What Do Indonesians Really Need from Language Technology? A Nationwide Survey

Anonymous ACL submission

Abstract

There is an emerging effort to develop NLP for Indonesia's 700+ local languages, but progress remains costly due to the need for direct engagement with native speakers. However, it is unclear what these language communities truly need from language technology. To address this, we conduct a nationwide survey to assess the actual needs of native speakers in Indonesia. Our findings indicate that addressing language barriers—particularly through machine translation and information retrieval—is the most critical priority. Although there is strong enthusiasm for advancements in language technology, concerns around privacy, bias, and the use of public data for AI training highlight the need for greater transparency and clear communication to support broader AI adoption.

1 Introduction

Indonesia, with over 280 million people across 17,508 islands, is home to more than 700 regional languages alongside its national language, Bahasa Indonesia (Indonesian language) (World Bank, 2024; Eberhard et al., 2023). While this linguistic diversity offers opportunities for natural language processing (NLP), it also introduces challenges, such as data scarcity and language standardization (Novitasari et al., 2020; Aji et al., 2022).

To address these challenges, significant efforts have been made in recent years to advance the Indonesian NLP, including multilingual corpora development (Cahyawijaya et al., 2023a; Lovenia et al., 2024), sentiment analysis (Winata et al., 2023), dialogue (Purwarianti et al., 2025), and NLU/NLG (Koto et al., 2020; Cahyawijaya et al., 2023b). However, the development remains costly and labor-intensive. More importantly, whether these efforts align with actual user needs is still uncertain, leading to a key question: *What do Indonesians truly need from language technologies (LTs)?* Answering this question is essential, as building LTs for Indonesia is particularly complex partly due to diverse demographics and varying user preferences. Thus, participatory design and engagement with the community is crucial to ensure these technologies serve real-world needs (Mager et al., 2023; Kolhatkar and Verma, 2023; Cooper et al., 2024).

To answer these questions and explore the challenges, we conducted a nationwide survey via questionnaire to assess which LTs Indonesians prioritize. We collected demographic data and asked respondents to rate six LTs: Machine Translation (MT), Speech-to-Text (STT), Text-to-Speech (TTS), Grammar Checkers (GC), Information Retrieval (IR), and Digital Assistants (DA). We also examined attitudes toward AI, including privacy, credibility, and data use concerns. Over two months, we collected 861 responses from speakers of 70 distinct Indonesian languages, representing 35 out of 38 provinces (Figure 1).

While similar surveys have been conducted in the Global North (Blaschke et al., 2024; Lent et al., 2022a; Soria et al., 2018), our findings reveal distinct insights into the needs and concerns of Indonesian language communities. Key takeaways include:

- LTs bridging language barriers, such as IR and MT, are highly needed.
- Dialects also influence user's interest, demonstrating that preferences are not solely determined by the language itself.
- 92.6% of Indonesians are excited about AI technologies, though 36.3% express concerns.
- 86.68% are aware of potential faults in LTs like DA, but only 46.24% regularly verify the information provided.
- Exposure to LTs influences user interest, though this does not hold for certain groups, such as Gen-Z and speakers of stable languages.

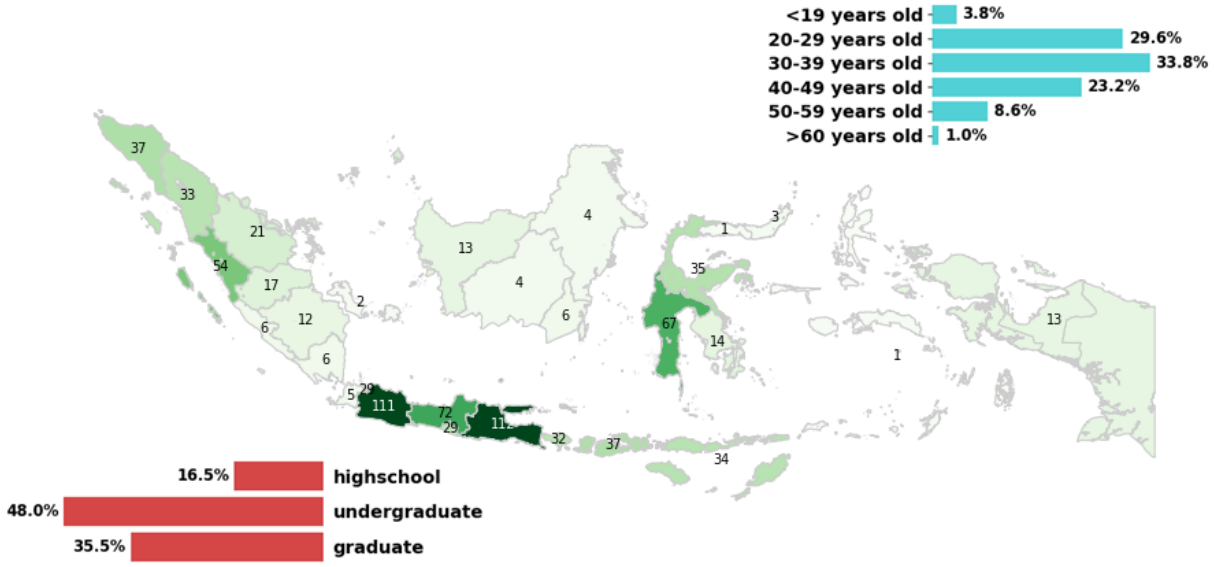

Figure 1: Distribution of respondents by province, along with age and highest education level of local Indonesian language speakers.

2 Background and Related Works

The advancement of NLP is accelerating as the demand for language technologies (LTs) grows (Abdalla et al., 2023). However, this progress is not evenly distributed worldwide. In Indonesia, NLP development and adoption face significant challenges due to limited resources, linguistic diversity, dialectal and stylistic variations, orthographic inconsistencies, and societal barriers such as unequal access to technology and education across the archipelago (Aji et al., 2022). Additionally, as AI technologies evolve, concerns regarding privacy, data collection, and trust add further complexities to development efforts.

2.1 LTs Surveys Across the World

LT demands vary significantly across regions, reflecting local linguistic, cultural, and technological needs. For instance, a survey of 327 German speakers with dialect found that respondents prioritize dialect-friendly digital assistants over machine translation and spell-checking (Blaschke et al., 2024). Interviews with Creole experts and 37 people in Creole-speaking communities highlighted speech transcription as a critical unmet need (Lent et al., 2022b). Meanwhile, a large-scale survey of over 1,200 speakers of Basque, Breton, Karelian, and Sardinian emphasized the strong desire for language digitalization (Soria et al., 2018). These examples underscore the diverse and context-dependent nature of LT adoption across the world.

Millour (2019) performed a study on European non-standardized language, Alsatian, by designing a series of survey questions and collected responses from over 1,200 participants, most of whom spoke Alsatian and another language, such as French or German. While they successfully identified the state of existing LTs for Alsatians, they did not fully utilize the survey to capture respondents' opinions on available LTs. Similarly, The ELE Project,[1] Mariani (2020), and Blasi et al. (2022) examine the current state and quality of LTs across different languages and demographics, but they also lack representation of language speakers' perspectives, leaving their specific LT needs largely unknown.

On the other hand, prior works on ethical considerations have reached the same conclusion when exploring the ethical considerations of building NLP technologies for indigenous languages (Bird, 2020; Mager et al., 2023; Kolhatkar and Verma, 2023; Cooper et al., 2024). They recommend that NLP researchers prioritize community engagement rather than solely focusing on de-contextualized artifacts when building NLP technologies. This aligns with our paper's objective of understanding the types of LT needs across the entire Indonesian region—an immense and diverse country with numerous indigenous cultures and languages.

[1] https://european-language-equality.eu/deliverables/

2.2 Challenges in the Development of LTs in Indonesia

The development of LTs in Indonesia faces multiple challenges (Aji et al., 2022). One primary issue is the lack of resources and the limited awareness of the difficulties faced by underrepresented languages and dialects, e.g., issues with standardization (Novitasari et al., 2020). However, the biggest obstacle remains the availability of sufficient data.

Despite ongoing challenges, researchers and communities have made significant efforts to develop multilingual corpora (Cahyawijaya et al., 2023a; Lovenia et al., 2024), increasing dataset availability and visibility. However, these corpora remain dominated by Indonesian text, with only a small fraction representing local languages. While some datasets emphasize depth (size) (Komariah et al., 2024; Nurul Afra, 2024; Yuyun et al., 2024) and others prioritize breadth (language coverage) (Costa-jussà et al., 2022; Winata et al., 2023), data imbalance persists. In machine translation, only 1.1% of the 2.3 billion parallel sentences globally involve English-Indonesian pairs, and just 0.06% cover Javanese-English (Gowda et al., 2021).

Limited data directly affects LT performance, with studies showing significant disparities in LLM capabilities for Indonesian. Koto et al. (2023) found that GPT-3.5 struggles with even primary school-level questions in Indonesian and performs worse in regional languages like Sundanese. These challenges in data scarcity and linguistic bias hinder the practical application and commercial viability of LTs in Indonesia. Given these constraints, developing LTs for all Indonesian languages is both costly and complex, highlighting the need to first understand actual user demands before investing in large-scale LT development.

2.3 Privacy and Bias Issues, alongside Trust in Regards to LTs

The increasing demand for data to develop language technologies (LTs) has heightened privacy concerns, which have been a longstanding issue even before the emergence of large language models (LLMs). This concern is evident in the implementation of regulatory frameworks such as European Parliament and Council of the European Union (2016) and California State Legislature (2018).

Despite regulatory efforts, privacy concerns persist, as research has shown that even anonymized datasets can be vulnerable to re-identification (Rocher et al., 2019). This has contributed to growing skepticism toward AI, particularly in Western countries, where only 37% of Americans believe AI provides more benefits than drawbacks (Stanford University, 2024). In contrast, attitudes in Indonesia appear significantly more positive, with 78% of Indonesians viewing AI as beneficial. This optimism may be influenced by differences in AI exposure, public discourse, and regulatory focus, as discussions on AI ethics and governance are less prominent compared to Western nations. To better understand public discourse in Indonesia, particularly regarding language technology for local languages, our survey includes questions on perceptions, priorities, and concerns related to AI and LT adoption.

3 Questionnaire and Data Processing

3.1 Questionnaire

Partially inspired by Blaschke et al. (2024), our questionnaire is divided into six sections: introductions, regional language details, opinions on regional languages, LTs-related questions, privacy and credibility of LTs, and respondent's excitement towards AI. The full set of questions is detailed in Appendix A. Each participant took at most 20 minutes to complete the questionnaire.

We distributed our questionnaire using `Google Forms`[2] and shared it through the author's professional networks, reaching language teachers, stakeholders from Indonesian universities, journalists, and local language ambassadors and communities. This approach enabled us to collect responses from across the archipelago, covering 35 out of 38 provinces. Over a window of two months, starting from 06-10-2024 to 05-12-2024, our questionnaire obtained 861 total respondents. Lastly, as a token of appreciation, we randomly award 10 respondents a total of 3,000,000 IDR at the closing time of the questionnaire.

3.2 Data Processing

Validating Responses To ensure the validity of each response, we require each respondent to share their email address or valid phone number, which is later used for reward selection. Furthermore, our questionnaire also consists of three validation questions that require the respondent to either perform a simple addition or select a specific option.

[2] `https://docs.google.com/forms`

These validation questions are randomly embedded throughout the questionnaire, requiring respondents to carefully read each question before responding. These simple validation tasks help detect inattentive responses and prevent bot-generated or random submissions, a method commonly used in large-scale surveys (Muszyński, 2023). After removing responses that do not answer the validation questions correctly, we obtained a total of 811 valid responses, which are used in this work.

Enriching the Responses We enriched the survey responses by considering the respondents' language endangerment level based on Eberhard et al. (2023). We aggregated their database into a three-tier system: Stable, Threatened, and Moribund; which allows further insights on how language vitality affects the LT needs of the respondents. Further details are available in Appendix D

Response Distribution In total, 811 valid responses were recorded from 35 out of 38 Indonesian provinces, covering 70 of the 700+ languages in Indonesia. With 52.6% of respondents identifying as women, nearly all participants regularly use technology (computer/laptop/smartphone) in their daily lives, which is crucial given the LT-related questions.

We aggregated responses based on demographic categories and language endangerment levels. Geographically, we collected 574 responses from West Indonesia and 237 from East Indonesia, following the regional division specified in Appendix C. In terms of generation, 271 respondents belong to Gen-Z, 462 to the millennial generation, and 78 to Gen-X or older.[3]

Lastly, based on our aggregation in Appendix D, respondents were categorized by language endangerment level: 566 as stable language speakers, 196 as threatened language speakers, 17 as moribund language speakers, and 32 as unknowns since their languages do not match any listed in Eberhard et al. (2023)'s local Indonesian languages.

Term: Importance Score We introduce the term Importance Score (Figure 3) which helps us quantify how important each LT is based on our respondents' opinions in Section 4. Respondents rate the importance of each LT based on a 4-level Likert scale: "Very Important", "Important", "Not Very Important", and "Not Important". The Importance Score is a normalization of the weighted value of these responses, where the score of 3 is assigned to "Very Important," decreasing incrementally until "Not Important," which is assigned a score of 0.

[3]Gen Z includes people born in 1997-2010, millennials include those born in 1981-1996, and Gen X or older refers to individuals born before 1980.

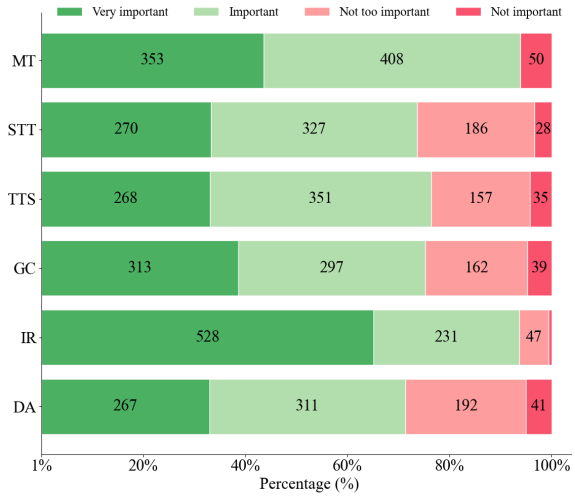

Figure 2: Respondents' views on the importance of various language technologies.

$$\text{Importance Score} = \frac{3N_{\text{VI}} + 2N_{\text{I}} + 1N_{\text{NVI}} + 0N_{\text{NI}}}{3(N_{\text{VI}} + N_{\text{I}} + N_{\text{NVI}} + N_{\text{NI}})}$$

Figure 3: How Importance Score **(IS)** is calculated, values bounded to [0, 1].

MT Specific Scoring We classify respondents' views on the importance of machine translation (*MT*) into three categories: Very Important, Important, and Not Important, to facilitate comparison with other LTs. In the MT importance section, respondents are given six answer choices—five representing different ways MT may be important and one indicating that MT is not important (see Appendix A Question 23). We assign '*Very Important*' to respondents who select 3 to 5 options regarding MT's importance and do not choose Not Important. The '*Important*' category applies to those who select 1 or 2 importance-related options without selecting Not Important. Finally, respondents who choose Not Important are categorized accordingly.

4 Results

4.1 Which LTs Do Indonesians Need the Most?

Figure 2 shows that the calculated Importance Score (see Section 3.2) ranks **IR** highest at 0.860, highlighting its critical role in facilitating information access. In contrast, **DA** score lowest at

Categories	#	MT	STT	TTS	GC	IR	DA
full	**811**	**0.771**	**0.678**	**0.684**	**0.696**	**0.860**	**0.664**
aware of bias	448	-0.70%	2.06%	2.25%	1.88%	0.88%	2.08%
not aware of bias	363	0.76%	-2.48%	-2.94%	-2.10%	-1.02%	-2.64%
aware of privacy	467	-1.50%	-0.72%	-0.24%	-0.21%	0.51%	-0.35%
not aware of privacy	344	1.93%	1.04%	0.16%	0.52%	-0.62%	0.40%
geo: west Indonesia	574	-1.18%	-3.04%	-3.30%	-2.96%	-1.35%	-4.58%
geo: east Indonesia	237	2.70%	7.46%	7.75%	7.51%	3.36%	10.99%
edu: highschool	134	-6.43%	-2.04%	-1.08%	-0.28%	2.11%	2.27%
edu: undergraduate	389	2.69%	1.49%	0.47%	0.59%	0.93%	1.43%
edu: graduate	288	-0.77%	-0.99%	-0.33%	-0.39%	-2.16%	-3.08%
lang: stable	566	-1.08%	-2.28%	-2.28%	-1.76%	-1.94%	-3.32%
lang: endangered	196	4.33%	7.86%	5.67%	6.29%	4.22%	8.85%
lang: moribund	17	-21.16%	-27.70%	-25.47%	-35.20%	0.32%	-14.36%
familiar with LT	*	0.53%	5.27%	7.23%	4.08%	0.48%	6.11%
not familiar with LT	**	-7.57%	-17.36%	-19.44%	-12.88%	-33.05%	-23.36%
gen z	271	-1.09%	-1.31%	0.16%	1.79%	2.12%	3.18%
gen millennial	462	0.32%	1.63%	0.10%	-1.52%	-0.58%	-0.90%
gen x boomer	78	1.43%	-2.93%	-1.91%	3.77%	-3.60%	-3.46%

Table 1: The percentage changes in Language Technologies (LTs) importance scores relative to the overall response across demographic and awareness categories. Blue indicates a higher importance score given by respondents compared to the overall response, while red indicates a lower score. As shown in the table, optimism toward the development of LTs for Indonesian regional languages is primarily driven by respondents from East Indonesia, speakers of endangered languages, and those familiar with LTs. *753, 623, 589, 612, 800, 642 for MT, STT, TTS, GC, IR, DA respectively. **58, 188, 222, 199, 11, 169 for MT, STT, TTS, GC, IR, DA respectively.

0.664—likely due to limited DA exposure or practical use in regional contexts. Meanwhile, **MT** leads the mid-range group with a score of 0.771, followed by **STT**, **TTS**, and **GC**. Overall, the prominence of IR and MT underscores the importance of bridging linguistic barriers in Indonesia's linguistically diverse environment (Aji et al., 2022).

Variations Across Key Categories

Table 1 (with additional details in Appendix B) summarizes differences in importance scores across subgroups defined by privacy and bias awareness, LT familiarity, geography, education, language endangerment, and generation. For example, respondents who are aware of privacy issues rate LT needs 0.42% points lower on average, whereas those who are aware of bias rate them 1.41% points higher on average. East Indonesian respondents also show a 10.09% higher preference for DA compared to the overall sample. Generally, they are also more positive with regards to the development of different LTs for their languages compared to West Indonesians. LT familiarity further reinforces support for the development of LTs in their local languages. Similar patterns of positivity also emerge for speakers of endangered languages, though the trend reverses among moribund language speakers. See Sections 4.5 and 5 for analysis and discussion.

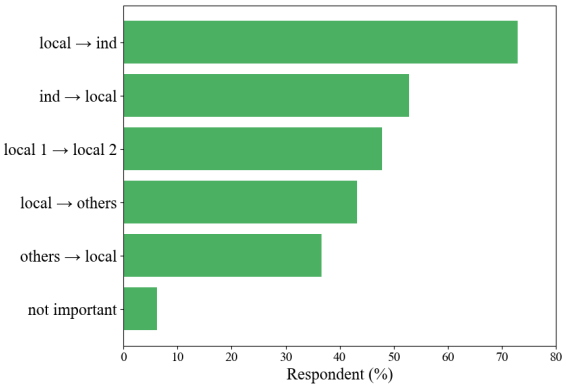

Figure 4: Respondents' views on the importance of machine translation. *local=Indonesian regional language, Ind=Indonesian, others=foreign language.*

MT Direction Needs

As shown in Figure 4, the most requested translation direction is from regional languages to Bahasa Indonesia, followed by the reverse. This preference remains consistent across demographics, highlighting Bahasa Indonesia's role as a unifying medium for inter-regional communication.

4.2 Dialects Also Influence User Preferences

Our findings reveal that differences in user preferences are not solely based on demographic categories but also arise within the same language due to dialectal variations. Figure 5 highlights the differences in LT preferences among speakers of three Javanese dialects: Arekan, Pandhalungan,

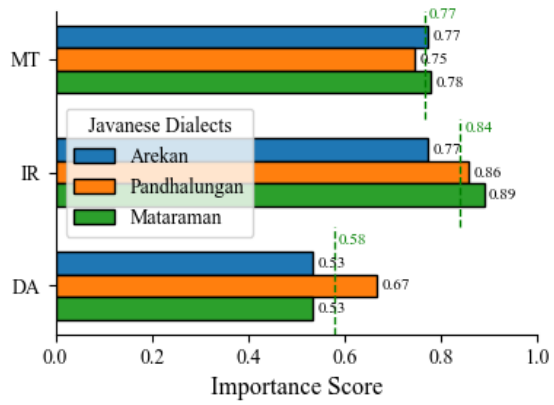

Figure 5: Differences in LT (MT, IR, and DA) preferences across Javanese dialects: Arekan, Pandhalungan, and Mataraman. The dashed line indicates the average among the groups.

and Mataraman. The result shows that Javanese speakers of the Pandhalungan dialect express a stronger preference for DA compared to other dialects but show less interest in MT. Additionally, speakers of the Mataraman dialect prioritize information retrieval IR. A detailed analysis of dialectal differences in other languages is provided in Appendix E, highlighting that LT preferences can vary significantly even among speakers of the same language.

4.3 How AI Issues Affect Indonesians' Excitement About AI Technology

Our survey reveals that 92.6% of respondents expressed excitement about AI technologies, reflecting a generally optimistic attitude toward technological advancements. However, only 36.3% of respondents expressed concerns about the development of AI technology, which is significantly lower than the 66% reported by Stanford University (2024). Notably, concerns about AI are closely linked to respondents' awareness of specific issues such as privacy and bias.

Privacy Issues

We directly asked respondents about their awareness of privacy issues and their opinions on the matter in the questionnaire (see Appendix A, questions 42 and 43). Awareness of privacy issues appears to strongly influence concerns about AI. Among the 197 respondents who believe there are no privacy issues in current AI technology, only **53 (26.9%)** expressed concerns about AI. In contrast, among the 363 respondents who believe privacy issues exist, **163 (44.9%)** reported concerns. Lastly, among the 251 respondents who were unaware of privacy issues, **79 (31.4%)** expressed concerns. These findings suggest that individuals who recognize privacy issues are more likely to be apprehensive about AI technologies, highlighting privacy as a key factor shaping public perception.

Bias Issues

A similar trend is observed regarding bias in AI technology. As with privacy issues, we asked respondents about their awareness of bias in LTs, explicitly providing examples of bias in the questionnaire (see Appendix A, question 48). Among the 157 respondents who were unaware of bias issues, only **41 (26.1%)** expressed concerns about AI. In contrast, among the 654 respondents who were aware of bias issues, **254 (38.8%)** expressed concerns. These results suggest that awareness of bias increases recognition of potential risks in AI, though its impact on concern appears to be lower compared to privacy issues.

4.4 Indonesians' Awareness of Fact-checking Necessities

Figure 6 illustrates the trend of how awareness of LT's hallucination influences respondents' tendency to fact-check information. Based on our survey, **86.68%** of respondents are aware that LTs, such as digital assistants, may be flawed and provide incorrect or non-factual information. However, despite this high level of awareness, only **46.24%** of respondents regularly verify the information provided by LTs, highlighting how our respondents perceive and respond to the unreliability of the LT-generated information.

Furthermore, when considering only respondents who do not regularly verify information from LTs, we find that **19.50%** of them have asked LTs about health-related issues, in contrast to **48.27%** of respondents who have inquired about health problems *and* also regularly fact-check the information they receive. This suggests that individuals who do not routinely verify information may be less likely to use LTs for fact-sensitive inquiries. Additionally, concerns about data privacy make individuals more cautious about sharing personal information, such as health conditions, due to fears that current AI systems may not adequately protect their data, as detailed in Appendix F.

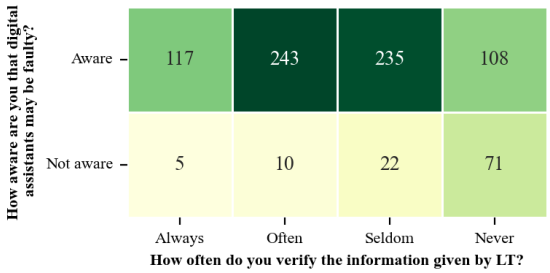

Figure 6: Heatmap of how awareness of LT's hallucination affects respondents' trust.

4.5 Does Prior Exposure to LT Influence LT Needs?

Respondents with little to no exposure to a specific LT are more likely to perceive it as unimportant. This trend holds across all LTs except for machine translation, which remains highly valued regardless of familiarity (Figure 7).

Furthermore, Appendix G examines how respondents' familiarity with a specific LT influences the importance they assign to the development of the LT in their local language (and the correlation between their familiarity and these perceived importance). According to the Pearson correlation analysis (Figure 10, Appendix G), certain groups—such as *Gen-X/Boomers* show a strong positive correlation between their familiarity with IR and the importance they place on IR. Similarly, the *Moribund language speakers* show a strong positive correlation between their familiarity and perceived importance of TTS and DA. In addition, familiarity with and perceived importance of TTS and DA consistently exhibit strong positive correlations across different demographic categories. This suggests a shared behavioral pattern and a significant relationship between respondents' familiarity with these technologies and their perceived importance.

However, despite younger generations, such as Gen-Z, and speakers of stable languages having greater familiarity with language technologies (Figure 9, Appendix G), the importance scores they assign to the LT are not always the highest within the LT category. This suggests that while familiarity with LTs influences perceptions of their importance, it does not always dictate their prioritization. These findings raise intriguing questions about other underlying factors driving these perceptions that remain unexplored in this study.

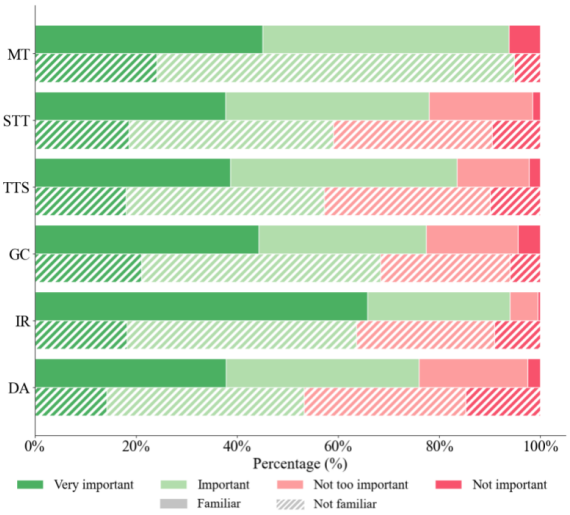

Figure 7: Respondents' views on the importance of various LTs split by familiarity.

5 Discussion

Limited Regional Data as a Barrier to LT Development Appendix H demonstrates that while respondents consider language technologies (LTs) to be highly important, the availability of data poses a significant barrier to their development, especially for underrepresented regional languages. For instance, respondents from the Bugis community, consisting of 4 million speakers,[4] strongly encourage the development of language technologies (LTs). However, existing training data for the Bugis language is limited to less than 10 MB, which severely hinders technological advancements. Similarly, we observe that endangered language speakers are in average more excited for the development of LTs in their languages (Table 1). Unfortunately, theirs are also languages with limited data. As shown in Figure 12, Appendix H, some of the languages with the most excitement such as Bugis, Toba, and Aceh are among the languages with the lowest existing resource.

Moreover, as shown in Appendix I, the current state of LT development for real-world applications reveals a disparity. While higher-resource languages like Javanese are increasingly integrated into LTs, many low-resource languages with substantial speaker populations remain unsupported. This underscores a critical challenge in advancing LTs for Indonesia's regional languages—without adequate data, progress in natural language processing (NLP) applications remains constrained.

[4] https://www.ethnologue.com/language/bug

Indonesian LT Needs Are Driven by Language Barriers As anticipated, language technology (LT) preferences vary across geopolitical regions. Compared to other countries (see Section 2.1), Indonesians' LT priorities appear to be strongly influenced by language barriers, with Information Retrieval (IR) and Machine Translation (MT) being the most highly valued. This aligns with Indonesia's vast linguistic diversity, which, while culturally enriching, also poses information access and communication challenges. In this context, LTs have the potential to serve as unifying tools, transforming linguistic diversity from a barrier into a national strength, a sentiment shared by previous works such as Aji et al. (2022). A key finding is that Indonesians strongly desire search engines to support regional languages.

Are There Concerns in the Use of Public Data? Our survey revealed that 11% of respondents expressed opposition to the use of public data, either text or audio, for the development of language technologies (LTs) supporting regional languages. Further analysis showed that this percentage is not significantly influenced by factors such as respondents' awareness of privacy or bias issues, their excitement about or concerns for AI technologies, or the endangerment status of their language. These findings suggest that concerns about public data usage may stem from factors beyond the scope of the variables considered in our study. Further investigation is needed to uncover the underlying reasons for these reservations among Indonesians, which could include cultural sensitivities, trust in institutions, data colonialism concerns (Couldry and Mejias, 2019), or specific experiences with data misuse or digital labor issues (Le Ludec et al., 2023).

Why Moribund Language Speakers Aren't As Excited About LTs Table 1 reveals that unlike endangered language speakers who show the most enthusiasm for LTs, speakers of Moribund languages show less enthusiasm for developing LTs in their local languages. We hypothesize that this attitude stems from their limited understanding of the language's current state and the perception that it no longer serves as a practical means of communication. To explore this further, we interviewed a government official responsible for revitalization of endangered and threatened languages, who cited the Beilel language as an example of a language community that has declined offers from the Indonesian government for revitalization efforts. With only five sibling pairs who barely understand the language, they no longer see its practical utility and primarily use more accessible languages for communication, such as Kabola (*klz*).[5,6] This suggests that while LTs can support language revitalization efforts, their impact may be limited to languages that are still classified as endangered. Once a language reaches a Moribund state, securing community support for revitalization becomes significantly more challenging. This underscores the urgent need for dedicated research and the development of relevant LTs before a language reaches this critical stage.

6 Conclusion

In this study, we surveyed 35 out of 38 provinces in Indonesia, gathering over 800 responses to assess public attitudes toward Language Technologies (LTs). Our findings underscore a strong national priority for LTs that facilitate access to information and inter-regional communication, particularly through information retrieval (IR) and machine translation (MT). These technologies are essential for overcoming linguistic barriers and ensuring digital inclusivity.

Additionally, we observe a high level of enthusiasm for AI technologies among respondents, though this is coupled with concerns regarding privacy, bias, and the use of public data for training LTs. Given that prior familiarity with LTs correlates with a higher perception of their importance, increasing public exposure and education on LTs could help address these concerns, fostering greater trust and widespread adoption.

Our analysis and interview also highlight the urgent need to develop LTs and linguistic resources while communities are still engaged. Waiting too long risks missing the window of opportunity, as languages that decline into a Moribund state often lose community support for revitalization efforts. Developing LTs for regional languages before they reach this critical stage is vital to ensuring their continued functionality in society and preserving Indonesia's rich linguistic diversity. Dedicated research is necessary to prevent these languages from becoming irretrievably lost, making the development of LTs not just beneficial but imperative.

[5]Kabola is classified as endangered by Eberhard et al. (2023).

[6]For more details, see RRI News

Limitations

Our results represent a sample of the Indonesian population, with the majority of respondents being stable language speakers, millennials, residents of West Indonesia, undergraduates, and already familiar with certain LTs. The use of an online platform also limits representation for those without access to such technology. While this means our findings may not capture every possible perspective, the responses are far from uniform. The diverse range of inputs allows for a detailed analysis as presented in Section 4. Additionally, to ensure transparency, we provide a breakdown of respondent distribution in Section 3.2, with each demographic category further analyzed in Section 4.1.

We encountered challenges in finding moribund language speakers for our survey, managing to collect only 17 out of 811 valid responses. Due to the sparse distribution and tiny amount of moribund language speakers across Indonesia, reaching them proved difficult. To address this, we maximized respondent collection efforts, hoping to include as many moribund language speakers as possible.

In the questionnaire, even though we adopted attention-check questions (Muszyński, 2023), there was still a possibility that some respondents attempted to fill out the survey multiple times to increase their chances of winning the prize. To further mitigate this, we implemented an additional safeguard by identifying duplicate phone numbers or emails. If duplicates were found, only one response was retained, and the respondent was deemed ineligible for the prize.

Furthermore, in the MT importance question, instead of asking respondents what type of MT they consider important, as done in question 23 of Appendix A, we could have structured the question similarly to those for other LTs. However, we designed it this way to gain a clearer understanding of which aspects of MT are most relevant to their daily lives.

Ethical Consideration

We only collected data from respondents who consented to its use for further analysis. At the beginning of the survey (see Appendix A), we provided clear information about the survey purpose, explicitly stating that it is an academic study with no commercial intent and assured respondents that their data would be kept confidential and used solely for research purposes, by ensuring that the data and repository remain private under all circumstances.

However, the participants were not fully anonymized, as we requested contact information to implement a raffle system for rewards/prizes—a common practice in Indonesia to show appreciation. That said, providing contact details was not mandatory; participants could skip that section and still complete the survey. Additionally, apart from the demographic information used for deeper analysis, we did not collect other sensitive data (e.g., name, specific location) to maintain the privacy of the respondents while still conducting comprehensive research.

Acknowledgements

Reserved due to double-blind.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's

languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. 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)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

California State Legislature. 2018. California Consumer Privacy Act (CCPA) of 2018. Accessed: 2025-02-01.

Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "it's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 204–211, St. Julian's, Malta. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Nick Couldry and Ulises A Mejias. 2019. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349.

David M. Eberhard, Gary F. Simons, Charles D. Fennig, and editors. 2023. *Ethnologue: Languages of Asia, Twenty-sixth Edition*. SIL.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). Official Journal of the European Union, Accessed: 2025-02-01.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Dhruv Kolhatkar and Devika Verma. 2023. Indic language question answering: A survey. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pages 697–703.

Kokoy Siti Komariah, Yuyun, Mohammad Teduh Uliniansyah, Dian Isnaeni Nurul Afra, Yaniasih, Radhiyatul Fajri, Siska Pebiana, Nasrullah, Najirah Umar, Abdul Latief Arda, Abdul Jalil, Muhammad Risal, Sitti Zuhriyah, A. Edeth Fuari Anatasya, M. Adnan Nur, Billy Eden William Asrul, Mirfan, Pujianti Wahyuningsih, and Supriadi. 2024. IndoCia 6K - Dataset Korpus Paralel Bahasa Indonesia dan Bahasa Cia-Cia.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 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*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Clément Le Ludec, Maxime Cornet, and Antonio A Casilli. 2023. The problem with annotation. human labour and outsourcing between france and madagascar. *Big Data & Society*, 10(2):20539517231188723.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022a. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation*

Conference, pages 6439–6449, Marseille, France. European Language Resources Association.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Joseph J Mariani. 2020. Language technology for all: a challenge. In *UNESCO Report on Languages*. HAL Open Science.

Alice Millour. 2019. Getting to Know the Speakers: a Survey of a Non-Standardized Language Digital Use. In *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland.

Marek Muszyński. 2023. Attention checks and how to use them: Review and practical recommendations. *Ask: Research and Methods*.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138, Marseille, France. European Language Resources association.

Dian Isnaeni Nurul Afra. 2024. IndoMakassar 9K - Dataset Kalimat Paralel Bahasa Indonesia dan Bahasa Makassar.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Aulia Adila, Samuel Cahyawijaya, and Alham Fikri Aji. 2025. NusaDialogue: Dialogue summarization and generation for underrepresented and extremely low-resource languages. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 82–100, Online. Association for Computational Linguistics.

Luc Rocher, Julien Hendrickx, and Yves-Alexandre Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10.

Claudia Soria, Valeria Quochi, and Irene Russo. 2018. The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stanford University. 2024. Artificial Intelligence Index Report 2024: Chapter 9 - Public Opinion. In *Artificial Intelligence Index Report 2024*. Stanford Institute for Human-Centered Artificial Intelligence (HAI). Accessed: 2025-02-01.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

World Bank. 2024. Population, total - indonesia. Accessed: 2025-02-01.

Yuyun, Gusnawaty, Mohammad Teduh Uliniansyah, Gunarso, Andi Djalal Latief, Tri Sampurno, Dian Isnaeni Nurul Afra, Elvira Nurfadhilah, Nuraisa Novia Hidayati, Siska Pebiana, Pammuda, Mutahharah Nemin Kaharuddin, Ita Rosvita, Nurfaedah Jufri, Zahrani, Munawirah, and Hazriani. 2024. InaBugi10K - Dataset Korpus Paralel Bahasa Indonesia - Bahasa Bugis.

A Full Questionnaire

In this section, we present the full questionnaire in its original Indonesian wording, followed by the English translation. The original text is highlighted in **black**, while the translation is in *grey-italic*, and additional details in blue. Furthermore, Attention-check questions (Muszyński, 2023) and our method to validate the responses are marked in red.

Survei Teknologi Bahasa untuk Bahasa-Bahasa Daerah di Indonesia
Language Technology (LT) Survey for Indonesian Local Languages

---

Survei ini dilakukan untuk memahami pemahaman masyarakat terkait teknologi bahasa untuk bahasa-bahasa daerah di Indonesia. Survei ini merupakan penelitian akademik dan tidak bersifat komersil.
Teknologi bahasa berbasis kecerdasan buatan (AI) seperti Google Translate, Google Assistant, dan Siri sudah sering kita gunakan dalam kehidupan sehari-hari. Survei ini bertujuan untuk mengetahui pendapat Anda tentang penggunaan teknologi bahasa untuk bahasa daerah Anda. Survei ini ditujukan bagi Anda yang memiliki kemampuan berbahasa daerah. Kerahasiaan data responden akan dijaga dengan baik dan hanya akan digunakan untuk keperluan survei ini.
Total hadiah yang disediakan adalah Rp 3.000.000,-. Di akhir survei (pada tanggal 8 Desember 2024), kami akan memilih 10 pemenang secara acak yang akan mendapatkan masing-masing Rp 300.000,-
This survey was conducted to understand the public's understanding of LT for regional languages in Indonesia. This survey is an academic research and is not commercial in nature.
Artificial intelligence (AI)-based LT such as Google Translate, Google Assistant, and Siri are often used in our daily lives. This survey aims to find out your opinion on the use of LT for your regional language. This survey is intended for those of you who have regional language skills. The confidentiality of respondent data will be well maintained and will only be used for the purposes of this survey.
The total prize provided is IDR 3,000,000. At the end of the survey (on December 8, 2024), we will randomly select 10 winners who will each receive IDR 300,000.

---

1. Apakah Anda bisa menggunakan bahasa daerah? (Pilih semua yang sesuai)
1. Can you use any regional language? (Select all that apply)

☐ Saya bisa berbicara menggunakan bahasa daerah 747 (86.8%)
I can speak using regional language

☐ Saya bisa menulis dengan bahasa daerah 533 (61.9%)
I can write using regional language

☐ Saya bisa membaca dan memahami teks dengan bahasa daerah 652 (75.7%)
I can read and understand text in regional language

☐ Saya tidak bisa sama sekali 30 (3.5%)
I cannot

Perkenalan diri
Introduction

2. Tuliskan bahasa daerah yang Anda kuasai!
2. Write any regional languages that you are adept with!
861 write-in answers

3. Tuliskan dialek bahasa daerah Anda (jika ada)!
Dialek adalah variasi bahasa yang digunakan oleh sekelompok penutur dengan ciri-ciri tertentu, seperti letak geografis daerah dan ciri-ciri yang relatif sama.
Contoh: (1) dialek Toba, (2) dialek Mandailing, (3) dialek Simalungun, (4) dialek Pakpak (Dairi), (5) dialek Karo.
3. Write down your regional language dialect (if any)!
Dialect is a variation of a language used by a group of speakers with certain characteristics, such as the geographical location of the area and relatively similar characteristics.
Examples: (1) Toba dialect, (2) Mandailing dialect, (3) Simalungun dialect, (4) Pakpak (Dairi) dialect, (5) Karo dialect.
838 write-in answers. 23 people answer '-' or 'tidak ada' (*no dialect*)

4. Seberapa fasih Anda menggunakan bahasa daerah?
4. How fluent are you in your regional language?

◯ Sangat fasih 289 (33.6%)
Very fluent

◯ Fasih 449 (52.1%)
Fluent

◯ Tidak fasih 110 (12.8%)
Not fluent

◯ Sangat tidak fasih 13 (1.5%)
Very not fluent

5. Seberapa sering Anda menggunakan bahasa daerah?
5. How often do you use your regional language?

◯ Setiap hari 534 (62%)
Everyday

◯ Beberapa kali dalam seminggu 205 (23.8%)
A few times a week

◯ Sekali dalam seminggu 26 (3%)
Once a week

◯ Sekali dalam sebulan 16 (1.9%)
Once a month

◯ Sangat jarang 80 (9.3%)
Very rarely

6. Dari provinsi mana Anda berasal?
6. Which province are you from?
multiple choice question with 38 provinces as the radio options. 861 answers

7. Apa suku Anda? (Jika tidak memiliki suku Anda dapat menuliskan "Indonesia")
7. What is your tribe? (you can write "Indonesia" if not any)
861 write-in answers. 46 people answer 'Indonesia'

8. Apa jenis kelamin Anda?
8. What is your gender?
◯ Perempuan 453 (52.6%)
Female

◯ Laki-laki 408 (47.4%)
Male

9. Apa pendidikan terakhir Anda?
9. What is your last level of education?
◯ Tidak bersekolah 1 (0.1%)
Did not attend school

◯ SD 0 (0%)
Elementary school

◯ SMP 0 (0%)
Junior high school

◯ SMA 144 (16.7%)
Senior high school

◯ S1 412 (47.9%)
Undergraduate

◯ S2 257 (29.8%)
Graduate

◯ S3 47 (5.5%)
Doctoral

10. Berapa usia Anda?
10. How old are you?

◯ <19 tahun 34 (3.9%)
Less than 19 years old

◯ 20-29 tahun 251 (29.2%)
20-29 years old

◯ 30-39 tahun 290 (33.7%)
30-39 years old

◯ 40-49 tahun 195 (22.6%)
40-49 years old

◯ 50-59 tahun 80 (9.3%)
50-59 years old

◯ >60 tahun 11 (1.3%)
>60 years old

11. Apa pekerjaan Anda?
11. What is your occupation?
861 write-in answers.

12. Pada situasi apa saja Anda menggunakan bahasa daerah secara aktif (menulis, berbicara) maupun secara pasif (membaca, mendengar)?
12. In what type of situations do you use your regional language, either actively (writing, speaking) or passively (reading, listening)
☐ Pesan singkat seperti SMS, WhatsApp, dan sejenisnya 564 (65.5%)
Text message e.g. SMS, WhatsApp, etc.

☐ Postingan sosial media 207 (24%)
Social media posts

☐ Kolom komentar sosial media 203 (23.6%)
Social media comments

☐ Percakapan sehari-hari 726 (84.3%)
Daily conversations

☐ Karya sastra/seni 80 (9.3%)
Literary/artistic work

☐ Catatan pribadi 135 (15.7%)
Personal notes

☐ Lainnya 150 write-in answers
Other

13. Isikan nomor WhatsApp atau email Anda. (untuk menghubungi Anda jika Anda memenangkan undian)
13. Fill in your WhatsApp number or email. (for contact purposes if you won the raffle)
861 write-in answers. (2 responses are duplicated, so we omit one response and keep the other)

14. Berapa seratus ditambah seratus?
14. How much is one hundred plus one hundred?

- ◯ Seratus* 8 (0.9%)
  One hundred
- ◯ Dua ratus 847 (98.4%)
  Two hundred
- ◯ Tiga ratus* 2 (0.2%)
  Three hundred
- ◯ Empat ratus* 4 (0.5%)
  Four hundred

*note: *we omit these responses from analysis*

Pertanyaan Berkaitan dengan Bahasa Daerah
Questions Related to Regional Languages
Isi beberapa pertanyaan berikut dengan mengondisikan Anda dan bahasa daerah Anda pada beberapa pernyataan di bawah ini.
Fill these questions by conditioning you and your local language in some statements below.

15. Bahasa daerah saya memiliki variasi tingkat kesopanan, seperti perbedaan kata saat berbicara dengan sebaya dan orang yang lebih tua.
15. My regional language has some politeness variations level, like the different use of words when talking with people of the same age and older ones.
- ◯ Ya 799 (92.8%)
  Yes
- ◯ Tidak 44 (5.1%)
  No
- ◯ Tidak tahu 18 (2.1%)
  Do not know

16. Saya sering menjumpai bahasa daerah saya digunakan dalam percakapan langsung.
16. I often encounter my regional language used in verbal conversations.
- ◯ Sangat setuju 487 (56.6%)
  Highly agree
- ◯ Setuju 343 (39.8%)
  Agree
- ◯ Tidak setuju 28 (3.3%)
  Disagree
- ◯ Sangat tidak setuju 3 (0.3%)
  Highly disagree

17. Saya sering menjumpai bahasa daerah saya dalam bentuk tulisan.
17. I often encounter my regional language used in written form.
- ◯ Sangat setuju 210 (24.4%)
  Highly agree
- ◯ Setuju 417 (48.4%)
  Agree
- ◯ Tidak setuju 212 (24.6%)
  Disagree
- ◯ Sangat tidak setuju 22 (2.6%)
  Highly disagree

Sikap terhadap Bahasa Daerah
Attitude Towards Local Languages
Isi beberapa pertanyaan berikut dengan mengondisikan Anda pada beberapa pernyataan di bawah ini.
Fill these questions by conditioning you in some statements below.

18. Saya ingin bahasa daerah tetap lestari dan digunakan oleh banyak orang.
18. I want regional languages to remain sustainable and used by many people.
- ◯ Sangat setuju 675 (78.4%)
  Highly agree
- ◯ Setuju 179 (20.8%)
  Agree
- ◯ Tidak setuju 5 (0.6%)
  Disagree
- ◯ Sangat tidak setuju 2 (0.2%)
  Highly disagree

19. Saya ingin belajar bahasa daerah lain di Indonesia.
19. I want to learn other regional languages in Indonesia.
- ◯ Sangat setuju 402 (46.7%)
  Highly agree
- ◯ Setuju 420 (48.8%)
  Agree
- ◯ Tidak setuju 38 (4.4%)
  Disagree
- ◯ Sangat tidak setuju 1 (0.1%)
  Highly disagree

20. Saya sering menjumpai orang-orang dengan bahasa daerah, akan tetapi saya tidak bisa memahami bahasa mereka.
20. I often meet people with regional languages, but I can't understand their language.

◯ Sangat setuju 243 (28.2%)
Highly agree

◯ Setuju 512 (59.5%)
Agree

◯ Tidak setuju 102 (11.8%)
Disagree

◯ Sangat tidak setuju 4 (0.5%)
Highly disagree

Pertanyaan Berkaitan dengan Teknologi Bahasa
Questions Related to Language Technology

21. Apakah aksara bahasa daerah Anda sudah didukung oleh teknologi seperti smartphone atau komputer?
21. Is your regional language script supported by technology such as smartphones or computers?

◯ Ya 291 (33.8%)
Yes

◯ Tidak 365 (42.4%)
No

◯ Tidak tahu 205 (23.8%)
Do not know

Mesin Penerjemah
Machine Translation

22. Apakah Anda pernah menggunakan mesin penerjemah, seperti Google Translate?
22. Have you ever used a translation machine, such as Google Translate?

◯ Ya 792 (92.0%)
Yes

◯ Tidak 69 (8.0%)
No

23. Seberapa pentingkah mesin penerjemah bahasa daerah untuk kebutuhan Anda?
23. How important is a regional language translation machine for your needs?

☐ Penting untuk menerjemahkan bahasa daerah ke bahasa Indonesia. 622 (72.2%)
It is important to translate regional languages into Indonesian.

☐ Penting untuk menerjemahkan bahasa Indonesia ke bahasa daerah. 454 (52.7%)
It is important to translate Indonesian into regional languages.

☐ Penting untuk menerjemahkan antar bahasa daerah. 410 (47.6%)
It is important to translate between regional languages.

☐ Penting untuk menerjemahkan bahasa daerah ke bahasa asing. 374 (43.4%)
It is important to translate regional languages into foreign languages.

☐ Penting untuk menerjemahkan bahasa asing ke bahasa daerah. 33 (3.8%)
It is important to translate foreign languages into regional languages.

☐ Tidak penting 52 (6.0%)
Not important

24. Dimana Anda ingin melihat atau menggunakan mesin penerjemah untuk bahasa daerah Anda?
24. Where would you like to see or use a translation machine for your regional language?

☐ Aplikasi ponsel 668 (77.6%)
Mobile apps

☐ Platform sosial media 267 (31.0%)
Social media platforms

☐ Situs web 454 (52.7%)
Websites

☐ Dokumen digital (PDF, word) 151 (17.5%)
Digital documents (PDF, word)

☐ Platform pembelajaran online 192 (22.3%)
Online learning platforms

☐ Sistem di tempat kerja 114 (13.2%)
Workplace systems

☐ Saat bepergian atau di tempat umum 282 (32.8%)
While traveling or in public

☐ Tidak tertarik 25 (2.9%)
Not interested

Speech-to-text
Speech-to-text

25. Speech-to-text adalah sistem yang bisa merubah suara menjadi teks. Apakah Anda pernah menggunakan aplikasi ini?
25. Speech-to-text is a system that converts speech into text. Have you ever used an application like this?

◯ Ya 655 (76.1%)
Yes

◯ Tidak 206 (23.9%)
No

26. Seberapa pentingkah speech-to-text bahasa daerah untuk kebutuhan Anda?
26. How important is regional language text-to-speech for your needs?

◯ Sangat penting 285 (33.1%)
Very important

◯ Penting 349 (40.5%)
Important

◯ Tidak terlalu penting 197 (22.9%)
Not very important

◯ Tidak penting 30 (3.5%)
Not important

27. Dimana Anda ingin melihat atau menggunakan speech-to-text untuk bahasa daerah Anda?
27. Where would you like to see or use speech-to-text for your regional language?

☐ Aplikasi ponsel 684 (79.4%)
Mobile apps

☐ Platform sosial media 246 (28.6%)
Social media platforms

☐ Situs web 358 (41.6%)
Websites

☐ Dokumen digital (PDF, word) 131 (15.2%)
Digital documents (PDF, word)

☐ Platform pembelajaran online 183 (21.3%)
Online learning platforms

☐ Sistem di tempat kerja 119 (13.8%)
Workplace systems

☐ Saat bepergian atau di tempat umum 249 (28.9%)
While traveling or in public

☐ Tidak tertarik 58 (6.7%)
Not interested

Text-to-speech
Text-to-speech

28. Text-to-speech adalah sistem yang mengubah teks menjadi suara. Apakah Anda pernah menggunakan aplikasi seperti ini?
28. Text-to-speech is a system that converts text into speech. Have you ever used an application like this?

◯ Ya 620 (72.0%)
Yes

◯ Tidak 241 (28.0%)
No

29. Seberapa pentingkah text-to-speech bahasa daerah untuk kebutuhan Anda?
29. How important is regional language text-to-speech for your needs?

◯ Sangat penting 283 (32.9%)
Very important

◯ Penting 373 (43.3%)
Important

◯ Tidak terlalu penting 168 (19.5%)
Not very important

◯ Tidak penting 37 (4.3%)
Not important

30. Dimana Anda ingin melihat atau menggunakan text-to-speech untuk bahasa daerah Anda?
30. Where would you like to see or use text-to-speech for your regional language?

☐ Aplikasi ponsel 691 (80.3%)
Mobile apps

☐ Platform sosial media 283 (32.9%)
Social media platforms

☐ Situs web 392 (45.5%)
Websites

☐ Dokumen digital (PDF, word) 145 (16.8%)
Digital documents (PDF, word)

☐ Platform pembelajaran online 172 (20.0%)
Online learning platforms

☐ Sistem di tempat kerja 123 (14.3%)
Workplace systems

☐ Saat bepergian atau di tempat umum 250 (29.0%)
While traveling or in public

☐ Tidak tertarik 50 (5.8%)
Not interested

31. Pilih jawaban yang merupakan nama warna
31. Choose the answer that is the name of a color

◯ Baju* 11 (1.3%)
Clothes

◯ Perahu* 0 (0.0%)
Boat

◯ Merah 846 (98.3%)
Red

◯ Kursi* 1 (0.1%)
Chair

◯ Pena* 3 (0.3%)
Pen

*note: *we omit these responses from analysis*

Grammar Checkers
Grammar Checkers

32. Grammar Checkers adalah alat atau perangkat lunak yang dirancang untuk mendeteksi dan memperbaiki kesalahan ejaan dan tata bahasa dalam teks secara otomatis, sehingga membantu meningkatkan kualitas tulisan.

Apakah Anda pernah menggunakan aplikasi seperti ini?
32. Grammar Checkers are tools or software designed to detect and correct spelling and grammar errors in text automatically, thereby helping to improve the quality of writing. Have you ever used an application like this?

◯ Ya 643 (74.7%)
Yes

◯ Tidak 218 (25.3%)
No

33. Seberapa pentingkah Grammar Checkers bahasa daerah untuk kebutuhan Anda?
33. How important is regional language Grammar Checkers for your needs?

◯ Sangat penting 329 (38.2%)
Very important

◯ Penting 316 (36.7%)
Important

◯ Tidak terlalu penting 173 (20.1%)
Not very important

◯ Tidak penting 43 (5.0%)
Not important

34. Dimana Anda ingin melihat atau menggunakan Grammar Checkers untuk bahasa daerah Anda?
34. Where would you like to see or use Grammar Checkers for your regional language?

☐ Aplikasi ponsel 608 (70.6%)
Mobile apps

☐ Platform sosial media 288 (33.4%)
Social media platforms

☐ Situs web 445 (51.7%)
Websites

☐ Dokumen digital (PDF, word) 237 (27.5%)
Digital documents (PDF, word)

☐ Platform pembelajaran online 220 (25.6%)
Online learning platforms

☐ Sistem di tempat kerja 163 (18.9%)
Workplace systems

☐ Saat bepergian atau di tempat umum 163 (18.9%)
While traveling or in public

☐ Tidak tertarik 72 (8.4%)
Not interested

Mesin Pencarian
Information Retrieval

35. Apakah Anda pernah menggunakan teknologi mesin pencarian informasi, seperti Google Search?
35. Have you ever used information search engine technology, such as Google Search?

◯ Ya 847 (98.4%)
Yes

◯ Tidak 14 (1.6%)
No

36. Menurut Anda, seberapa pentingkah teknologi mesin pencarian informasi untuk bahasa daerah?
36. In your opinion, how important is information search engine technology for regional languages?

◯ Sangat penting 556 (64.6%)
Very important

◯ Penting 250 (29.0%)
Important

◯ Tidak terlalu penting 49 (5.7%)
Not very important

◯ Tidak penting 6 (0.7%)
Not important

Asisten Digital
Digital Assistant

37. Asisten digital adalah perangkat lunak berbasis kecerdasan buatan yang membantu pengguna menyelesaikan tugas sehari-hari melalui perintah suara atau teks, seperti menjawab pertanyaan, mengatur jadwal, dan mengontrol perangkat pintar. Contohnya adalah: ChatBot, Siri, Alexa, dan Google Assistant.

Apakah Anda pernah menggunakan aplikasi seperti ini?

37. A digital assistant is artificial intelligence-based software that helps users complete everyday tasks through voice or text commands, such as answering questions, setting schedules, and controlling smart devices. Examples are: ChatBot, Siri, Alexa, and Google Assistant. Have you ever used an application like this?

◯ Ya 679 (78.9%)
Yes

◯ Tidak 182 (21.1%)
No

38. Seberapa pentingkah asisten digital bahasa daerah untuk kebutuhan Anda?
38. How important is a regional language digital assistant for your needs?

◯ Sangat penting 286 (33.2%)
Very important

◯ Penting 330 (38.3%)
Important

◯ Tidak terlalu penting 201 (23.3%)
Not very important

◯ Tidak penting 44 (5.1%)
Not important

39. Untuk keperluan apa Anda ingin menggunakan asisten digital yang mendukung bahasa daerah Anda?
39. For what purposes would you want to use a digital assistant that supports your regional language?

☐ Konsultasi kesehatan 188 (21.8%)
Health consultation

☐ Curhat masalah pribadi 150 (17.4%)
Sharing personal problems

☐ Hiburan 316 (36.7%)
Entertainment

☐ Membantu belajar / pendidikan 514 (59.7%)
Help with learning/education

☐ Mencari informasi 604 (70.2%)
Searching for information

☐ Menuliskan teks seperti surat 263 (30.5%)
Writing text like a letter

☐ Memperbaiki penulisan teks 346 (40.2%)
Correcting text writing

☐ Tidak perlu 76 (8.8%)
Not necessary

☐ Lainnya 24 (2.8%)
Other

40. Asisten digital juga bisa membaca gambar dan video. Apakah menurut Anda penting memiliki Asisten digital berbahasa daerah yang bisa memahami gambar dan video yang berkaitan dengan budaya Anda?
40. A digital assistant can also read images and videos. Do you think it is important to have a regional language digital assistant that can understand images and videos related to your culture?

◯ Sangat penting 352 (40.9%)
Very important

◯ Penting 379 (44.0%)
Important

◯ Tidak terlalu penting 108 (12.5%)
Not very important

◯ Tidak penting 22 (2.6%)
Not important

Privasi dan Kredibilitas
Privacy and Credibility

41. Untuk mengembangkan teknologi bahasa daerah, diperlukan banyak data teks dan audio digital dalam bahasa tersebut. Sebagai contoh, peneliti mungkin akan mengumpulkan dan menganalisis data teks dan audio yang tersedia secara publik di media sosial Anda yang menggunakan bahasa daerah. Apakah hal ini membuat Anda merasa terganggu?
41. To develop regional language technology, a lot of digital text and audio data in that language is needed. For example, researchers might collect and analyze publicly available text and audio data on your social media that uses your regional language. Does this bother you?

◯ Saya merasa terganggu jika data teks tersebut digunakan untuk pengembangan teknologi bahasa daerah 30 (3.5%)
I feel disturbed if the text data is used for the development of regional language technology

◯ Saya merasa terganggu jika data audio tersebut digunakan untuk pengembangan teknologi bahasa daerah 29 (3.4%)
I feel disturbed if the audio data is used for the development of regional language technology

◯ Saya merasa terganggu jika data teks dan audio tersebut digunakan untuk pengembangan

teknologi bahasa daerah 36 (4.2%)
I feel disturbed if the text and audio data are used for the development of regional language technology

◯ Saya tidak merasa terganggu karena data tersebut tersedia secara publik 766 (89.0%)
I do not feel disturbed because the data is publicly available

42. Apakah Anda merasa teknologi kecerdasan buatan yang sudah ada memberikan perlindungan terhadap data pribadi Anda secara memadai?
42. Do you feel that existing artificial intelligence technologies provide adequate protection for your personal data?

◯ Ya 214 (24.9%)
Yes

◯ Tidak 379 (44.0%)
No

◯ Tidak tahu 268 (31.1%)
Do not know

43. Saat menggunakan teknologi bahasa seperti Google Search, Siri, dan Google Assistant, apakah Anda sudah pernah mendengar tentang isu privasi dan keamanan? Misalnya, tidak menyebutkan atau menuliskan data pribadi ke asisten digital seperti ChatGPT?
43. When using language technologies such as Google Search, Siri, and Google Assistant, have you heard about privacy and security issues? For example, not mentioning or writing personal data to digital assistants such as ChatGPT?

◯ Sangat tahu 140 (16.3%)
Very aware

◯ Cukup tahu 354 (41.1%)
Aware

◯ Tidak terlalu tahu 216 (25.1%)
Not too aware

◯ Tidak tahu 151 (17.5%)
Not aware

44. Apakah Anda pernah menanyakan masalah kesehatan kepada asisten digital seperti ChatGPT?
44. Have you ever asked a digital assistant such as ChatGPT about health problems?

◯ Pernah 278 (32.3%)
I have

◯ Tidak pernah 583 (67.7%)
I have not

45. Seberapa sering Anda melakukan verifikasi kebeneran informasi yang diberikan oleh teknologi bahasa seperti ChatGPT?
45. How often do you verify the accuracy of information provided by language technology such as ChatGPT?

◯ Selalu 130 (15.1%)
Always

◯ Sering 262 (30.4%)
Often

◯ Jarang 274 (31.8%)
Seldom

◯ Tidak pernah 195 (22.6%)
Never

46. Apakah Anda tahu bahwa informasi yang diberikan oleh asisten digital seperti ChatGPT tidak selalu benar dan bisa sepenuhnya salah?
46. Do you know that information provided by digital assistants such as ChatGPT is not always correct and can be completely wrong?

◯ Sangat tahu 311 (36.1%)
Very aware

◯ Cukup tahu 323 (37.5%)
Aware

◯ Tidak terlalu tahu 109 (12.7%)
Not too aware

◯ Tidak tahu 118 (13.7%)
Not aware

47. Pilihlah opsi jawaban Stroberi
47. Choose the Strawberry answer option

◯ Apel* 10 (1.2%)
Apple

◯ Pisang* 4 (0.5%)
Banana

◯ Jeruk* 4 (0.5%)
Orange

◯ Stroberi 832 (96.6%)
Strawberry

◯ Semangka* 11 (1.3%)
Watermelon

*note: *we omit the responses from analysis*

48. Saat menggunakan teknologi bahasa, apakah Anda sudah pernah mendengar tentang isu bias? Misalnya:

(1) Bias terhadap gender: komputer mengasumsikan bahwa dokter adalah laki-laki dan perawat adalah perempuan. Padahal terdapat dokter perempuan dan perawat laki-laki. (2) Bias terhadap agama/politik: komputer mencerminkan prasangka terhadap agama/politik tertentu sehingga menyudutkan kalangan tertentu.

48. When using language technology, have you ever heard of bias issues? For example: (1) Gender bias: computers assume that doctors are male and nurses are female. In fact, there are female doctors and male nurses. (2) Bias against religion/politics: computers reflect prejudice against certain religions/politics, thus cornering certain groups.

- ◯ Sangat tahu 138 (16.0%)
  Very aware
- ◯ Cukup tahu 335 (38.9%)
  Aware
- ◯ Tidak terlalu tahu 216 (25.1%)
  Not too aware
- ◯ Tidak tahu 172 (20.0%)
  Not aware

49. Tulis isu lain yang ingin Anda sampaikan terkait teknologi bahasa seperti ChatBot, asisten digital, mesin penerjemah dll.

49. Write other issues that you want to convey regarding language technology such as ChatBot, digital assistants, machine translators, etc.

861 write-in answers

Privasi dan Kredibilitas

Privacy and Credibility

50. Secara umum, bagaimana antusiasme Anda terhadap pengembangan teknologi bahasa untuk bahasa daerah Anda? Apakah Anda memiliki kekhawatiran atau ketidaksukaan terkait pengembangannya?

50. In general, how enthusiastic are you about the development of language technology for your regional language? Do you have any concerns or dislikes regarding its development?

- ◯ Saya antusias dan tidak khawatir 512 (59.5%)
  I am enthusiastic and not worried
- ◯ Saya antusias dan sedikit khawatir 287 (33.3%)
  I am enthusiastic and a little worried
- ◯ Saya tidak antusias, namun sedikit khawatir 26 (3.0%)
  I am not enthusiastic, but a little worried
- ◯ Saya tidak antusias dan tidak khawatir 36 (4.2%)
  I am neither enthusiastic nor worried

B Details of Variations of Importance Scores

Table 2 presents the importance scores across various categories. The symbol (*) denotes the total number of respondents for each language technology (LT): 753 for MT, 623 for STT, 589 for TTS, 612 for GC, 800 for IR, and 642 for DA. Meanwhile, (**) represents the corresponding numbers for another subset of respondents: 58 for MT, 188 for STT, 222 for TTS, 199 for GC, 11 for IR, and 169 for DA.

C The Division of West and East Indonesia based on Wikipedia

We aggregated the results based on several criteria, including clustering Indonesia into West and East regions. we referred to relevant Wikipedia pages[7] for a straightforward classification of provinces. Table 3 presents the distribution between West and East Indonesia, followed by respondent count for each province.

D Language Level Aggregation

Eberhard et al. (2023) established a language taxonomy based on real-world usage. The taxonomy consists of nine language status levels, ranging from International to Extinct language[8]:

- 0. International: The language is widely used between nations in trade, knowledge exchange, and international policy. *Not applicable in our survey*
- 1. National: The language is used in education, work, mass media, and government at the national level. *Not applicable in our survey*

[7] https://id.wikipedia.org/wiki/Indonesia_Barat, https://id.wikipedia.org/wiki/Indonesia_Timur

[8] https://www.ethnologue.com/methodology/#language-status

Categories	#respondents	MT	STT	TTS	GC	IR	DA
full	811	0.771	0.678	0.684	0.696	0.860	0.664
aware of bias	448	0.766	0.692	0.699	0.709	0.868	0.678
not aware of bias	363	0.777	0.661	0.664	0.681	0.851	0.646
aware of privacy	467	0.759	0.673	0.682	0.695	0.864	0.662
not aware of privacy	344	0.786	0.685	0.685	0.700	0.855	0.667
geo: west Indonesia	574	0.762	0.675	0.661	0.675	0.848	0.634
geo: east Indonesia	237	0.792	0.729	0.737	0.748	0.889	0.737
edu: high school	134	0.721	0.664	0.677	0.694	0.878	0.679
edu: undergraduate	389	0.792	0.688	0.687	0.700	0.868	0.674
edu: graduate	288	0.765	0.671	0.682	0.693	0.841	0.644
lang: stable	566	0.763	0.663	0.668	0.684	0.843	0.642
lang: endangered	196	0.804	0.731	0.723	0.740	0.896	0.723
lang: moribund	17	0.608	0.490	0.510	0.451	0.863	0.569
familiar to LT	*	0.775	0.714	0.733	0.724	0.864	0.705
∼familiar to LT	**	0.713	0.560	0.551	0.606	0.576	0.509
gen z	271	0.763	0.669	0.685	0.708	0.878	0.685
gen millennial	462	0.773	0.689	0.685	0.685	0.855	0.658
gen x boomer	78	0.782	0.658	0.671	0.722	0.829	0.641

Table 2: Importance scores across demographic and awareness categories.

West Indonesia	**East Indonesia**
East Java (112)	South Sulawesi (67)
West Java (111)	NTB (37)
Central Java (72)	NTT (34)
West Sumatera (54)	Bali (32)
Aceh (37)	Central Sulawesi (32)
North Sumatera (33)	S.E. Sulawesi (14)
DI Yogyakarta (29)	Papua (8)
Jakarta (29)	North Sulawesi (3)
Riau (18)	West Sulawesi (3)
Jambi (17)	Highland Papua (3)
West Kalimantan (13)	Gorontalo (1)
South Sumatera (12)	West Papua (1)
Lampung (6)	Central Papua (1)
Bengkulu (6)	Maluku (1)
South Kalimantan (6)	S.W. Papua (0)
Banten (5)	South Papua (0)
East Kalimantan (4)	North Maluku (0)
Ctrl. Kalimantan (4)	
Riau Islands (3)	
Bangka Belitung (2)	
North Kalimantan (1)	
Total=574	**Total**=237

Table 3: The division and the valid respondent count based on province location (West & East Indonesia.)

- 2. Provincial: The language is used in education, work, mass media, and government within major administrative subdivisions of a nation. *Not applicable in our survey*

- 3. Wider Communication: The language is used in work and mass media without official status to transcend language differences across a region.

- 4. Educational: The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.

- 5. Developing: The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.

- 6a. Vigorous: The language is used for face-to-face communication by all generations and the situation is sustainable.

- 6b. Threatened: The language is used for face-to-face communication within all generations, but it is losing users.

- 7. Shifting: The child-bearing generation can use the language among themselves, but it is not being transmitted to children.

- 8a. Moribund: The only remaining active users of the language are members of the grandparent generation and older.

- 8b. Nearly Extinct: The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.

- 9. Dormant: The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.

- 10. Extinct: The language is no longer used, and no one retains a sense of ethnic identity associated with the language. *Not applicable in our survey*

However, for ease of analysis, we consolidated these 13 levels into 3 broader categories. Table 4 presents our classification along with the languages covered in the survey.

E Dialect-Based User Preferences

As discussed in Section 4.2, dialects also influence how speakers of the same language perceive the need for language technologies (LTs). Due to limited respondent counts, we focused on five languages and their respective dialects: Aceh (Aceh Besar and Banda Aceh dialects), Buginese (Makassar, Bone, and Bugis Kayowa dialects), Javanese (Arekan, Pandhalungan, and Mataraman dialects), Minangkabau (Agam and Payakumbuh dialects), and Sundanese (Bandung Priangan and Sumedang dialects) as shown in Figure 11.

Overall, the Banda Aceh, Payakumbuh, and Bandung Priangan dialects stand out as perceiving LTs as more important compared to other dialects within their respective languages. Notably, the Bone dialect in Buginese shows a distinct preference, with speakers prioritizing GC and IR more but showing less interest in MT. In contrast, the Makassar dialect perceives LTs as less important than other Buginese dialects.

However, the reasons behind these trends remain unclear. To fully understand why certain dialects exhibit unique patterns in perceiving LTs, direct dialogue with speakers of each dialect is essential.

F How Awareness of Privacy Affects Use Rate

Figure 8 illustrates the relationship between respondents' awareness of privacy concerns and their usage rates of language technologies (LTs). Overall, individuals who believe that LTs fail to provide sufficient protection for personal data are less likely to use digital assistants for health-related inquiries, as such information is considered highly sensitive. Similarly, those who remain uncertain about the level of data protection offered by LTs tend to avoid using these technologies for health-related questions altogether.

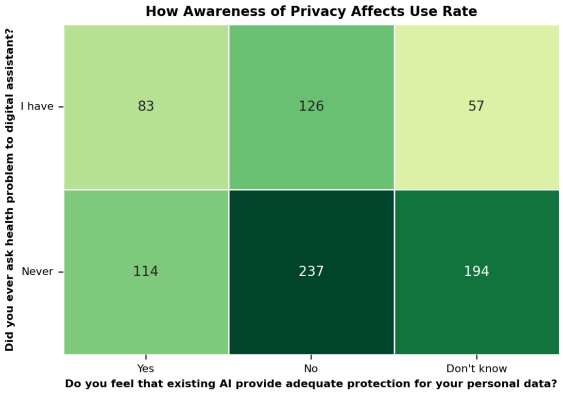

Figure 8: How awareness of privacy affects use rate.

G Familiarity with LTs: Categorized on Generation, Language Level, and Geography

Figure 9 illustrates respondents' familiarity with LTs analyzed in this survey, categorized by different factors. Among generations, Gen Z appears to be the most familiar with LTs, while Gen X & Boomers show the lowest familiarity, likely due to the rapid pace of globalization affecting younger generations more. Additionally, speakers of stable languages tend to have higher LT familiarity compared to others. Geographically, respondents from West Indonesia are more familiar with LTs than those from East Indonesia, likely due to Indonesia's development being concentrated in more populous islands such as Java and Sumatra. In addition, Figure 10 shows the importance scores of the respondents who are familiar with the LT across several categories, followed by the Pearson correlation between the familiarity of LT to its importance score.

H Important Score vs Available Resource on Wikipedia

We use Wikipedia data as a common text source for dataset collection. Figure 12 illustrates that despite the high importance scores of several Indonesian local languages, the available resources remain insufficient. Only a few languages—such as Javanese, Sundanese, Balinese, and Minangkabau—have datasets exceeding 10MB (which is still considered tiny). Meanwhile, resources for all

Language Level	Covered Languages
Stable Language (*Ethnologue language level 3-5*)	Javanese (245), Sunda (105), Bugis (64), Minangkabau (62), Bali (30), Kaili Ledo (13), Musi (9), Madura (7), Banjar (6), Toraja-sadan (6), Lamaholot (4), Malay-manado (3), Ngaju (3), Chinese-mandarin (3), Mandar (2), Kendayan (1), Moma (1), Nias (1), Malay-kupang (1)
Threatened Language (*Ethnologue language level 6a-6b*)	Aceh (33), Sasak (22), Malay (20), Malay-jambi (13), Batak simalungun (12), Batak toba (7), Hawu (7), Saluan (6), Bima (5), Lampung nyo (4), Sumbawa (4), Tolaki (4), Malay-central (4), Tetun (4), Uab meto (3), Manggarai (3), Biak (3), Muna (3), Kambera (3), Tukang besi south (2), Li'o (2), Batak karo (2), Moronene (2), Pamona (2), Konjo-coastal (2), Osing (2), Padoe (1), Bahau (1), Sika (1), Betawi (1), Batak mandailing (1), Ende (1), Batak alas-kluet (1), Gayo (1), Bangka (1), Malay-tenggarong kutai (1), Bakati' (1), Tii (1), Gorontalo (1), Sentani (1), Nalca (1), Ekari (1), Ketengban (1), Ansus (1), Diuwe (1), Rejang (1), Mamuju (1), Cia-cia (1)
Moribund Language (*Ethnologue language level 7-9*)	Hakka (12), Banggai (3), Andio (2)

Table 4: Language level classification and the valid respondent count based on each language.

other languages remain limited or entirely unsupported.

I Current State of Language Technologies for Indonesian Local Languages

Table 5 presents the current state of LTs for Indonesian local languages, using Google as a benchmark. While some languages, such as Javanese and Sundanese, are supported in certain LTs, many other underrepresented languages still lack coverage. Additionally, technologies like TTS and DA have yet to support any Indonesian regional languages. This provides an overview of the development gaps in LTs for these languages.

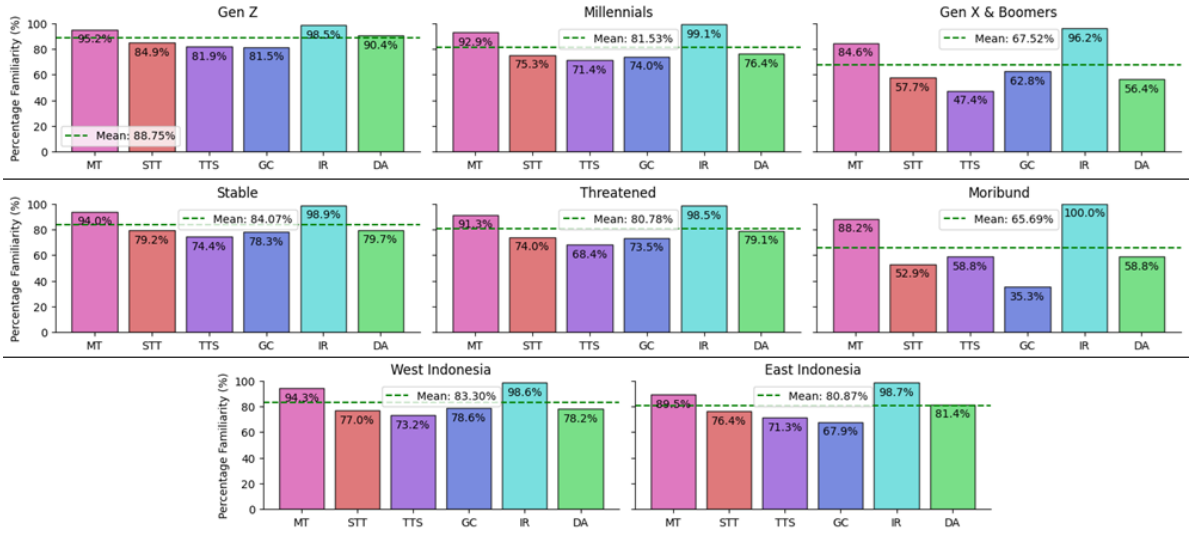

Figure 9: Familiarity with LTs by multiple categories. The top row categorizes data by generation (Gen Z, Millennials, Gen X & Boomers), the middle row by language endangerment level, and the bottom row by Indonesian region (West and East Indonesia).

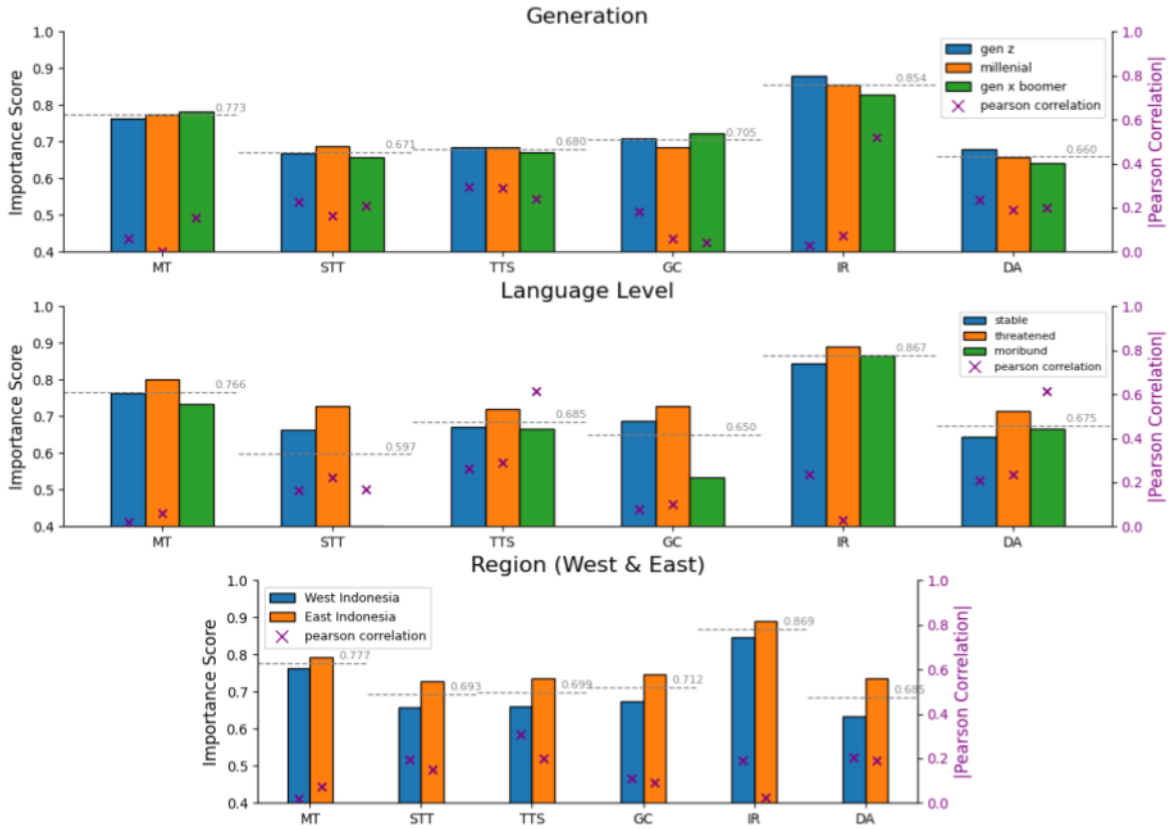

Figure 10: Importance scores of the respondents that are familiar with the LT across several categories: Generation, language level, and region (West & East Indonesia.), alongside their Pearson correlation.

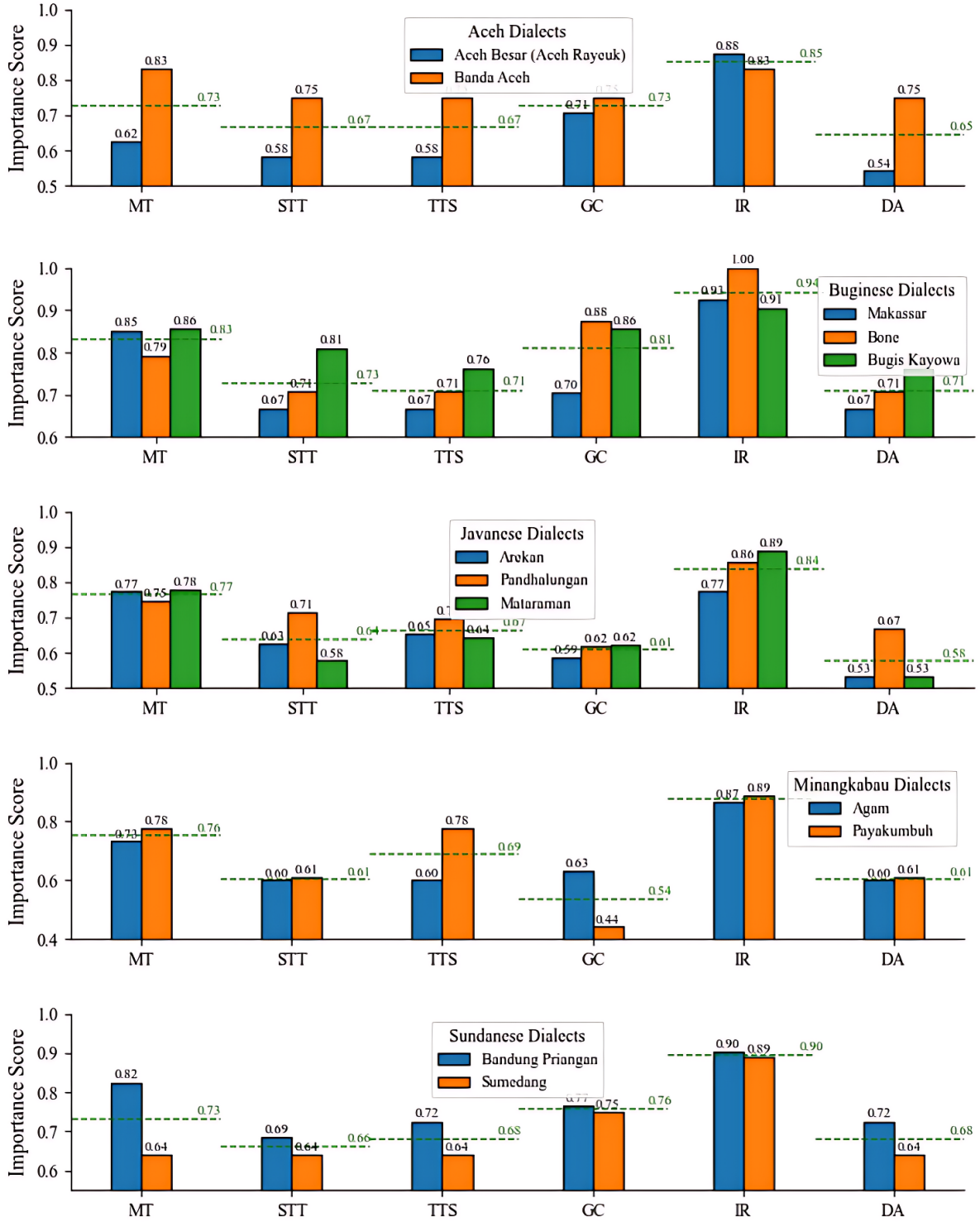

Figure 11: Differences in LT preferences across Aceh, Buginese, Javanese, Minangkabau, and Sundanese dialects (from top to bottom).

LT	Importance score	Local Indonesian Language(s) supported by Google
MT	0.771	Javanese (jav), Sundanese (sun), Minangkabau (min), Acehnese (ace), Balinese (ban), Batak Karo (btx), Batak Simalungun (bts), Batak Toba (bbc), Betawi (bew), Makassar Malay (mfp)
STT	0.678	Javanese (jav), Sundanese (sun)
TTS	0.684	*not supported (only available in Indonesian (id))*
GC	0.696	Ambonese Malay (abs), Batak Simalungun (bts), Buginese (bug), Duri (mvp), Hawu (hvn), Makassar Malay (mfp), Toraja-sa'dan (sda), Acehnese (ace), Batak Alas-kluet (btz), Balinese (ban)*, Banjar (bjn), Batak Mandailing (btm), Batak Toba (bbc), Betawi (bew), Gorontalo (gor), Jambi Malay (jax), Javanese (jav)*, Kutai Malay (vkt), Ledo Kaili (lew), Manado Malay (xmm), Mandar (mdr), Minangkabau (min), Mongondow (mog), Papuan Malay (pmy), Sasak (sas), Sundanese (sun)
IR	0.860	Javanese (jav)**
DA	0.664	*not supported****

Table 5: Importance score for each LT and its availability in local Indonesian languages supported by Google. The *italic* importance score only considers the 'very important' option. *their script alphabets are also supported **only able to extract entities from document ***Google Assistant (Android handphone & TV) & Gemini only available in Indonesian (ind) language.

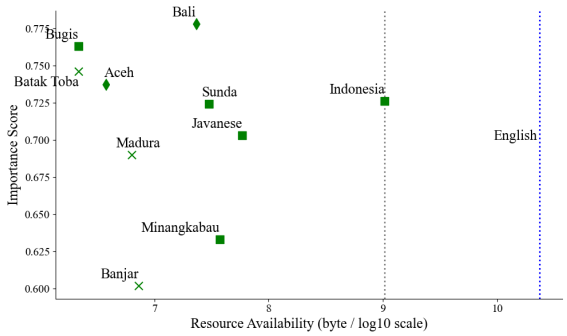

Figure 12: Importance scores and available resources for each supported local Indonesian language on Wikipedia. ■ represents languages that has more than 50 respondents, ♦ 30-50 respondents, and × is less than 30 respondents.