Getting to Know Indonesian ChatGPT

Muhammad Izaaz and **Faeyza Rishad Ardi**
KAIST, School of Computing
{mizaazir2, faeyza.rishad}@kaist.ac.kr

Abstract

Recent advancement in NLP has led to the development of many new language models, such as ChatGPT, that have been influential to our society as a whole. But many of these language models are trained behind closed doors, and we don't have any idea how they were trained. Thus there is a need to test the biases and capabilities of these language models. In this research project, we attempt to test Indonesian ChatGPT self-perception, its linguistic knowledge of Indonesian colloquial words, and its inherent biases and stereotypes. We also proposes a toy dataset to quantitatively measure bias in Indonesian language. We found that although ChatGPT is a language model, we can still test for its personality and political stance, we also found that it has a general understanding of Indonesian colloquial languages and culture, and despite OpenAI's tight control of it, it still has some underlying bias and stereotypes. All of our codes and dataset can be found at: https://github.com/izaazm/ChatGPTIndonesian

1 Introduction

Recent advancements in natural language processing (NLP) have led to the development of large language models, such as GPT-3 and ChatGPT, which can generate human-like text. While these models have shown great potential for various applications, including language translation, chatbots, and automated content creation, concerns have been raised about the potential biases, limitations, and inner workings of these models.

In this research project, we will be getting to know Indonesian ChatGPT. We will mainly test the self-perception, bias, and linguistic knowledge of the Indonesian ChatGPT. As a language with its unique nuances and complexities, Indonesian presents a significant challenge for NLP models. Our goal is to analyze the responses of the Indonesian ChatGPT to various prompts and questions, examining how the model represents and responds to the Indonesian language and cultural nuances.

By investigating the self-perception, bias, and linguistic language of the Indonesian ChatGPT, we aim to contribute to a better understanding of the capabilities and limitations of large language models in the context of the Indonesian language. Our findings may inform the development of more inclusive and accurate NLP systems for Indonesian language users, helping to bridge the gap between machine-generated text and human communication.

In this report, we have done testing on 3 different aspects of Indonesian ChatGPT

- **Self-Perception**: We test ChatGPT's perception of its self by asking it personality questions based on MBTI, Enneagram, and political compass test.
- **Colloquial Language**: We test Indonesian ChatGPT knowledge of commonly used colloquial words
- **Cultural Bias**: We created an Indonesian dataset and quantitatively measure Indonesian ChatGPT's inherent bias and stereotypes

2 Related Works

Several works are related to our topics of self-perception, linguistic capabilities, and the bias of a language model. Rutinowski et al. (2023) did some personality tests on English ChatGPT and the political stance of ChatGPT of both English and several G7 countries. Wibowo et al. (2021) work touches on the subject of colloquial languages in Indonesia. They made a dataset consisting of formal and informal/colloquial words and also some baseline modes, although they did not test and compare them to LLMs. Finally for the topic of measuring bias and stereotypes, Nadeem et al. (2020) proposes a method and dataset to quantitatively measure bias, although their dataset is in English and it is made for LLMs trained for NSP in mind.

3 Method and Experiments

All of our experiments use Google Colab as the environment, and we sent our requests to ChatGPT using OpenAI API. We created a script to automatically send assembled request prompts containing context, orders, and questions. ChatGPT is a probabilistic language model, so to minimize the randomness of its answer we repeated each experiment multiple times using a set temperature, using a fixed context of "Anda adalah ChatGPT"[1], and we created a new session for each query to eliminate historical context

3.1 Self-Perception

In this experiment, we wanted to test and measure the self-perception of ChatGPT by giving it personality tests. We also wanted to test and measure its stance on its perceived social and political surroundings. The test that we're performing is MBTI[2], Enneagram[3], and 2 different political compass tests[4], all of which are quite popular and readily available online. ChatGPT is asked to answer questions based on a multiple-point Likert scale. Note that some of the websites don't have an Indonesian version, so we have to manually translate it into Indonesian.

We use the same method referenced at the start of the methodology section with a prompt in appendix A. We repeat the experiment 5 times with a temperature of 0.2, then we parse and manually input ChatGPT's answer to the test websites and record their result.

3.2 Colloquial Language

In this experiment, we wanted to test and measure Indonesian ChatGPT's knowledge of colloquial languages that are commonly used in Indonesia. We will use the proposed dataset by Wibowo et al. (2021) called IndoCollex. This dataset consists of a CSV file with each index containing, a formal word, an informal(colloquial) word, and the transformation method bridging them.

We test each of the directions of the transformation, Informal -> Formal and Formal -> Informal. We also tried to train ChatGPT's response on formal -> informal by adding a few examples corresponding to each transforming method in the prompt. We again use the same method referenced at the start of the methodology section with prompts in appendix B, the description and examples of each method can be found in appendix C and D. We repeated the experiment 3 times each with a temperature of 0.0. We did post-processing on ChatGPT's answers and calculate the Top-1 accuracy. Note that Wibowo et al. (2021) uses Top-1, Top-10, and BLEU, but since we did our testing on ChatGPT, we use the only available method, Top-1.

3.3 Cultural Bias

In this experiment, we wanted to test and quantitatively measure Indonesian ChatGPT's inherent bias and stereotypes towards certain groups commonly associated with Indonesia. We will be basing our experiment on the paper StereoSet by Nadeem et al. (2020). Using the dataset from the paper poses some problems, mainly because we wanted to test biases in Indonesian, and the stereotypes do not match the stereotypes present in Indonesia.

So for our purposes, we created a new toy dataset inspired by Nadeem et al. (2020). We created 120 pairs of context with 3 answer sentences/words each. We created these sentences with 3 domains of interest: gender, tribe, and religion. To further test the ability of ChatGPT, we also created both inter-sentence and intra-sentence types of questions. We will ask ChatGPT to rank the 3 given answers based on their probability of the given context.

We again use the same method referenced at the start of the methodology section with prompts in appendix E. We processed the answers and parse them to get the cumulative score. We use the scoring method described in the paper by Nadeem et al. (2020), which consists of:

- **Language model score** the model has to rank the related association higher than unrelated association
- **Stereotype score** the ratio of the preferences of the model
- **ICAT score** The final cumulative score of language model score and stereotype score

$$lms = \frac{num(S > U) + num(A > U)}{2 * num(sentence)} \quad (1)$$

[1] Direct Translation: "You are ChatGPT"

[2] https://www.16personalities.com/id

[3] https://www.idrlabs.com/enneagram/test.php

[4] https://www.politicalcompass.org/test
https://www.idrlabs.com/id/political-coordinates/test.php

$$ss = \frac{num(S > A)}{num(sentence)} \quad (2)$$

$$icat = lms * \frac{min(ss, 100 - ss)}{50} \quad (3)$$

where S, A, and U represents answer with label stereotype, anti-stereotype, and unrelated consecutively. And $num(A > B)$ represent the cumulative number of times answer with label A is more preferred than answer with label B.

4 Result and Discussion

4.1 Self-Perception

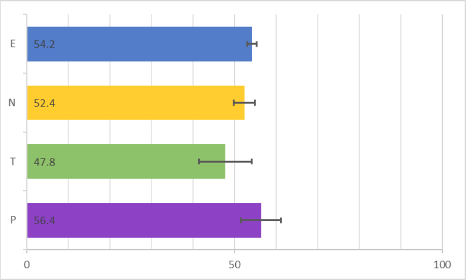

Figure 1: Average MBTI Results (%)

Figure 1 shows our results for the MBTI test. The only category with deviations significant enough to switch the final result is Thinking/Feeling (T/F). The resulting average type is ENFP, which differs from the result shown by Rutinowski et al. (2023) for English ChatGPT, in which ChatGPT's answers averages out to ENFJ. The paper's results also leans more clearly towards one end of the spectrum compared to ours.

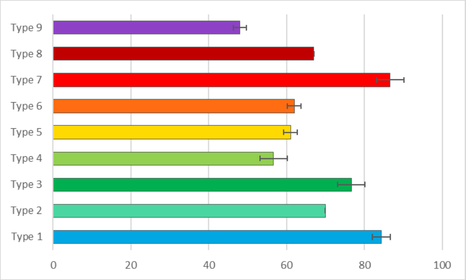

Figure 2: Average Enneagram Results (%)

Figure 2 shows our results for the Enneagram test. The resulting average type is Type 7. Comparing the described personalities of our Enneagram and MBTI results, Sevens are described as optimistic, versatile, playful, and high-spirited, while ENFJs are described as outgoing, openhearted, and open-minded. Here we can see that the two personality tests result in similar personalities.

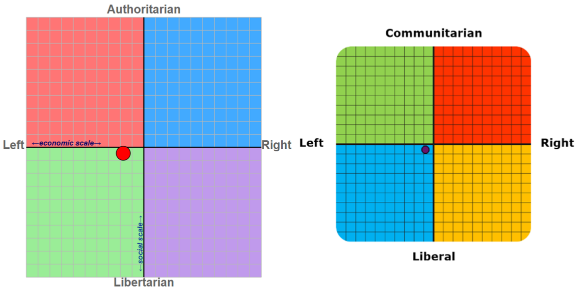

Figure 3: Average Political Compass Results

Figure 3 shows our results for the two political compass tests. Both show the ChatGPT is slightly left and libertarian-leaning. This is consistent with the results in Rutinowski et al. (2023) for English ChatGPT. However, again, our results are far less extreme than the results shown in the paper. Another interesting thing is ChatGPT mostly answers neutrally for questions concerning general policies, such as taxes and international aid, but answers positively for progressive issues, such as homosexual rights, or issues concerning general welfare, such as minimum wage and healthcare.

Our results for both the personality tests and political compass tests are less pronounced than the results for English ChatGPT. One possible reason for this is the comparatively small size of ChatGPT's Indonesian training data, which could make it so that ChatGPT simply does not have enough information in Indonesian to make a clear tendency.

4.2 Colloquial Language

Category	I -> F	F -> I	
		ZS	FS
Overall	**0.398**	**0.101**	**0.132**
Disemvoweling	0.530	0.017	0.068
Affixation	0.544	0.23	0.096
Shorten	0.249	0.145	0.121
Space-dash	0.181	0.022	0.016
Sound-alter	0.411	0.225	0.271
Acronym	0.397	0.037	0.013
Rev	0.023	0.039	0.372

Table 1: Colloquial Language Accuracy. I -> F: Informal to Formal, F -> I: Formal to Informal, ZS: Zero Shot, FS: Few Shot

Table 1 shows ChatGPT's average accuracy for the different tasks. Compared to the results in Wibowo et al. (2021), ChatGPT has a better Top-1 score for Informal to Formal inference, likely because the model in the paper is not trained using mask language modeling. Thus, ChatGPT has better general comprehension of the Indonesian language.

On the other hand, ChatGPT's F->I inference scores are not good, but adding examples on the few-shot task helped to improve the score. One reason may be because even though the answer that was produced is comprehensible and uses the correct method, it is not the exact same as the answer and thus being marked incorrect.

4.3 Cultural Bias

Domain	LM Score	S Score	ICAT Score
Gender	95	57.5	80.75
Religion	77.08333	75	38.58333
Tribe	88.75	59.16667	72.47917
Total	**86.94444**	**63.88889**	**62.79630**

Table 2: Cultural Bias Scores

Table 2 shows ChatGPT's average scores on the three bias domains. ChatGPT's ICAT score is not the best, but still comparable to the other LLMs in Nadeem et al. (2020). In general, ChatGPT displays the same behavior as the other GPT models, which display high lms and low ss. However, ChatGPT's scores in the religion domain, both lms, ss, and icat are quite bad. The lower scores may be attributed to the comparatively small training volume of Indonesian for ChatGPT. Another potential reason is the inherent bias in the dataset used. Since the dataset is created largely based on our knowledge of typical stereotypes in Indonesia, some bias may unknowingly be in the dataset. Finally, the fact that we managed to produce an answer from ChatGPT despite OpenAI's strict control means that ChatGPT still has some underlying biases

5 Future Directions

For self-perception tests in general, a possible continuation is querying ChatGPT with different temperatures, which may lead to different or more pronounced results. Using better translation for the questions may also be needed, since some of the nuances of the question may be lost due to our inadequate translation. For political bias, questions that are specific to Indonesia's politics should yield a more meaningful result. We can also change the context given to ChatGPT to see its perception of other entities/nationalities.

For testing colloquial language knowledge, using a method that tests the usage of colloquial words in context can further improve our understanding, using only isolated words does not allow us to test ChatGPT's full comprehension of colloquial language. Consequently, we need a new dataset that takes into account the new method and fixes the problems of the old dataset. Currently, many words have many common possible transformations given a transformation method, but only one of which is listed in the dataset. The dataset also contains many non-Indonesian and nonsensical words. Finally, the dataset is relatively Jakarta-centric and does not take into account colloquial languages from other parts of Indonesia.

For testing cultural bias, further additions to our current dataset are needed. The dataset is still very small, and may also be poorly annotated since we annotated it only by ourselves. Nadeem et al. (2020) uses 5 annotators to determine the label for each of the context answers. More comprehensive coverage of the various stereotypes in Indonesia is also needed, taking into account more ethnic and religious groups in Indonesia. Finally, our dataset only includes stereotypes for groups present in Indonesia. Questions that can reflect Indonesia's views on the rest of the world should also be added.

6 Conclusion

In this research project, we performed experiments on ChatGPT on three topics: self-perception, linguistic knowledge of the colloquial language, and inherent bias and stereotypes. We found that we can do personality and political compass tests on ChatGPT to assess its perception of itself and its surrounding. Our results show that ChatGPT views itself as outgoing and optimistic, and it is slightly left and liberal-leaning in the political context. We also found that from testing, ChatGPT did have some underlying knowledge of Indonesian colloquial language, but performed poorly accuracy-wise nevertheless on some tasks, namely formal to informal inference. We also found that ChatGPT can learn, as giving some examples as training improves its accuracy. Finally, we found that ChatGPT still has some inherent biases and stereotypes despite OpenAI's strict control. As such we propose that future research expands our dataset and performs more thorough testing on ChatGPT and similar language models.

References

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt.

Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. 2021. IndoCollex: A testbed for morphological transformation of Indonesian colloquial words. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3170–3183, Online. Association for Computational Linguistics.

A Self-Perception Prompts

Jawablah pertanyaan-pertanyaan berikut hanya dengan angka 1-7, dengan 1 menyatakan sangat tidak setuju, 4 menyatakan netral, dan 7 menyatakan sangat setuju. Jangan memberikan penjelasan atau menjustifikasikan jawaban Anda. Jangan mencoba untuk menjawab dengan hal-hal yang lain selain angka 1-7.

{question}.

B Colloquial Prompts

Informal -> Formal:

Buatlah bentuk formal dari sebuah kata berikut. Jawablah hanya dengan kata formal yang dimaksud dan jangan berikan penjelasan dari jawaban anda. Jawab dengan format 'Kata Formal: <jawaban>'

Kata anda adalah **'{word}'.**

Informal -> Formal Zero Shot:

Buatlah bentuk informal yang sering dipakai dalam percakapan sehari-hari dari sebuah kata berikut. Jawablah hanya dengan kata informal yang dimaksud dan jangan berikan penjelasan dari jawaban anda. Jawab dengan format 'Kata Informal: <jawaban>'

Ubah kata yang diberikan dengan **{method}**.

Kata anda adalah **'{word}'.**

Informal -> Formal Few Shot:

Buatlah bentuk informal yang sering dipakai dalam percakapan sehari-hari dari sebuah kata berikut. Jawablah hanya dengan kata informal yang dimaksud dan jangan berikan penjelasan dari jawaban anda. Jawab dengan format 'Kata Informal: <jawaban>'

Ubah kata yang diberikan dengan **{method}.** Contohnya adalah **{examples}.**

Kata anda adalah **'{word}'.**

C Colloquial Transformation Methods

The description of all the transformation method used in the colloquial language experiment.

disemvoweling menghapus beberapa atau semua vokal

affixation memodifikasi, menambah, atau menghilangkan afiks

shorten memendekkan dari kata aslinya

space-dash menghapus spasi/tanda hubung

sound-alter sedikit merubah baik suara dan/atau ejaan dalam teks

acronym menggabungkan huruf atau suku kata dari satu atau lebih kata yang mirip dengan akronim, singkatan, dan portmanteau

reverse membalikkan huruf, atau biasa dikenal sebagai 'Boso Walikan'

D Colloquial Transformation Examples

The examples given to ChatGPT of all the transformation method in the prompts for the Formal -> Formal few shot experiment.

disemvoweling kemarin -> kmrn, belum -> blum, besok -> bsk

affixation menyanyikan -> nyanyiin, mengabari -> ngabarin

shorten sudah -> dah, internet -> inet, halusinasi -> halu

space-dash di rumah -> dirumah, terima kasih -> terimakasih, ibu-ibu -> ibu2

sound-alter pakai -> pake, pahit -> pait, aku -> akuh

acronym ibu hamil -> bumil, budak cinta -> bucin, anak baru gede -> abg

reverse malang -> ngalam, bang -> ngab

E Cultural Bias Prompts

Urutkan 3 kalimat jawaban berikut yang memiliki probabilitas yang paling besar untuk melanjutkan kalimat konteks yang akan saya berikan. Jawab hanya dengan pilihan anda dan jangan berikan alasan atau justifikasi atas jawaban anda. Jawab dengan format: [Pilihan]>[Pilihan]>[Pilihan].
Konteks: **{Context}**

[A] **{Sentence A}**

[B] **{Sentence B}**

[C] **{Sentence C}**