

印尼人真正需要的语言技术是什么？ 全国范围的调查

Muhammad Dehan Al Kautsar¹, Lucky Susanto², Derry Wijaya², Fajri Koto¹

¹MBZUAI ²Monash University

muhammad.dehan@mbzuai.ac.ae

Abstract

目前正在努力开发针对印度尼西亚 700 多种地方语言的自然语言处理技术，但由于需要与当地语言的使用者直接互动，进展仍然耗资巨大。然而，这些语言社群真正需要的语言技术仍不明确。为了解决这一问题，我们进行了全国性的调查，以评估印度尼西亚语言使用者的实际需求。我们的调查结果表明，解决语言障碍——特别是通过机器翻译和信息检索——是最重要的优先事项。虽然对语言技术进步有强烈的热情，但关于隐私、偏见以及将公共数据用于人工智能训练的担忧，突显了需要更大的透明度和清晰的沟通，以支持人工智能的广泛应用。

1 介绍

印度尼西亚拥有超过 2.8 亿人口，分布在 17,508 个岛屿上，除了其国家语言印尼语 (Bahasa Indonesia) 外，还拥有 700 多种地区语言。这种语言的多样性为自然语言处理 (NLP) 提供了机会，但也带来了挑战，比如数据匮乏和语言标准化问题。

为了应对这些挑战，近年来已经进行了大量努力来推动印度尼西亚自然语言处理的发展，包括多语言语料库开发 (Cahyawijaya et al., 2023a; Lovenia et al., 2024)、情感分析 (Winata et al., 2023)、对话 (Purwarianti et al., 2025) 以及自然语言理解/生成 (Koto et al., 2020; Cahyawijaya et al., 2023b)。然而，这些发展的成本仍然很高且劳动密集。更重要的是，这些努力是否符合实际用户需求仍然不确定，这引发了一个关键问题：印度尼西亚人真正需要语言技术 (LT) 提供什么？回答这个问题至关重要，因为为印度尼西亚构建语言技术特别复杂，部分原因是人口多样化以及用户偏好各异。因此，参与性设计与与社区的互动至关重要，以确保这些技术能够满足实际需求 (Mager et al., 2023; Kolhatkar and Verma, 2023; Cooper et al., 2024)。

为了回答这些问题并探索这些挑战，我们通过问卷进行了一项全国性调查，评估印尼人优先考虑哪些语言技术 (LTs)。我们收集了人口

统计数据，并要求受访者对六种语言技术进行评分：机器翻译 (MT)、语音转文本 (STT)、文本转语音 (TTS)、语法检查器 (GC)、信息检索 (IR) 以及数字助手 (DA)。我们还考察了对人工智能的态度，包括隐私、可信度和数据使用的关注。在两个月的时间内，我们从 70 种不同印尼语言的受访者中收集了 861 份回馈，涵盖了 38 个省中的 35 个 (图 1)。

虽然在全球北方已经进行了类似的调查 (Blaschke et al., 2024; Lent et al., 2022a; Soria et al., 2018)，但我们的研究揭示了关于印尼语言社区的需要和关注的独特见解。关键点包括：

- 诸如 IR 和 MT 之类的跨越语言障碍的 LTs 是非常需要的。
- 方言也影响用户的兴趣，表明偏好不仅仅由语言本身决定。
- 92.6 % 的印度尼西亚人对人工智能技术感到兴奋，但 36.3 % 表示担忧。
- 86.68 % 意识到像 DA 这样的 LTs 中可能存在的故障，但只有 46.24 % 定期验证所提供的信息。
- 接触 LTs 会影响用户兴趣，但对于某些群体，如 Z 世代和使用稳定语言者，这一结论并不成立。

2 背景与相关工作

随着对语言技术 (LTs) 的需求增长，NLP 的发展正在加速 (Abdalla et al., 2023)。然而，这一进展在全球范围内并不均衡。在印度尼西亚，由于资源有限、语言多样性、方言和风格差异、拼写不一致以及全岛技术和教育获取不平等社会障碍，NLP 的开发和采用面临着重大挑战 (Aji et al., 2022)。此外，随着人工智能技术的发展，有关隐私、数据收集和信任的问题给开发工作增加了更多的复杂性。

2.1 全球范围内的 LTs 调查

语言技术需求在不同地区之间差异显著，反映了本地语言、文化和技术需求。例如，对 327 名使用方言的德国人的调查发现，受访者优先

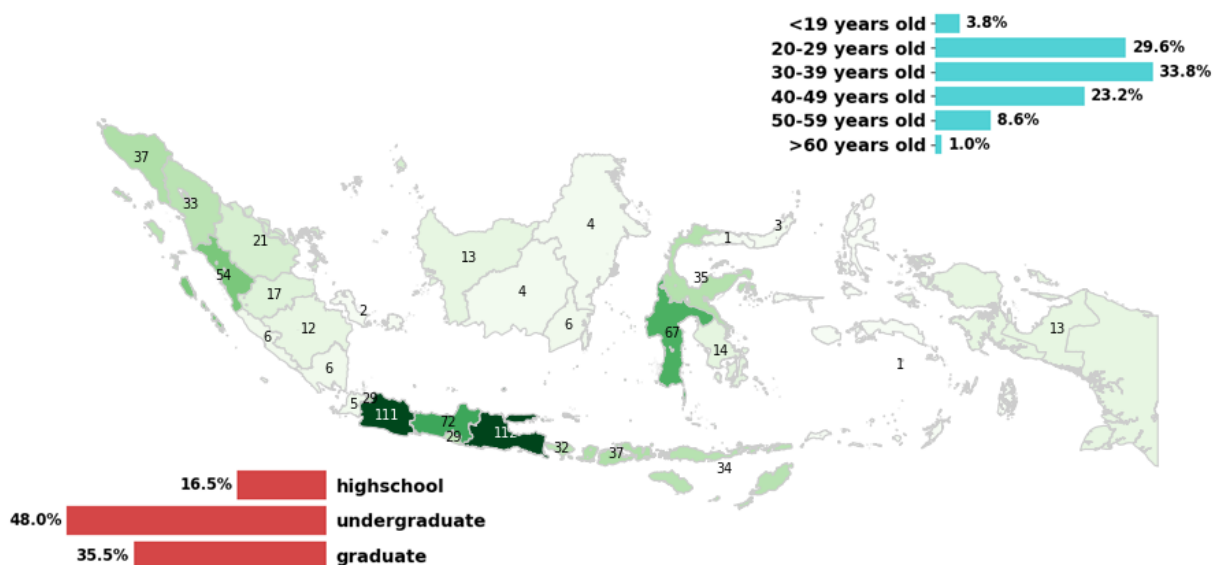


Figure 1: 按省份分布的受访者，以及当地印尼语使用者的年龄和最高教育水平。

考虑方言友好的数字助理，而不是机器翻译和拼写检查 (Blaschke et al., 2024)。与克里奥尔语专家和 37 名讲克里奥尔语社区的人的访谈强调了语音转录作为一个关键未满足的需求 (Lent et al., 2022b)。同时，对超过 1200 名说巴斯克语、布列塔尼语、卡累利阿语和撒丁语的人的大规模调查强调了对语言数字化的强烈渴望 (Soria et al., 2018)。这些例子强调了语言技术采用在全球范围内多样化和依靠具体情境化的性质。Millour (2019) 进行了一项关于欧洲非标准化语言，阿尔萨斯语的研究，通过设计一系列调查问题，收集了超过 1200 名参与者的回应，其中多数人说阿尔萨斯语以及另一种语言，比如法语或德语。虽然他们成功地识别了阿尔萨斯语现有语言技术的状态，但他们没有充分利用调查来捕捉受访者对可用语言技术的意见。同样地，ELE 项目¹、Mariani (2020) 和 Blasi et al. (2022) 也检查了不同语言和人口中的语言技术的现状和质量，但它们同样缺乏语言使用者观点的表现，导致其具体的语言技术需求仍然未知。

另一方面，之前关于伦理考虑的研究在探索为土著语言构建 NLP 技术的伦理考虑时也得出了相同的结论。他们建议 NLP 研究人员在构建 NLP 技术时，优先考虑社区参与，而不仅仅关注去语境化的人工制品。这与我们论文的目标一致，即了解整个印尼地区的语言技术需求类型——这是一个拥有众多土著文化和语言的巨大而多样化的国家。

¹<https://european-language-equality.eu/deliverables/>

2.2 印度尼西亚语言技术开发中的挑战

印度尼西亚的语言技术发展面临多重挑战 (Aji et al., 2022)。一个主要问题是资源的缺乏和对未被充分代表的语言和方言所面临困难的有限认识，例如标准化问题 (Novitasari et al., 2020)。然而，最大障碍仍然是足够数据的可用性。

尽管面临持续的挑战，研究人员和社区为开发多语言语料库 (Cahyawijaya et al., 2023a; Lovenia et al., 2024) 做出了显著努力，增加了数据集的可用性和可见性。然而，这些语料库仍然主要由印度尼西亚文本主导，只有一小部分代表本地语言。虽然一些数据集强调深度 (规模) (Komariah et al., 2024; Nurul Afra, 2024; Yuyun et al., 2024)，而另一些则优先考虑广度 (语言覆盖) (Costa-jussà et al., 2022; Winata et al., 2023)，但数据不平衡依然存在。在机器翻译中，全球 23 亿个平行句子中只有 1.1 % 涉及英语-印度尼西亚语对，只有 0.06 % 涉及爪哇语-英语 (Gowda et al., 2021)。

数据的有限性直接影响着 LT 的性能，研究显示 LLM 在印度尼西亚的能力存在显著差异。Koto et al. (2023) 发现 GPT-3.5 在印尼语中甚至难以处理小学水平的问题，并且在像巽他语这样的地区语言中表现更差。这些数据稀缺和语言偏差的挑战阻碍了 LT 在印度尼西亚的实际应用和商业可行性。鉴于这些限制，为所有印度尼西亚语言开发 LT 既昂贵又复杂，这突显出在投资大规模 LT 开发之前，首先需要了解实际用户需求的必要性。

2.3 隐私和偏见问题，以及对 LTs 的信任

对开发语言技术 (LTs) 的数据需求增加，加剧了人们对隐私的担忧，这一问题早在大型语言模型 (LLMs) 出现之前就已存在。这种担忧在 [European Parliament and Council of the European Union \(2016\)](#) 和 [California State Legislature \(2018\)](#) 等法规框架的implement中表现得很明显。

尽管有监管努力，隐私问题仍然存在，因为研究表明，即使是匿名化的数据集也可能易于重新识别 ([Rocher et al., 2019](#))。这导致了对人工智能日益增长的怀疑，尤其是在西方国家，那里的只有 37% 的美国人认为人工智能提供的好处多于弊端 ([Stanford University, 2024](#))。相比之下，印度尼西亚的态度显然更为积极，78% 的印度尼西亚人认为人工智能是有益的。这种乐观态度可能受到人工智能曝光、公共讨论和监管重点差异的影响，因为关于人工智能伦理和治理的讨论不如西方国家那么突出。为了更好地了解印度尼西亚的公共讨论，特别是关于当地语言的语言技术，我们的调查包括关于与人工智能和语言技术采用相关的看法、优先事项和关注的问题。

3 问卷与数据处理

3.1 问卷

部分受到 [Blaschke et al. \(2024\)](#) 的启发，我们的问卷分为六个部分：介绍、区域语言详细信息、对地方语言的看法、与 LTs 相关的问题、LTs 的隐私和可信度，以及受访者对 AI 的兴趣。完整的问题集合详见附录 ??。每位参与者完成问卷最多需要 20 分钟。

我们使用 Google Forms² 分发了我们的问卷，并通过作者的专业网络分享，触及到了语言教师、印度尼西亚大学的相关利益者、记者以及本地语言大使和社区。这种方法使我们能够收集到来自整个群岛的回复，涵盖了 38 个省中的 35 个。在从 2024 年 10 月 6 日到 2024 年 12 月 5 日为期两个月的时间内，我们的问卷共获得了 861 名受访者。最后，作为一种感谢的表示，我们在问卷结束时随机给 10 名受访者总共 3,000,000 印尼盾的奖励。

为了确保每个回答的有效性，我们要求每位受访者分享他们的电子邮件地址或有效的电话号码，这些信息将用于后续的奖励选择。此外，我们的问卷还包含三个验证问题，要求受访者执行简单的加法或选择一个特定选项。这些验证问题会随机嵌入在问卷中，要求受访者在回复前仔细阅读每个问题。这些简单的验证任务有助于检测不认真的回答并防止机器人生

成或随机提交的回答，这是一种在大型调查中常用的方法。在移除未正确回答验证问题的回答后，我们总共获得了 811 份有效回答，这些回答用于本研究。

我们通过根据 [Eberhard et al. \(2023\)](#) 考虑受访者的语言濒危等级来丰富调查问卷的回复。我们将他们的数据库归纳为一个三层体系：稳定、受威胁和濒死；这使我们能够深入了解语言活力如何影响受访者的 LT 需求。更多细节见附录 C。

总共有 811 份有效问卷来自印度尼西亚 38 个省中的 35 个，涵盖了印度尼西亚 700 多种语言中的 70 种。52.6% 的受访者自认为是女性，几乎所有参与者都在日常生活中经常使用技术（电脑/笔记本电脑/智能手机），这在与语言技术相关的问题中特别重要。我们根据人口统计类别和语言濒危程度对问卷结果进行了汇总。地域上，我们收集了来自印度尼西亚西部的 574 份问卷，以及来自印度尼西亚东部的 237 份问卷，这一区域划分按照附录 B 中的说明进行。在代际方面，271 名受访者属于 Z 世代，462 名属于千禧一代，78 名属于 X 世代或更老。³

最后，基于我们在附录 C 中的汇总，受访者根据语言濒危程度被分类：566 人为稳定语言使用者，196 人为受威胁语言使用者，17 人为濒死语言使用者，还有 32 人因其所使用的语言不在 [Eberhard et al. \(2023\)](#) 所列的印度尼西亚地方语言中，而被归类为未知。

我们引入了术语“重要性评分”（图 2），这帮助我们在第 4 节根据受访者的意见量化每个 LT 的重要性。受访者基于一个四级 Likert 量表评估每个 LT 的重要性：“非常重要”、“重要”、“不太重要”和“不重要”。重要性评分是这些响应加权值的归一化，其中“非常重要”的评分为 3，逐渐减少到“不重要”，其评分为 0。

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

Figure 2: 重要性评分 (IS) 的计算方法，取值范围在 [0, 1] 之内。

机器翻译特定评分 我们将受访者对机器翻译 (MT) 重要性的观点分为三个类别：非常重要、重要和不重要，以便与其他语言技术进行比较。在 MT 重要性部分，受访者有六个回答选择——其中五个代表 MT 可能重要的不同方式，一个表示 MT 不重要（参见附录 ?? 问

³Z 世代包括 1997 年至 2010 年出生的人，千禧一代包括 1981 年至 1996 年出生的人，X 世代或更早世代指的是 1980 年之前出生的个体。

²<https://docs.google.com/forms>

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 %	4.48 %	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: 语言技术 (LTs) 重要性分数相对于整体回应在不同人口统计和认知类别中的百分比变化。蓝色表示相比整体回应, 受访者给予更高的重要性分数, 而红色表示较低的分数。如表所示, 对于开发印尼地方语言的语言技术的乐观态度主要来自东印尼的受访者、濒危语言的使用者, 以及那些熟悉语言技术的人。*753, 623, 589, 612, 800, 642 分别对应于机器翻译、语音转文字、文字转语音、语音合成、信息检索、数据分析。**58, 188, 222, 199, 11, 169 分别对应于机器翻译、语音转文字、文字转语音、语音合成、信息检索、数据分析。

题 23)。我们将选择 3 到 5 个有关 MT 重要性选项且不选择不重要的受访者归类为“非常重要”。“重要”类别适用于选择 1 或 2 个相关选项且未选择不重要的受访者。最后, 选择不重要的受访者则归入相应类别。

4 结果

4.1 印尼人最需要哪些 LT ?

图 ?? 显示, 计算得出的重要性评分 (见第 ?? 节) 中, IR 以 0.860 的分数名列前茅, 突出了其在促进信息访问中的关键作用。相反, 由于 DA 暴露或在地区背景中的实际使用有限, DA 的分数最低, 为 0.664。同时, MT 在中间组中以 0.771 分领先, 接下来是 STT、TTS 和 GC。总体而言, IR 和 MT 的显著性突显了在印度尼西亚语言多样化环境中弥合语言障碍的重要性。(Aji et al., 2022)

4.1.1

关键类别的变化 表格 1 (附加细节见附录 A) 总结了在隐私和偏见意识、LT 熟悉度、地理位置、教育、语言濒危和世代定义的子组中重要性得分的差异。例如, 意识到隐私问题的受访者平均给 LT 需求打分低 0.42 % 分, 而意识到偏见的受访者平均给其打分高 1.41 % 分。东印尼的受访者对 DA 的偏好也比总体样本高出 10.09 %。一般而言, 他们对为他们的语言发展不同的 LT 持更加积极的态度, 相较于西印尼人。LT 熟悉度进一步支持他们本地语言的 LT

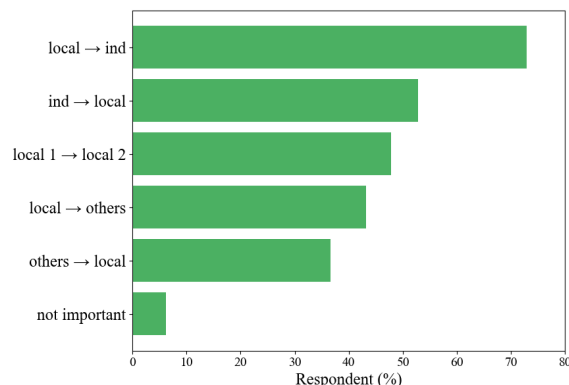


Figure 3: 受访者对机器翻译重要性的看法。local= 印度尼西亚区域语言, Ind= 印度尼西亚语, others= 外语。

发展。相似的积极模式也出现在濒危语言的使用者中, 尽管这一趋势在式微语言使用者中有所逆转。分析和讨论见章节 4.3 和 5。

如图 3 所示, 需求最多的翻译方向是从区域语言翻译成印尼语, 其次是相反方向。这种偏好在不同的人口统计学中保持一致, 突显了印尼语作为跨区域交流统一媒介的角色。

我们发现, 用户偏好的差异不仅仅基于人口统计类别, 还由于方言差异在同一种语言中产生。图 ?? 突出显示了三种爪哇方言使用者——Arekan、Pandhalungan 和 Mataraman——在 LT 偏好上的差异。结果表明, 相较于其他方言, Pandhalungan 方言的爪哇语使用者对 DA 表现出更强的偏好, 但对 MT 的兴趣较少。此外, Mataraman 方言的使用者优先考虑信息检索

IR。附录 ?? 中提供了其他语言中方言差异的详细分析，突出显示即使在同一种语言的使用者中，LT 偏好也可能存在显著差异。

我们的调查显示，92.6 % 的受访者对人工智能技术表示兴奋，反映出对技术进步普遍乐观的态度。然而，仅有 36.3 % 的受访者对人工智能技术的发展表示担忧，这明显低于 [Stanford University \(2024\)](#) 报告的 66 %。值得注意的是，对人工智能的担忧紧密关联于受访者对特定问题如隐私和偏见的认知。

我们在问卷中直接询问受访者关于他们对隐私问题的认知及其对此的看法（见附录 ??，问题 42 和 43）。对隐私问题的认知似乎强烈影响对人工智能的担忧。在认为当前人工智能技术不存在隐私问题的 197 名受访者中，仅有 53 人 (26.9 %) 表示担忧。相比之下，在认为存在隐私问题的 363 名受访者中，有 163 人 (44.9 %) 表示担忧。最后，在对隐私问题不了解的 251 名受访者中，有 79 人 (31.4 %) 表示担忧。这些发现表明，认识到隐私问题的个人更有可能对人工智能技术感到担忧，突显出隐私是影响公众认知的关键因素。

一个类似的趋势也出现在对 AI 技术的偏见上。与隐私问题相似，我们询问了受访者对语言技术偏见的认识，并在问卷中明确提供了偏见的例子（见附录 ??，第 48 题）。在 157 名不知道偏见问题的受访者中，只有 41 人 (26.1 %) 表达了对 AI 的担忧。相比之下，在 654 名知道偏见问题的受访者中，有 254 人 (38.8 %) 表达了担忧。这些结果表明，对偏见的认识提高了对 AI 潜在风险的识别，尽管相比于隐私问题，其对担忧的影响似乎较低。

4.2 印度尼西亚人对事实核查必要性的意识

图 4 展示了对 LT 幻觉的认知如何影响受访者核实信息倾向的趋势。根据我们的调查，86.68 % 的受访者意识到 LT，例如数字助手，可能存在缺陷并提供不正确或非事实性的信息。然而，尽管有这样高的认知水平，仅有 46.24 % 的受访者会定期核实 LT 提供的信息，这突显了受访者对 LT 生成信息不可靠性的认知和反应。

此外，仅考虑那些不经常验证来自 LT 的信息的受访者，我们发现 19.50 % 的人曾向 LT 询问与健康相关的问题，而这与 48.27 % 的受访者形成对比——这些受访者既询问过健康问题，又定期查证他们收到的信息。这表明，不定期验证信息的个体可能不太可能使用 LT 进行与事实相关的查询。此外，对数据隐私的担忧让个人在分享个人信息（比如健康状况）时更加谨慎，因为他们担心当前的 AI 系统可能

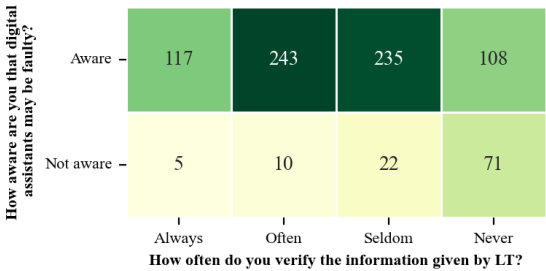


Figure 4: 关于对 LT 的幻觉的意识如何影响受访者的信任的热图。

无法充分保护他们的数据，详见附录 D。

4.3 之前接触 LT 是否会影响 LT 需求？

对某一特定 LT 接触较少甚至没有接触的受访者更有可能将其视为不重要。这一趋势在所有 LT 中都适用，除了机器翻译，无论熟悉与否，它始终被高度重视（图 5）。

此外，附录 E 探讨了受访者对特定语言技术（LT）的熟悉程度如何影响他们对该语言技术在其本地语言中发展的重视程度（以及他们的熟悉程度与这些感知重要性之间的相关性）。根据皮尔逊相关性分析（图 8，附录 E），某些群体——例如 X 世代/婴儿潮一代与他们对信息检索（IR）的熟悉程度和他们对 IR 重视程度之间表现出强烈的正相关。类似地，濒危语言使用者对 TTS 和 DA 的熟悉程度与其感知的重要性之间也表现出强烈的正相关。此外，TTS 和 DA 的熟悉程度和感知的重要性在不同人口类别之间一致表现出强烈的正相关。这表明了一种共同的行为模式，以及受访者对这些技术的熟悉程度与其感知重要性之间的显著关系。

然而，尽管年轻一代，如 Z 世代，以及稳定语言的使用者对语言技术有更高的熟悉度（图 7，附录 E），他们赋予语言技术的重要性评分并不总是在语言技术类别中最高。这表明虽然对语言技术的熟悉度会影响其重要性的感知，但并不总是决定其优先级。这些发现引发了关于其他驱动这些感知的潜在因素的有趣问题，而这些因素在本研究中仍未被探索。

5 讨论

有限的区域数据是 LT 发展的障碍 附录 ?? 显示，尽管受访者认为语言技术（LTs）非常重要，数据的可用性却对其发展构成了重大障碍，特别是对那些不被广泛使用的地区语言而言。例如，来自布吉社区（拥有 400 万使用者）⁴的受访者极力推动语言技术的发展。然而，现有的布吉语训练数据仅有不到 10 MB，这严重

⁴<https://www.ethnologue.com/language/bug>

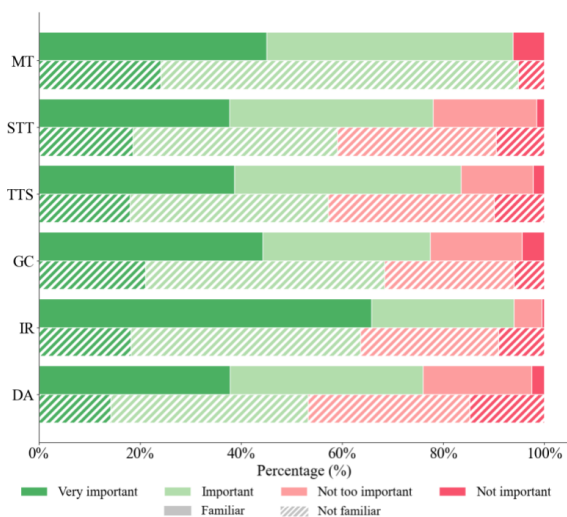


Figure 5: 受访者对不同语言技术重要性的看法按熟悉程度划分。

制约了技术进步。同样地，我们发现濒危语言的使用者平均而言对其语言的 LT 的发展更为兴奋（表 1）。不幸的是，濒危语言的数据也通常有限。正如图 10、中附录 ?? 所示，布吉语、托巴语和亚齐语等一些受访者最为期待的语言却是现有资源最少的语言之一。

此外，如附录 ?? 所示，现实应用领域的语言技术（LT）发展现状显示出一种不平衡。尽管像爪哇语这样的高资源语言正越来越多地被整合进入语言技术中，许多拥有大量使用者的低资源语言仍然得不到支持。这强调了推动印尼地方语言的语言技术的一个关键挑战——没有足够的数据，自然语言处理（NLP）应用的发展将受到限制。正如预期的那样，语言技术（LT）的偏好因地缘政治区域而异。与其他国家相比（见第 2.1 节），印尼人的语言技术优先事项似乎受到语言障碍的强烈影响，其中信息检索（IR）和机器翻译（MT）被认为是最有价值的。这与印尼的丰富语言多样性相符，虽然丰富了文化，但也对信息获取和沟通带来了挑战。在这种背景下，语言技术有潜力作为统一工具，将语言多样性从障碍转变为国家优势，这一观点也得到了 Aji et al. (2022) 等先前研究的支持。一个关键发现是，印尼人强烈希望搜索引擎能够支持地方语言。

我们的调查显示，11% 的受访者反对使用公共数据（包括文本和音频）来开发支持区域语言的语言技术。进一步分析表明，这一比例并未受到受访者对隐私或偏见问题的认识、他们对人工智能技术的热情或担忧，或是他们语言的濒危状态等因素的显著影响。这些发现表明，关于公共数据使用的顾虑可能源于我们研究中未考虑的变量之外的因素。需要进一步的调查来揭示印尼人对这些保留意见的深

层原因，可能包括文化敏感性、对机构的信任度、数据殖民主义的担忧 (Couldry and Mejias, 2019)，或是与数据滥用或数字劳动问题的特定经验 (Le Ludec et al., 2023)。

为什么濒危语言的使用者对语言技术不感兴趣
表 1 表明，与对语言技术表现出最大热情的濒危语言使用者不同，濒死语言的使用者对发展以其本地语言为基础的语言技术缺乏热情。我们假设这种态度源于他们对该语言现状的有限理解以及认为它不再是实用交流工具的看法。为了进一步探索这一点，我们采访了一位负责濒危和受威胁语言复兴的政府官员，作为语言群体拒绝印尼政府复兴努力的例子，他提到贝利尔语言。由于只有五对兄弟姐妹几乎不懂这种语言，他们不再认为它有实用性，主要使用更易于交流的语言，如 Kabola (klz)⁵⁶。这表明虽然语言技术能够支持语言复兴努力，但其影响可能仅限于仍被归类为濒危的语言。一旦语言达到濒死状态，获得社区支持进行复兴就会变得非常困难。这强调了在语言达到这一关键阶段之前进行专门研究和开发相关语言技术的紧迫需求。

6 结论

在这项研究中，我们调查了印度尼西亚 38 个省中的 35 个，共收集 800 多份回应，以评估公众对语言技术（LTs）的态度。我们的研究结果强调了一项强烈的国家优先事项，即通过信息检索（IR）和机器翻译（MT）等语言技术促进信息获取和区域间沟通。这些技术对于克服语言障碍和确保数字包容性至关重要。

此外，我们观察到受访者对人工智能技术表现出高水平的热情，但这也伴随着对隐私、偏见以及使用公共数据来训练语言技术的担忧。鉴于先前对语言技术的熟悉度与其重要性的更高认知相关，提高公众对语言技术的接触和教育有助于解决这些担忧，促进更大的信任和广泛采纳。

我们的分析和访谈也强调了在社区仍然参与的情况下开发语言技术和语言资源的紧迫性。等待太久有可能错失良机，因为语言一旦陷入濒危状态，往往会失去社区对复兴工作的支持。在区域语言达到这个关键阶段之前开发语言技术，对于确保它们在社会中的持续功能和保护印尼丰富的语言多样性至关重要。必须进行专门的研究，以防止这些语言不可挽回地消失，使得开发语言技术不仅是有益的，更是势在必行。

⁵Kabola 被 Eberhard et al. (2023) 列为濒危物种。

⁶更多细节，请参见 RRI 新闻

局限性 我们的研究结果代表了印尼人口的一个样本，其中大多数受访者是语言能力稳定的人群，千禧一代，西印尼居民，本科生，并且已经熟悉某些语言技术。使用在线平台也限制了那些无法接触此类技术的人的代表性。虽然这意味着我们的发现可能未能捕捉到每一个可能的视角，但回应远非一致。多样化的输入范围允许进行详细的分析，如第 4 节所示。此外，为了确保透明性，我们在第 ?? 节中提供了受访者分布的细目，且在第 4.1.1 节中进一步分析每个人口统计类别。我们在寻找濒危语言使用者时遇到了挑战，仅在 811 个有效回应中收集到 17 个。由于印尼各地濒危语言使用者分布稀疏且数量极少，接触他们非常困难。为此，我们最大化收集受访者的努力，以期能尽可能多地包括濒危语言使用者。

在问卷调查中，即使我们采用了注意力检查问题 (Muszynski, 2023)，仍有可能有些受访者试图多次填写调查问卷以增加赢得奖品的机会。为了进一步缓解这一问题，我们通过识别重复的电话号码或电子邮件来实施额外的安全措施。如果发现重复，则仅保留一个回复，并且该受访者被视为无资格获得奖品。

此外，在机器翻译重要性的问题中，我们没有像附录 ?? 中的问题 23 那样询问受访者认为哪种类型的机器翻译重要，而是可以将该问题设计得类似于其他语言技术的问题。然而，我们这样设计以便更清楚地了解机器翻译的哪些方面与他们的日常生活最相关。

我们仅收集同意将其数据用于进一步分析的受访者的数据。在调查开始时（见附录 ??），我们提供了关于调查目的明确信息，明确指出这是一个没有商业意图的学术研究，并向受访者保证，他们的数据将被保密，并且仅用于研究目的，确保数据和存储库在任何情况下都保持私密。

然而，参与者并未完全匿名化，因为我们要求提供联系方式，以便实施奖励/奖金抽奖系统——这在印度尼西亚是一种常见的表示感谢的做法。尽管如此，提供联系信息并非强制性要求；参与者可以跳过该部分并仍然完成调查。此外，除了用于更深入分析的人口统计信息外，我们没有收集其他敏感数据（例如，姓名、具体位置），以在进行全面研究的同时保护被调查者的隐私。

致谢 由于双盲而保留。

References

- Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névél, 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 Prasoj, 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.
- 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 Alfinia, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kausubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Dandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapusita, 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, Pujiyanti 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, Orevaghene 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, Orevaghene 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 Fadhillah, 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 Muszyski. 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 Prasajo, 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.
- 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](#).

在本节中，我们展示了完整的问卷，其中包含其原始的印尼语措辞，随后是英文翻译。原始文本以黑色显示，翻译内容以 灰斜体 显示，附加细节在 蓝色 中。此外，注意力检查问题 (Muszyski, 2023) 以及我们用于验证回答的方法标记在 红色 。

磅_____

印尼地方语言的语言技术调查
印尼当地语言的语言技术调查

这项调查是为了理解社区对印度尼西亚地方语言技术的理解而进行的。该调查是学术研究，不具有商业性质。

基于人工智能 (AI) 的语言技术，例如 Google 翻译、Google 助手和 Siri，在日常生活中我们已经经常使用。这项调查旨在了解您对使用语言技术于您地方语言的看法。该调查针对具有地方语言能力的人。受访者的数据保密将被妥善保护，且仅用于此调查的目的。

总奖金额为 3,000,000 印尼盾。调查结束时 (2024 年 12 月 8 日)，我们将随机选出 10 名获奖者，每人将获得 300,000 印尼盾。

这项调查旨在了解公众对印度尼西亚地区语言 LT 的理解。此调查属于学术研究，并非商业性质。

基于人工智能 (AI) 的 LT 工具, 如谷歌翻译、谷歌助手和 Siri, 在我们的日常生活中经常被使用。此调查的目的是了解您对 LT 在您的地区语言中使用的看法。此调查专为那些拥有地区语言技能的人士而设计。调查数据的保密性将得到良好保护, 仅用于本次调查的目的。总奖金额为 300 万印尼盾。在调查结束后 (于 2024 年 12 月 8 日), 我们将随机选出 10 名获奖者, 每人将获得 30 万印尼盾。

点_____

1. 您能使用地方语言吗? (请选择所有符合项)

1. 您会使用任何区域语言吗? (选择所有适用的选项)

☐ Saya bisa berbicara menggunakan bahasa daerah 747 (86.8 %)
我可以使用方言交流

☐ Saya bisa menulis dengan bahasa daerah 533 (61.9 %)
我可以使用地方语言写作

☐ Saya bisa membaca dan memahami teks dengan bahasa daerah 652 (75.7 %)
我能够阅读和理解区域语言的文本

☐ Saya tidak bisa sama sekali 30 (3.5 %)
我不能

点_____

自我介绍

介绍

2. 用您掌握的方言书写!

2. 写下任何您擅长的地区语言!

861 个书写答案

3. 请写出您的地方方言 (如果有的话)!

方言是由一个特定人群使用的语言变体, 这些人群在地理位置和其他特征上有一定的共同点。

例如: (1) 拖巴方言, (2) 曼达林方言, (3) 西马隆及方言, (4) 普珀 (达伊里) 方言, (5) 加罗方言。

3. 写下您的地方语言方言 (如果有的话)!

方言是一种语言的变体, 用于具有某些特征的一组讲话者, 例如该地区的地理位置和相对相似的特征。

示例: (1) 多巴方言, (2) 曼达岭方言, (3) 锡马隆贡方言, (4) 帕帕克 (戴里) 方言, (5) 卡罗方言。

838 个填写的答案。23 人回答 “-” 或 “tidak ada” (没有方言)

4. 您使用地方语言的流利程度如何?

4. 你的地方语言流利程度如何?

☐ Sangat fasih 289 (33.6 %)
非常流利

☐ Fasih 449 (52.1 %)
流利

☐ Tidak fasih 110 (12.8 %)
不流利

☐ Sangat tidak fasih 13 (1.5 %)
非常不流利

5. 您使用方言的频率是多少?

5. 你多久使用一次你的地方语言?

☐ Setiap hari 534 (62 %)
每天

☐ Beberapa kali dalam seminggu 205 (23.8 %)
每周几次

☐ Sekali dalam seminggu 26 (3 %)
每周一次

☐ Sekali dalam sebulan 16 (1.9 %)
每月一次

☐ Sangat jarang 80 (9.3 %)
极少

6. 您来自哪个省份?

6. 你来自哪个省?

包含 38 个省作为单选选项的多项选择题。861 个答案

7. 您的民族是什么? (如果没有民族, 您可以写 “印度尼西亚”)

7. 你的部落是什么? (如果没有, 可以写 “印尼”)

861 个自填答案。46 人回答 “印度尼西亚”

8. 你的性别是?

8. 你的性别是什么?

☐ Perempuan 453 (52.6 %)
女性

☐ Laki-laki 408 (47.4 %)
男性

9. 您的最后学历是什么?

9. 您的最高学历是什么?

☐ Tidak bersekolah 1 (0.1 %)
未上学

☐ SD 0 (0 %)
小学

☐ SMP 0 (0 %)
初中

☐ SMA 144 (16.7 %)

高中

☐ S1 412 (47.9 %)

本科生

☐ S2 257 (29.8 %)

研究生

☐ S3 47 (5.5 %)

博士

10. 您的年龄是多少?

10. 你多大了?

☐ <19 tahun 34 (3.9 %)

小于 19 岁

☐ 20-29 tahun 251 (29.2 %)

20 到 29 岁

☐ 30-39 tahun 290 (33.7 %)

30-39 岁

☐ 40-49 tahun 195 (22.6 %)

40-49 岁

☐ 50-59 tahun 80 (9.3 %)

50-59 岁

☐ >60 tahun 11 (1.3 %)

60 岁以上

11. 您的职业是什么?

11. 你的职业是什么?

861 个书写答案。

12. 在什么情况下您会积极 (写、说) 或被动 (读、听) 使用本地方言?

13. 输入您的 WhatsApp 号码或邮箱。(方便联系您如果您中奖了)

14. 注释

☐ Seratus* 8 (0.9 %)

一百

☐ Dua ratus 847 (98.4 %)

二百

☐ Tiga ratus* 2 (0.2 %)

三百

☐ Empat ratus* 4 (0.5 %)

四百

: 一百加一百是多少? 14. 一百加一百是多少?

我们省略了这些反应进行分析

磅_____

与方言相关的问题

与地区语言相关的问题

请根据以下陈述填写一些问题, 将您和您的方

言置于一种情境中。

根据您的情况和当地语言填写以下陈述中的问题。

15. 我们的地方语言有不同的礼貌等级, 比如在与同龄人和年长者交谈时词语的差异。

16. 我经常在直接对话中看到我的地方语言被使用。

16. 我经常在口头交流中遇到使用我的地区语言。

☐ Sangat setuju 487 (56.6 %)

非常同意

☐ Setuju 343 (39.8 %)

同意

☐ Tidak setuju 28 (3.3 %)

不同意

☐ Sangat tidak setuju 3 (0.3 %)

非常不同意

17. 我经常在书面形式中遇到我的地方语言。

17. 我经常遇到以书面形式使用的我的地方语言。

☐ Sangat setuju 210 (24.4 %)

非常同意

☐ Setuju 417 (48.4 %)

同意

☐ Tidak setuju 212 (24.6 %)

不同意

☐ Sangat tidak setuju 22 (2.6 %)

非常不同意

点_____

对地区语言的态度

对本地语言的态度

请根据以下几种陈述, 填写下列问题。

根据下面的一些陈述来回答这些问题。

18. 我希望地方语言能够持续存在, 并被许多人使用。

19. 我想学习印尼其他地区的方言。

20. 我经常遇到讲地方语言的人, 但我不能理解他们的语言。

与语言技术相关的问题

与语言技术相关的问题

21. 您的地方语言文字是否已经被智能手机或电脑等技术所支持?

21. 您的地区语言文字是否得到智能手机或电脑等技术支持?

- ☐ Ya 291 (33.8 %)
是
- ☐ Tidak 365 (42.4 %)
否
- ☐ Tidak tahu 205 (23.8 %)
不知道

翻译机
机器翻译

22. 您是否曾经使用过翻译机器, 例如 Google 翻译?

23. 地方语言翻译机器对您的需求有多重要?

23. 地区语言翻译机对您的需求有多重要?

- ☐ Penting untuk menerjemahkan bahasa daerah ke bahasa Indonesia. 622 (72.2 %)
将区域语言翻译成印尼语是很重要的。
- ☐ Penting untuk menerjemahkan bahasa Indonesia ke bahasa daerah. 454 (52.7 %)
将印度尼西亚语翻译成当地语言是很重要的。
- ☐ Penting untuk menerjemahkan antar bahasa daerah. 410 (47.6 %)
在区域语言之间进行翻译是重要的。
- ☐ Penting untuk menerjemahkan bahasa daerah ke bahasa asing. 374 (43.4%)
将区域语言翻译成外国语言是很重要的。
- ☐ Penting untuk menerjemahkan bahasa asing ke bahasa daerah. 33 (3.8 %)
将外语翻译为本地语言是很重要的。
- ☐ Tidak penting 52 (6.0 %)
不重要

24. 您希望在哪里看到或使用针对您所在地区语言的翻译机?

语音转文本
语音转文字

25. 语音转文字是可将声音转换为文本的系统。您是否曾经使用过这个应用程序?

25. 语音转文本是一个将语音转换为文本的系统。你曾经使用过这样的应用程序吗?

- ☐ Ya 655 (76.1 %)
是
- ☐ Tidak 206 (23.9 %)
否

26. 对于您的需求而言, 方言语音转文本有多重要?

26. 地区语言文本转语音对您的需求有多重要?

- ☐ Sangat penting 285 (33.1 %)
非常重要
- ☐ Penting 349 (40.5 %)
重要
- ☐ Tidak terlalu penting 197 (22.9 %)
不太重要
- ☐ Tidak penting 30 (3.5 %)
不重要

27. 您希望在哪里查看或使用适用于您的地方语言的语音转文本?

文本到语音

28. Text-to-speech 是一个将文本转换为语音的系统。您曾经使用过这样的应用程序吗?

28. 文字转语音是一个将文字转换为语音的系统。你曾经使用过这样的应用程序吗?

- ☐ Ya 620 (72.0 %)
是
- ☐ Tidak 241 (28.0 %)
否

29. 对于您的需求来说, 方言文本转语音有多重要?

29. 地区语言的文本转语音对您的需求有多重要?

- ☐ Sangat penting 283 (32.9 %)
非常重要
- ☐ Penting 373 (43.3 %)
重要
- ☐ Tidak terlalu penting 168 (19.5 %)
不是很重要
- ☐ Tidak penting 37 (4.3 %)
不重要

30. 你想在哪里查看或使用针对您本地语言的语音合成功能?

30. 您希望在哪里看到或使用针对您所在地区语言的文本转语音功能?

- ☐ Aplikasi ponsel 691 (80.3 %)
移动应用
- ☐ Platform sosial media 283 (32.9 %)
社交媒体平台
- ☐ Situs web 392 (45.5 %)
网站

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

数字文档 (PDF、word)

☐ Platform pembelajaran online 172 (20.0 %)

在线学习平台

☐ Sistem di tempat kerja 123 (14.3 %)

职场系统

☐ Saat bepergian atau di tempat umum 250 (29.0 %)

在旅行或公共场合时

☐ Tidak tertarik 50 (5.8 %)

不感兴趣

31. 选择作为颜色名称的答案

31. 选择颜色名称的答案

☐ Baju* 11 (1.3 %)

衣物

☐ Perahu* 0 (0.0 %)

船

☐ Merah 846 (98.3 %)

红色

☐ Kursi* 1 (0.1 %)

椅子

☐ Pena* 3 (0.3 %)

笔

注: 我们省略了这些回答的分析

语法检查器

32. 语法检查器是一种工具或软件, 旨在自动检测和纠正文本中的拼写和语法错误, 从而帮助提高写作质量。

您是否曾经使用过这样的应用程序?

32. 语法检查器是一种工具或软件, 旨在自动检测和纠正文本中的拼写和语法错误, 从而帮助提高写作质量。你曾经使用过这样的应用程序吗?

☐ Ya 643 (74.7 %)

是

☐ Tidak 218 (25.3 %)

否

33. 对于您的需求来说, 地方语言语法检查工具有多重要?

34. 您想在哪里查看或使用本地语言的语法检查器?

搜索引擎
信息检索

35. 您是否曾使用过信息搜索技术, 例如 Google 搜索?

36. 在您看来, 信息检索技术对于地方语言有多重要?

数字助手

数字助手

37. 数字助手是一种基于人工智能的软件, 帮助用户通过语音或文本命令完成日常任务, 如回答问题、安排日程、控制智能设备。例子包括: ChatBot、Siri、Alexa 和 Google Assistant。

您是否曾使用过类似的应用程序?

37. 数字助理是基于人工智能的软件, 通过语音或文本命令帮助用户完成日常任务, 例如回答问题、设置时间表和控制智能设备。例子有: ChatBot、Siri、Alexa 和 Google Assistant。你曾使用过这样的应用程序吗?

☐ Ya 679 (78.9 %)

是

☐ Tidak 182 (21.1 %)

否

38. 您认为地方语言的数字助理对您的需求有多重要?

38. 对于您的需求来说, 地区语言数字助手有多重要?

☐ Sangat penting 286 (33.2 %)

非常重要

☐ Penting 330 (38.3 %)

重要

☐ Tidak terlalu penting 201 (23.3 %)

不太重要

☐ Tidak penting 44 (5.1 %)

不重要

39. 您想要在什么情况下使用支持您的方言的数字助手?

39. 您希望出于什么目的使用支持您所在地区语言的数字助理?

☐ Konsultasi kesehatan 188 (21.8 %)

健康咨询

☐ Curhat masalah pribadi 150 (17.4 %)

分享个人问题

☐ Hiburan 316 (36.7 %)

娱乐

☐ Membantu belajar / pendidikan 514 (59.7 %)

帮助学习/教育

☐ Mencari informasi 604 (70.2 %)

搜索信息

- ☐ Menuliskan teks seperti surat 263 (30.5 %)
书写像信件一样的文字
- ☐ Memperbaiki penulisan teks 346 (40.2%)
修正文本写作
- ☐ Tidak perlu 76 (8.8 %)
没有必要
- ☐ Lainnya 24 (2.8 %)
其他

40. 数字助理也可以读取图片和视频。您认为拥有能理解与您文化相关的图片和视频的方言数字助理重要吗？

40. 数字助手也可以读取图像和视频。您认为拥有能够理解与您的文化相关的图像和视频的区域语言数字助手重要吗？

- ☐ Sangat penting 352 (40.9 %)
非常重要
- ☐ Penting 379 (44.0 %)
重要
- ☐ Tidak terlalu penting 108 (12.5 %)
不是很重要
- ☐ Tidak penting 22 (2.6 %)
不重要

隐私和可信度
隐私和可信度

41. 为了开发本地语言技术，需要该语言的大量数字文本和音频数据。例如，研究人员可能会收集和分析您在社交媒体上使用的本地语言的公开可用文本和音频数据。这是否会让您感到困扰？

41. 为了开发地区语言技术，需要大量该语言的数字文本和音频数据。例如，研究人员可能会收集和分析社交媒体上使用您地区语言的公开文本和音频数据。这会让您感到困扰吗？

- ☐ Saya merasa terganggu jika data teks tersebut digunakan untuk pengembangan teknologi bahasa daerah 30 (3.5 %)
如果文本数据被用于发展区域语言技术，我会感到不安
- ☐ Saya merasa terganggu jika data audio tersebut digunakan untuk pengembangan teknologi bahasa daerah 29 (3.4 %)
如果音频数据被用于区域语言技术的发展，我会感到不安
- ☐ Saya merasa terganggu jika data teks dan audio tersebut digunakan untuk pengembangan teknologi bahasa daerah 36 (4.2 %)

如果文本和音频数据用于开发区域语言技术，我会感到不安

- ☐ Saya tidak merasa terganggu karena data tersebut tersedia secara publik 766 (89.0 %)
我并不感到不安，因为这些数据是公开可用的

42. 您是否认为现有的人工智能技术能够充分保护您的个人数据？

43. 在使用语言技术如 Google 搜索、Siri 和 Google 助手时，您是否听说过有关隐私和安全性的问题？例如，不向像 ChatGPT 这样的数字助手提及或写下个人数据？

43. 在使用诸如谷歌搜索、Siri 和谷歌助手等语言技术时，您是否听说过隐私和安全问题？例如，不向像 ChatGPT 这样的数字助手提及或输入个人数据？

- ☐ Sangat tahu 140 (16.3 %)
非常清楚
- ☐ Cukup tahu 354 (41.1 %)
感知
- ☐ Tidak terlalu tahu 216 (25.1 %)
不太清楚
- ☐ Tidak tahu 151 (17.5 %)
不知道

44. 您是否曾向像 ChatGPT 这样的数字助手咨询健康问题？

44. 你有没有询问过像 ChatGPT 这样的数字助手关于健康问题？

- ☐ Pernah 278 (32.3 %)
我有
- ☐ Tidak pernah 583 (67.7 %)
我没有

45. 您多长时间进行一次对由语言技术（如 ChatGPT）提供的信息的真实性验证？

45. 您多久核实一次由 ChatGPT 等语言技术提供的信息的准确性？

- ☐ Selalu 130 (15.1 %)
始终
- ☐ Sering 262 (30.4 %)
经常
- ☐ Jarang 274 (31.8 %)
很少
- ☐ Tidak pernah 195 (22.6 %)
从不

46. 您知道像 ChatGPT 这样的数字助理提供的信息不总是正确的，并且可能完全错误吗？

46. 你知道诸如 ChatGPT 之类的数字助理提供的信息并不总是正确的，有时甚至可能完全错误吗？

- ☐ Sangat tahu 311 (36.1 %)
非常清楚
- ☐ Cukup tahu 323 (37.5 %)
意识到
- ☐ Tidak terlalu tahu 109 (12.7 %)
不是特别清楚
- ☐ Tidak tahu 118 (13.7 %)
不知道

47. 选择草莓选项

47. 选择草莓答案选项

- ☐ Apel* 10 (1.2 %)
苹果
- ☐ Pisang* 4 (0.5 %)
香蕉
- ☐ Jeruk* 4 (0.5 %)
橙子
- ☐ Stroberi 832 (96.6 %)
草莓
- ☐ Semangka* 11 (1.3 %)
西瓜

注记：我们在分析中省略了这些响应

48. 当使用语言技术时，您是否曾听说过偏见问题？例如：

(1) 性别偏见：计算机假设医生是男性，护士是女性。事实上，存在女性医生和男性护士。(2) 宗教/政治偏见：计算机会反映对某些宗教/政治的偏见，从而使某些群体处于不利地位。

48. 在使用语言技术时，你是否听说过偏见问题？例如：(1) 性别偏见：计算机假定医生是男性而护士是女性。事实上，有女性医生和男性护士。(2) 对宗教/政治的偏见：计算机反映了对某些宗教/政治的偏见，从而使某些群体处于不利地位。

- ☐ Sangat tahu 138 (16.0 %)
非常清楚
- ☐ Cukup tahu 335 (38.9 %)
知晓
- ☐ Tidak terlalu tahu 216 (25.1 %)
不太意识到
- ☐ Tidak tahu 172 (20.0 %)
不知道

49. 请写下您希望表达的其他关于语言技术的问题，如聊天机器人、数字助理、翻译机等等。

49. 写下您希望传达的关于语言技术的问题，例如聊天机器人、数字助理、机器翻译等。

861 个书写答案

隐私和可信度

隐私和可信度

50. 一般来说，您对发展地方语言的语言技术的热情如何？您对其开发有担忧或不满吗？

50. 总的来说，您对您所在地区语言的 language 技术发展有多热衷？您对其发展有什么担忧或不喜欢的地方吗？

- ☐ Saya antusias dan tidak khawatir 512 (59.5 %)
我很热情，并不担心
- ☐ Saya antusias dan sedikit khawatir 287 (33.3 %)
我既感到热情又有点担心
- ☐ Saya tidak antusias, namun sedikit khawatir 26 (3.0 %)
我不是很热情，但有点担心
- ☐ Saya tidak antusias dan tidak khawatir 36 (4.2 %)
我既不热情也不担心

A 重要性评分变化的详细信息

表 2 显示了在各个类别中的重要性评分。符号 (*) 表示每种语言技术 (LT) 的调查者总数：机器翻译 (MT) 为 753，语音转文本 (STT) 为 623，文本转语音 (TTS) 为 589，语音控制 (GC) 为 612，信息检索 (IR) 为 800，对话代理 (DA) 为 642。同时，符号 (**) 表示另一个调查者子集的相应数字：机器翻译 (MT) 为 58，语音转文本 (STT) 为 188，文本转语音 (TTS) 为 222，语音控制 (GC) 为 199，信息检索 (IR) 为 11，以及对话代理 (DA) 为 169。

B 根据维基百科的印度尼西亚西部和东部分区

我们根据多个标准汇总了结果，包括将印度尼西亚分为西部和东部地区。我们参考了相关的维基百科页面⁷来对省份进行简单的分类。表 3 展示了西印度尼西亚和东印度尼西亚之间的分布，随后是每个省的受访者数量。

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

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: 在不同人口统计和意识类别中的重要性评分。

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: 根据省份所在地（印尼西部 & 东部）划分和有效受访者数量。

C 语言层次聚合

Eberhard et al. (2023) 建立了一个基于实际使用情况的语言分类系统。该系统包括九个语言状态级别，从国际语言到灭绝语言⁸：

- 0. 国际: 这种语言在国家之间的贸易、知

⁸<https://www.ethnologue.com/methodology/#language-status>

识交流和国际政策中被广泛使用。在我们的调查中不适用

- 1. 国家: 该语言在国家层面的教育、工作、大众传媒和政府中使用。在我们的调查中不适用
- 2. 省级: 该语言用于一个国家主要行政区划内的教育、工作、大众媒体和政府。此项不适用于我们的调查
- 3. 更广泛的交流: 这种语言在工作和大众媒体中不具备官方地位，但用于跨越一个地区的语言差异。
- 4. 教育方面: 这种语言在广泛的制度支持的教育体系中得到了持续使用，其标准化和文学得以维持。
- 5. 发展中: 该语言正在积极使用，其中一些使用经过标准化形式的文献，尽管这还没有广泛传播或可持续。
- 6a. 活力: 这种语言被所有世代用于面对面的交流，并且这种情况是可持续的。
- 6b. 受威胁: 这种语言在所有代际之间用于面对面交流，但其使用者正在减少。
- 7. 转移: 生育期的一代可以在他们之间使用该语言，但没有将其传授给孩子。
- 8a. 垂死状态: 语言的唯一现存活跃使用者是祖父母辈及更年长的成员。
- 8b. 濒临绝迹: 该语言唯一剩余的使用者是祖父母及更年长的一代人，他们几乎没有机会使用这种语言。

- 9. 休眠: 该语言作为一个民族社区的遗产身份的提醒, 但没有人能掌握它超过象征性的熟练程度。
- 10. 灭绝: 语言不再使用, 没有人保留与该语言相关的民族认同感。在我们的调查中不适用

然而, 为了便于分析, 我们将这 13 个级别合并为 3 个更广泛的类别。表 4 展示了我们的分类以及调查中涉及的语言。

如同在第 ?? 节中讨论的, 方言也影响同一语言的使用者对语言技术 (LTs) 需求的认知。由于受访者数量有限, 我们集中在五种语言及其相应的方言: 亚齐语 (大亚齐和万达亚齐方言)、布吉语 (望加锡、邦和布吉斯卡尤瓦方言)、爪哇语 (阿雷坎、潘达鲁岸和玛塔拉曼方言)、米南加保语 (阿甘和巴雅贡布方言), 以及巽他语 (万隆普里安甘和松江方言), 如图 9 所示。

总的来说, 班达亚齐、巴亚孔布和万隆 Priangan 方言在它们各自的语言中比其他方言更为重视 LTs。值得注意的是, 布吉语中的 Bone 方言表现出一种明显的偏好, 讲者更优先考虑 GC 和 IR, 而对 MT 的兴趣较小。相比之下, 望加锡方言认为 LTs 的重要性低于其他布吉语方言。

然而, 这些趋势背后的原因仍不清楚。要完全理解为何某些方言在感知 LTs 方面表现出独特模式, 与每种方言的使用者进行直接对话是必不可少的。

D 隐私意识如何影响使用率

图 6 说明了受访者对隐私问题的关注程度与他们使用语言技术 (LTs) 的频率之间的关系。总体而言, 认为 LTs 未能为个人数据提供足够保护的人不太可能使用数字助手进行健康相关的咨询, 因为此类信息被认为是高度敏感的。同样, 对于 LTs 所提供的数据保护水平仍然不确定的人, 往往会完全避免使用这些技术进行健康相关的问题查询。

E 对语言技术的熟悉程度: 按世代、语言水平和地理位置分类

图 7 显示了在本调查中分析的受访者对语言技术 (LTs) 的熟悉程度, 并根据不同因素进行分类。在各代人中, Z 世代对 LTs 的熟悉程度最高, 而 X 世代和婴儿潮一代的熟悉程度最低, 这可能是由于快速的全球化进程对年轻一代的影响更大。此外, 使用稳定语言的人往往比其他人对 LTs 的熟悉程度更高。地理上来看, 来自印度尼西亚西部的受访者对 LTs 的熟悉程度

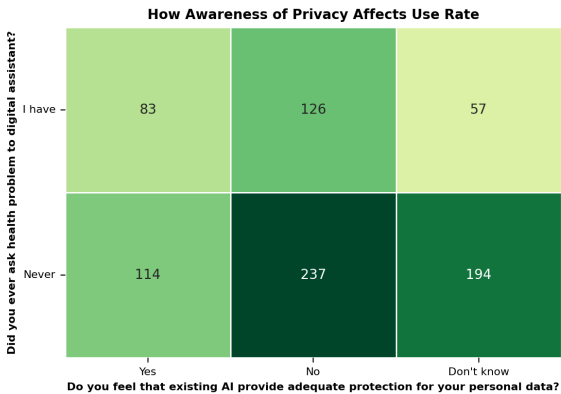


Figure 6: 隐私意识如何影响使用率。

高于来自印度尼西亚东部的受访者, 这可能是由于印度尼西亚的发展集中在爪哇和苏门答腊等人口较多的岛屿。此外, 图 8 显示了熟悉 LT 的受访者在几个类别中的重要性评分, 接着是 LT 熟悉度与其重要性评分之间的皮尔逊相关性。

我们使用维基百科数据作为数据集收集的通用文本来源。图 10 说明了尽管一些印度尼西亚地方语言的重要性评分很高, 但可用的资源仍然不足。只有少数几种语言——如爪哇语、巽他语、巴厘语和米南加保语——的数据集超过 10MB (这仍然被视为微小)。同时, 所有其他语言的资源仍然有限或完全不支持。

表 5 使用谷歌作为基准, 展示了印度尼西亚地方语言的语言技术的当前状态。虽然某些语言, 如爪哇语和巽他语, 在某些语言技术中得到支持, 但许多其他欠代表的语言仍然缺乏覆盖。此外, 像 TTS 和 DA 这样的技术尚未支持任何印度尼西亚地区的语言。这概述了这些语言的语言技术发展差距。

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: 基于每种语言的语言水平分类和有效回答者数量。

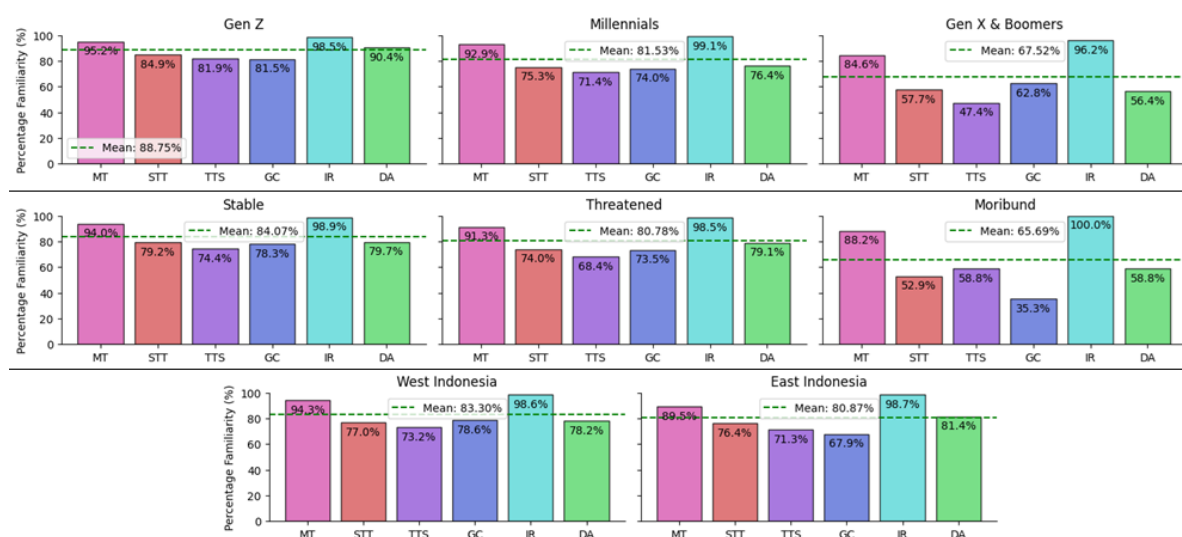


Figure 7: 按多种类别熟悉 LTs。顶行按代际分类（Z 世代、千禧一代、X 世代、& 婴儿潮一代），中行按语言濒危程度分类，底行按印度尼西亚地区分类（印尼西部和东部）。

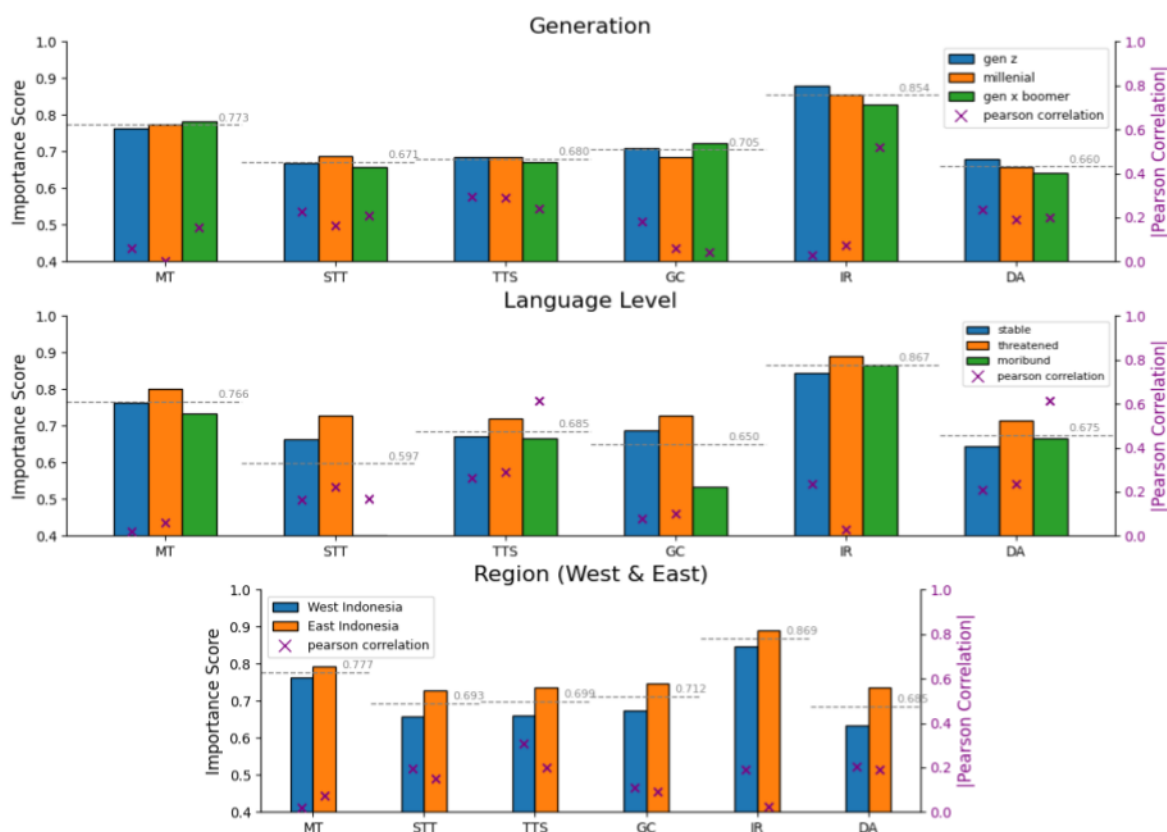


Figure 8: 熟悉 LT 的受访者在几个类别之间的重要性评分：世代、语言水平和地区（西部 & 东部印度尼西亚），以及它们的皮尔逊相关系数。

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: 每个 LT 的重要性得分及其在谷歌支持的印尼本地语言中的可用性。斜体的重要性得分仅考虑“非常重要”选项。* 他们的文字字母也被支持 ** 仅能够从文档中提取实体 *** 谷歌助手（安卓手机 & 电视）& Gemini 仅在印尼语（ind）中可用。

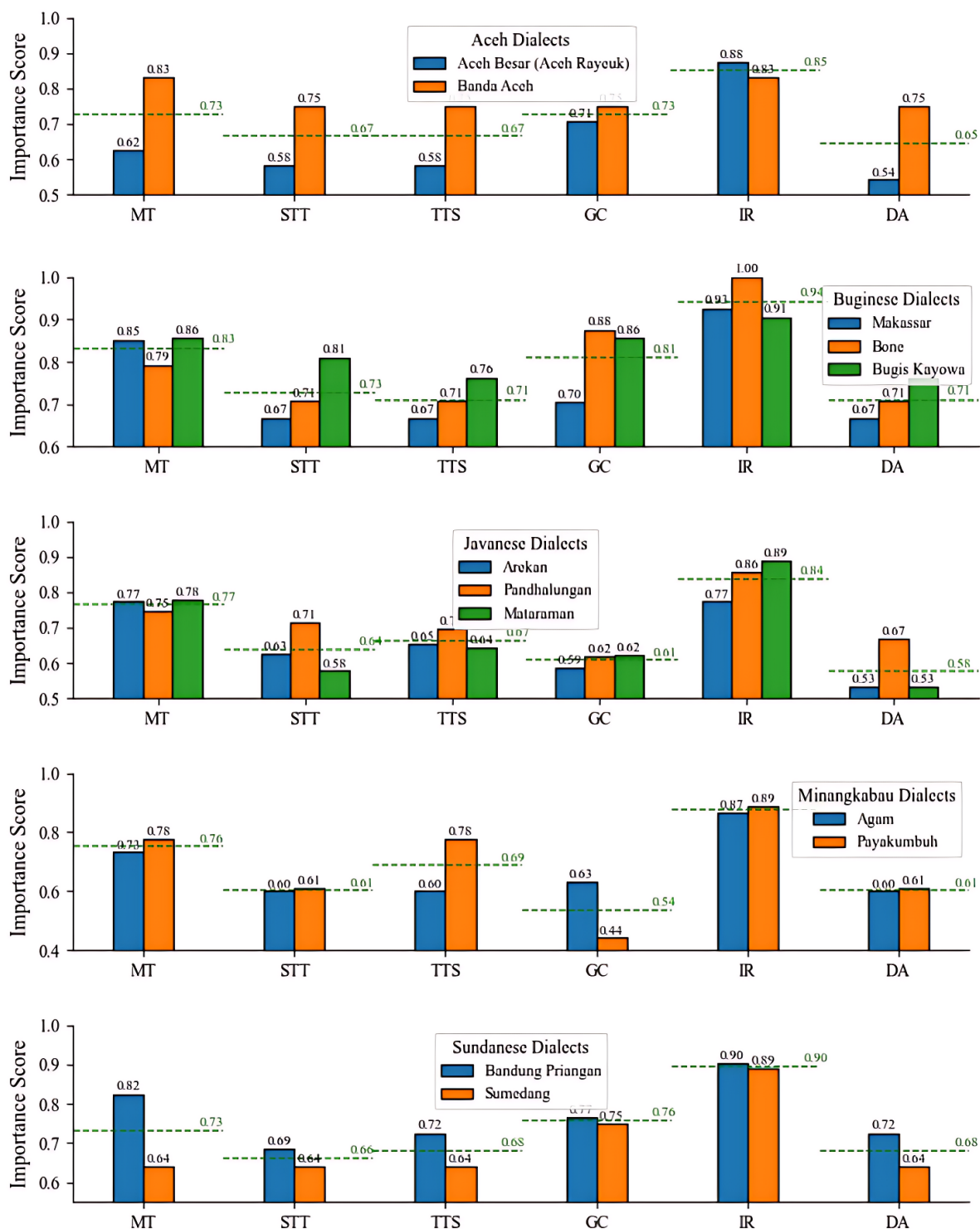


Figure 9: 亚齐、布吉、爪哇、米南加保和巽他方言在 LT 倾向上的差异 (从上到下)。

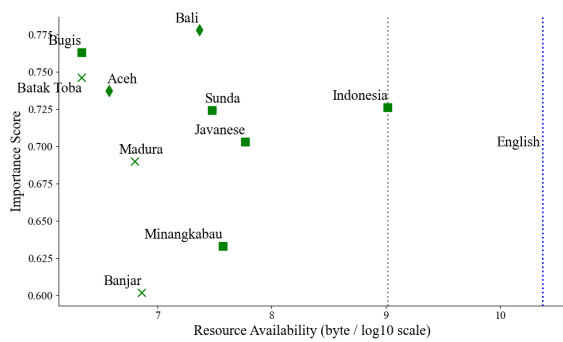


Figure 10: 维基百科上每种支持的印尼本地语言的重要性得分和可用资源。■ 代表有超过 50 名受访者的语言，◆ 代表有 30-50 名受访者的语言，× 代表少于 30 名受访者的语言。