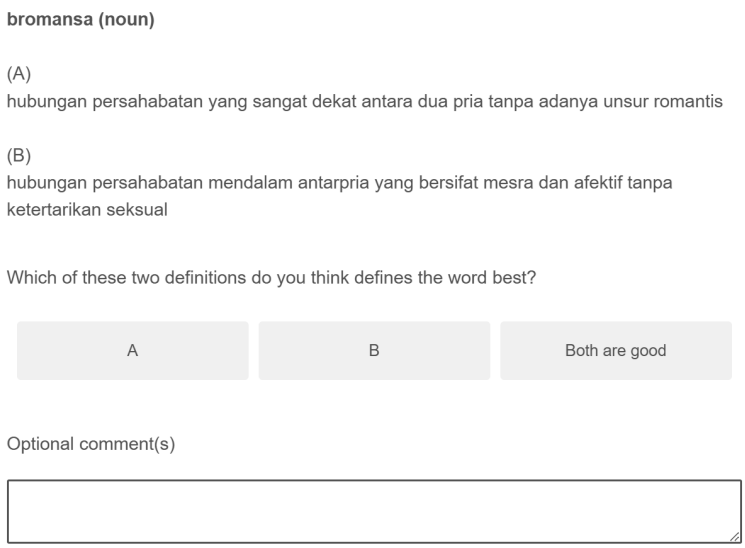


Figure 2: Survey Question Sample Third Task, user's preference on definitions

References

Arvan, Mohammad, Luís Pina, and Natalie Parde (Dec. 2022). "Reproducibility in Computational Linguistics: Is Source Code Enough?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2350–2361. DOI: 10.18653/v1/2022.emnlp-main.150. URL: https://aclanthology.org/2022.emnlp-main.150/.

Atkins, Beryl T. and Michael. Rundell (2008). *The Oxford guide to practical lexicography*. eng. Oxford [etc: Oxford University Press. ISBN: 9780199277704.

Aziz, E. Aminudin (Oct. 2023). *Kata Pengantar dan Prakata KBBI Edisi Keenam*. KBBI Daring. Accessed: 2025-05-26. URL: https://kbbi.kemdikbud.go.id/Beranda/Prakata.

Badan Pengembangan dan Pembinaan Bahasa (n.d.). *Peta Bahasa*. https://petabahasa.kemdikbud.go.id/databahasa.php. Accessed 15-06-2025.

— (n.d.). *Tentang Kami - KBBI Daring*. https://kbbi.kemdikbud.go.id/Beranda/TentangKami. Accessed: 2025-04-14.

Bahasa, Badan (2017). *Kamus Besar Bahasa Indonesia*. Accessed: 2025-03-28. URL: %5Curl%7Bhttps://badanbahasa.kemendikdasmen.go.id/artikel-detail/97/sejarah-kamus-besar-bahasa-indonesia%7D.

Bolinger, Dwight (1965). "The Atomization of Meaning." In: *Language* 41.4, pp. 555–573. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/411524 (visited on 06/10/2025).

Cahyawijaya, Samuel, Holy Lovenia, Alham Fikri Aji, et al. (July 2023). "NusaCrowd: Open Source Initiative for Indonesian NLP Resources." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 13745–13818. DOI: 10.18653/v1/2023.findings-acl.868. URL: https://aclanthology.org/2023.findings-acl.868/.

Cahyawijaya, Samuel, Holy Lovenia, Fajri Koto, et al. (Nov. 2023). "NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages." In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Jong C. Park et al. Nusa Dua, Bali: Association for Computational Lin-

guistics, pp. 921–945. DOI: 10.18653/v1/2023.ijcnlp-main.60. URL: https://aclanthology.org/2023.ijcnlp-main.60/.

Cai, Bill et al. (June 2024). "Low-Cost Generation and Evaluation of Dictionary Example Sentences." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 3538–3549. DOI: 10.18653/v1/2024.naacl-long.194. URL: https://aclanthology.org/2024.naacl-long.194/.

Conklin, Harold (1975). "L. Zgusta Manual of lexicography. (Janua Linguarum, Series Maior 39) Prague: Academia; The Hague: Mouton." In: *Language in Society* 4.2, pp. 241–243. DOI: 10.1017/S0047404500004711.

Cruse, D. Alan. (1986). *Lexical semantics*. eng. Cambridge textbooks in linguistics. Cambridge [etc: Cambridge University Press. ISBN: 052125678X.

Diana, Denaya (2023). "IndoSBERT: Indonesian SBERT for Semantic Textual Similarity tasks." In: URL: https://huggingface.co/denaya/indoSBERT-large.

Durkin, Philip, ed. (2016). *The Oxford Handbook of Lexicography*. 1st. Oxford Handbooks in Linguistics. Oxford; New York: Oxford University Press, pp. xxiii, 698. ISBN: 9780199691630. DOI: 10.1093/ijl/ecw043. URL: https://doi.org/10.1093/ijl/ecw043.

Endarmoko, Eko and Anton M. Moeliono (May 2008). "W. J. S. Poerwadarminta: Bapak Kamus Indonesia." In: *Tempo*. Accessed: 07-06-2025.

Fagbohun, Oluwole, Rachel M. Harrison, and Anton Dereventsov (2024). "An Empirical Categorization of Prompting Techniques for Large Language Models: A Practitioner's Guide." In: *Journal of Artificial Intelligence Machine Learning & Data Science* 1.4, pp. 142–152. DOI: 10.51219/JAIMLD/Oluwole-Fagbohun/18.

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Grehenson, Gusti (May 2022). *Penutur Bahasa Indonesia Capai 300 Juta Jiwa*. UGM Berita. Accessed: 2025-05-29. URL: https://ugm.ac.id/id/berita/22527-penutur-bahasa-indonesia-capai-300-juta-jiwa/.

Holy, Adib (2024). *Bagaimana Editor KBBI Daring Bekerja?* Ed. by Rifka Az-zahra. Accessed: 2025-04-11. URL: %5Curl%7Bhttps://narabahasa.id/artikel/linguistik-terapan/perkamusan/bagaimana-editor-kbbi-daring-bekerja/%7D.

Jakubiček, Miloš and Kilgarriff, Adam and Kovář, Vojtěch and Baisa, Vít and Suchomel, Vít (2017). *Indonesian web corpus (idTenTen)*. Sketch Engine. Accessed: 2025-04-06. URL: https://www.sketchengine.eu/indonesian-corpus/.

Johnson, Samuel (2021). "Preface to A Dictionary of the English Language (1755)." In: *Samuel Johnson*. New Haven: Yale University Press, pp. 397–417. ISBN: 9780300258004. DOI: doi:10.12987/9780300258004-038. URL: https://doi.org/10.12987/9780300258004-038.

Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kamajaya, Ian, David Moeljadi, and Dora Amalia (Sept. 2017). "KBBI Daring: A Revolution in the Indonesian Lexicography." In: *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, pp. 513–530. URL: https://davidmoeljadi.github.io/papers/ASIALEX2017_davidmoeljadi.pdf.

Kaye, Alan S. (1999). "R.R.K. Hartmann and Gregory James. Dictionary of Lexicography. London: Routledge. 1998. Pp. xv + 176. 105.00(*hardcover*).." In: *Canadian Journal of Linguistics/Revue canadienne de linguistique* 44.3, pp. 299–300. DOI: 10.1017/S0008413100017369.

Koto, Fajri et al. (Dec. 2020). "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 757–770. DOI: 10.18653/v1/2020.coling-main.66. URL: https://aclanthology.org/2020.coling-main.66/.

Lew, Robert (June 2023). *ChatGPT as a COBUILD lexicographer*. DOI: 10.1057/s41599-023-02119-6. URL: osf.io/t9mbu_v1.

Lin, Chin-Yew (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text summarization branches out: Proceedings of the ACL-04 workshop*, pp. 74–81.

Liu, Pengfei et al. (2021). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." In: *arXiv preprint arXiv:2107.13586*. Submitted July 28, 2021. URL: https://arxiv.org/abs/2107.13586.

Marcus, Gary and Ernest Davis (Aug. 2020). "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about." In: *MIT Technology Review*. URL: https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

McKean, Erin and Will Fitzgerald (2023). "The ROI of AI in Lexicography." In: *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*. Seoul: Yonsei University, pp. 10–20.

Moeljadi, David, Ian Kamajaya, and Dora Amalia (June 2017). "Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications." In: *Proceedings of the 11th International Conference of the Asian Association for Lexicography (ASIALEX 2017)*. Guangzhou, China, pp. 64–80. URL: https://davidmoeljadi.github.io/papers/ASIALEX2017_davidmoeljadi.pdf.

Nurjaman, Wisnu (May 2024). "Peran Kamus Besar Bahasa Indonesia (KBBI) dalam Peningkatan Kualitas Berbahasa dalam Pendidikan." In: *Semantik : Jurnal Riset Ilmu Pendidikan, Bahasa dan Budaya* 2.2, pp. 230–237. DOI: 10.61132/semantik.v2i2.643. URL: https://doi.org/10.61132/semantik.v2i2.643.

Office of Assistant to Deputy Cabinet Secretary for State Documents Translation (Nov. 2023). *Bahasa Indonesia Named UNESCO General Conference Official Language*. Sekretariat Kabinet Republik Indonesia. Accessed: 2025-05-29. URL: https://setkab.go.id/en/bahasa-indonesia-named-unesco-general-conference-official-language/.

References

OpenAI et al. (2024). *GPT-4o System Card*. arXiv: 2410.21276 [cs.CL]. URL: https://arxiv.org/abs/2410.21276.

Pelindungan Bahasa dan Sastra, Pusat Pengembangan dan (2019). *Petunjuk Teknis Penggunaan KBBI Daring*. Pusat Pengembangan dan Pelindungan Bahasa dan Sastra.

Pelindungan Bahasa dan Sastra Badan Pengembangan Bahasa dan Perbukuan Kementerian Pendidikan dan Kebudayaan, Pusat Pengembangan dan (2019). *Petunjuk Teknis Penyusunan Kamus Ekabahasa*. Pusat Pengembangan dan Pelindungan Bahasa dan Sastra Badan Pengembangan Bahasa dan Perbukuan Kementerian Pendidikan dan Kebudayaan.

Peristilahan, KKLP Perkamusan Dan (2022). *Sejarah Kamus Besar Bahasa Indonesia*. Accessed: 2025-04-11. URL: %5Curl%7Bhttps://badanbahasa.kemendikdasmen.go.id/artikel-detail/97/sejarah-kamus-besar-bahasa-indonesia%7D.

Pham, Bach et al. (2025). *Word Definitions from Large Language Models*. arXiv: 2311.06362 [cs.CL]. URL: https://arxiv.org/abs/2311.06362.

Rees, Geraint Paul and Robert Lew (Dec. 2023). "The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task." In: *International Journal of Lexicography* 37.1, pp. 50–74. ISSN: 0950-3846. DOI: 10.1093/ijl/ecad030. eprint: https://academic.oup.com/ijl/article-pdf/37/1/50/56778107/ecad030.pdf. URL: https://doi.org/10.1093/ijl/ecad030.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: https://doi.org/10.48550/arXiv.1908.10084.

Rundell, Michael (2023). "Automating the Creation of Dictionaries: Are We Nearly There?" In: *Proceedings of the 16th International Conference of the Asian Association for Lexicography (AsiaLex 2023)*. Seoul, Korea, pp. 9–17.

Schryver, Gilles-Maurice de (Oct. 2023). "Generative AI and Lexicography: The Current State of the Art Using ChatGPT." In: *International Journal of Lexicography* 36.4, pp. 355–387. ISSN: 0950-3846. DOI: 10.1093/ijl/ecad021. eprint: https://academic.oup.com/ijl/article-pdf/36/4/355/57174270/ecad021.pdf. URL: https://doi.org/10.1093/ijl/ecad021.

Schuff, Hendrik et al. (2023). "How to do human evaluation: A brief introduction to user studies in NLP." In: *Natural Language Engineering* 29.5, pp. 1199–1222. DOI: 10.1017/S1351324922000535.

Sunendar, Dadan (2018). *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan.

Team, Gemini et al. (2025). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

Tim Penyusun Kamus, Pusat Pembinaan dan Pengembangan Bahasa (1988). *Kamus Besar Bahasa Indonesia*. 1st ed. Jakarta: Departemen Pendidikan dan Kebudayaan Republik Indonesia / Balai Pustaka, pp. xix, 1090. ISBN: 9794620653.

Tran, Hanh Thi Hong et al. (2023). "Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT." In: *Proceedings of the eLex 2023 Conference:*

Electronic Lexicography in the 21st Century. Ed. by Marek Medveď et al. Brno: Lexical Computing, pp. 19–38. URL: https://www.youtube.com/watch?v=rQC3Rz04b20.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Wei, Jason et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL]. URL: https://doi.org/10.48550/arXiv.2201.11903.

Wilie, Bryan et al. (Dec. 2020). "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Ed. by Kam-Fai Wong, Kevin Knight, and Hua Wu. Suzhou, China: Association for Computational Linguistics, pp. 843–857. DOI: 10.18653/v1/2020.aacl-main.85. URL: https://aclanthology.org/2020.aacl-main.85/.