Euphemism-SRH: A Multilingual Benchmark for Evaluating LLMs on Indirect Women's Health Queries

Anonymous ACL submission

Abstract

Medical question answering (QA) benchmarks typically evaluate models on explicit and clinically transparent health questions, despite real patient communication often being indirect. This is particularly common in women's sexual and reproductive health (SRH), where stigma, taboo, and sociocultural norms encourage the use of euphemistic language to describe sensitive symptoms, body parts, or sexual experiences. As a result, current benchmarks rarely evaluate whether LLMs can recover patients' intended medical meanings when concerns are expressed indirectly. To address this gap, we introduce EUPHEMISM-SRH, a multilingual benchmark for culturally grounded euphemistic women's SRH QA. The dataset contains 1,050 Indonesian doctor–patient QA pairs aligned with six regional-language translations, yielding 7,350 language-specific instances across seven languages. Each instance is annotated by women native-speaker annotators for euphemism presence, euphemistic spans, and corresponding literal meanings. We evaluate LLMs under euphemistic, non-euphemistic, and de-euphemized settings using factual precision, factual recall, hallucination rate, and omission rate. Our results show that the raw impact of euphemism is highly language-dependent and does not uniformly degrade performance across languages. However, replacing euphemistic expressions with literal meanings consistently improves factuality while reducing hallucinated and omitted information. These findings highlight euphemism as an important yet underexplored challenge in medical QA, and motivate culturally grounded SRH benchmarks for safer and more reliable medical LLMs in low-resource settings.

1 Introduction

Euphemistic expressions replace language considered inappropriate, offensive, or discomforting, enabling more socially acceptable communication around taboo or sensitive topics (Firsich and Rios, 2024; Kohli et al., 2022; Danescu-Niculescu-Mizil et al., 2013; Magu and Luo, 2018). Their use is shaped by social context, gender, and communicative setting (Kapron-King and Xu, 2021). Prior sociolinguistic work suggests that women often use, and are described through, more diverse and frequent euphemistic expressions than men (Lakoff, 1973, 1977; Jespersen, 2013). This is especially relevant in women's sexual and reproductive health (SRH), where stigma, modesty norms, and cultural constraints can encourage patients to discuss menstruation, contraception, sexual activity, reproductive organs, and genital symptoms indirectly (Chaudhri et al., 2022). Euphemism is therefore not merely a stylistic device: in doctor–patient communication, it can help patients raise sensitive concerns, but it can also obscure clinically important information (Khisamova and Abdullina, 2022; Khaydarova, 2024; Ezeugo and Chukwu, 2024). Because clear and precise language is central to effective doctor–patient relationships and therapeutic outcomes (Odebunmi, 2010), euphemistic SRH communication creates a tension between sociocultural acceptability and clinical clarity.

This tension becomes increasingly important as patients turn to large language models (LLMs) for health information. LLMs are attractive in digital health settings because they are accessible, practical, and capable of simplifying medical terminology (Zaretsky et al., 2024); some evaluations even report no strong preference between physician and LLM responses in specific health-information settings (Yan et al., 2025). However, existing medical question-answering benchmarks (Ben Abacha and Demner-Fushman, 2019; Pal et al., 2022) largely assume that patients ask explicit, clinically transparent questions. This assumption does not hold in women's SRH, where patients may avoid direct references to sensitive symptoms, body parts, or sexual experiences. As a result, current medical

QA evaluations leave open a critical benchmark gap: they do not test whether LLMs can infer what patients actually mean when medically relevant concerns are expressed euphemistically rather than directly.

Euphemistic SRH communication is particularly challenging in multilingual and low-resource settings. Since LLMs are primarily trained on literal and high-resource text, they often struggle with implicit and culturally grounded meanings (Liu et al., 2021). Because euphemisms vary across languages and sociocultural contexts (Sammartino et al., 2025), expressions that are interpretable in one community may become ambiguous or misleading in another. These issues are further amplified in low-resource languages, where both linguistic coverage and exposure to SRH-related euphemisms remain limited. Existing low-resource QA benchmarks also rarely include gendered health topics or euphemistic variants, leaving substantial gaps in both language coverage and content relevance. Evaluating LLMs on euphemistic SRH questions therefore requires not only medical QA capabilities, but also cultural inference, low-resource language understanding, and factual medical response generation. To address this gap, our contributions are threefold: **First,** we introduce EUPHEMISM-SRH, a multilingual women's SRH QA dataset containing 1,050 Indonesian doctor–patient QA pairs aligned with six regional-language translations, resulting in 7,350 language-specific instances annotated with euphemism labels, spans, and literal meanings. **Second,** we analyze how euphemism prevalence, diversity, and density vary across languages and SRH topics, showing that euphemism usage is culturally and linguistically uneven. **Third,** we benchmark LLMs under euphemistic, non-euphemistic, and de-euphemized settings, showing that euphemism effects are language-dependent while de-euphemization consistently improves factuality.

2 Related Work

Euphemism and gendered health communication. Healthcare communication is shaped by social, cultural, and psychological factors, and non-direct language is often used in sensitive domains to reduce discomfort, maintain politeness, and support patient autonomy (Jamet, 2018; Khisamova and Abdullina, 2022). Euphemism is functionally embedded in doctor–patient interactions, influencing how symptoms, diagnoses, advice, and risks are conveyed and understood (Ezeugo and Chukwu, 2024; Humayun et al., 2025). While such language can make taboo topics more discussable, it can also obscure prognosis, treatment goals, or risks when clarity is needed for informed decision-making (Chou et al., 2017; Tayler and Ogden, 2005). This issue is particularly important in women's SRH: prior work suggests that women's speech and discourse about women often involve indirect or euphemistic language, especially for sensitive topics (Lakoff, 1973; Jespersen, 2013; Hysi, 2011), and gendered patterns in healthcare can contribute to women's symptoms being dismissed, minimized, or filtered through assumptions (Samulowitz et al., 2018; Bullo and Weckesser, 2021).

Multilingual and low-resource health NLP. Effective communication is essential for equitable healthcare access, yet language barriers, cultural norms, and health literacy differences affect how medical information is understood (Cate, 2025). In multilingual and low-resource settings, these challenges are amplified by non-standard terminology and culturally specific expressions (Dew et al., 2018; Rodríguez-Miret et al., 2024). Public health research also highlights the importance of culturally sensitive and sometimes indirect language for communicating about stigmatized SRH issues (Agyei and Kaura, 2023; Tuteja et al., 2022). Prior work on euphemism detection has developed methods for identifying euphemistic or coded language (Firsich and Rios, 2024; Kohli et al., 2022; Magu and Luo, 2018), including in cross-lingual settings (Sammartino et al., 2025). However, these resources are typically framed as detection or classification tasks, rather than as downstream medical QA benchmarks that test whether models can recover the intended clinical meaning and generate factually correct answers.

Medical QA and benchmark gap. Medical question answering has become a central task for evaluating whether NLP systems and LLMs can retrieve, reason over, and communicate health information. Existing benchmarks such as MedQuAD and MedMCQA primarily evaluate explicit medical knowledge and clinical reasoning (Ben Abacha and Demner-Fushman, 2019; Pal et al., 2022), while recent work studies the quality and usefulness of LLM-generated health information (Yan et al., 2025; Zaretsky et al., 2024). However, most existing resources assume that medically relevant

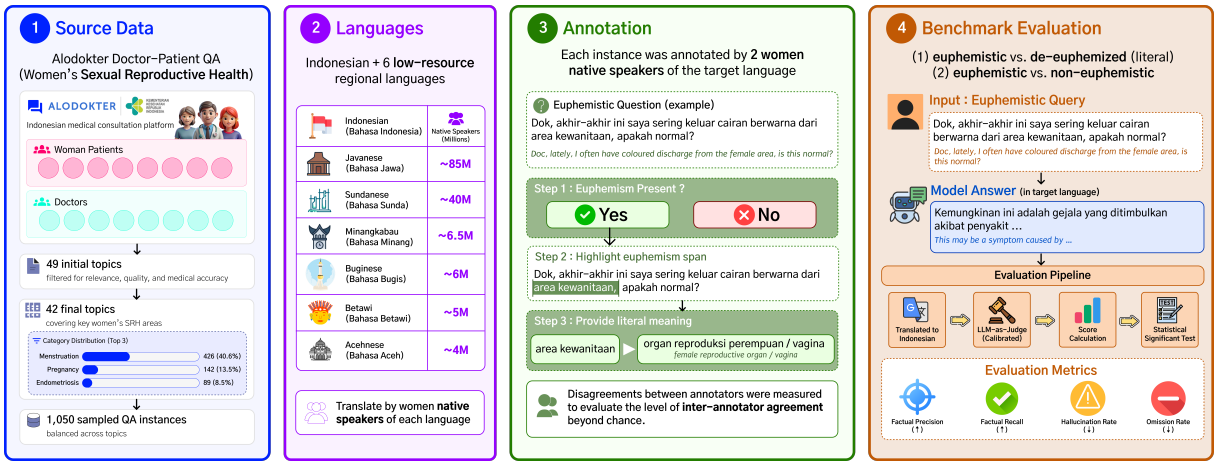

Figure 1: Overview of EUPHEMISM-SRH as a dataset and benchmark. We curate women's SRH doctor–patient QA from Alodokter, sample 1,050 instances across 42 topics, translate them into six Indonesian regional languages, annotate euphemistic spans and literal meanings, and evaluate LLMs on whether they can answer euphemistic medical questions with factual, complete, and non-hallucinated responses.

information is explicitly stated and directly recoverable from the surface form. As a result, they only partially capture challenges arising from culturally grounded and euphemistic language in women's SRH. Related resources similarly address only isolated aspects of the problem: euphemism datasets focus on detection rather than downstream medical QA (Firsich and Rios, 2024; Kohli et al., 2022; Magu and Luo, 2018; Sammartino et al., 2025), low-resource benchmarks often emphasize general-domain understanding (Winata et al., 2023), and multilingual medical NLP datasets typically prioritize translation or terminology mapping over patient-authored indirect communication (Zhou et al., 2023). Unlike prior work, EUPHEMISM-SRH combines patient-authored SRH questions, regional Indonesian languages, euphemism span and literal-meaning annotations, doctor-authored references, and factual QA evaluation, enabling a benchmark that evaluates whether models can preserve factuality when patients communicate sensitive health concerns indirectly.

3 EUPHEMISM-SRH Dataset

3.1 Dataset Construction

EUPHEMISM-SRH is designed to instantiate the benchmark gap by pairing women's SRH questions with euphemism labels, euphemistic spans, and literal meanings. Figure 1 illustrates the overall construction pipeline. Each dataset item contains a patient-authored question, a doctor-authored reference response, an SRH topic label, the language of the question, a binary euphemism label, and, when applicable, the annotated euphemistic span(s) and their literal meaning(s).

We first scrape Indonesian QA pairs from the Alodokter[1] platform. On Alodokter, patients can anonymously submit health-related questions, which are then answered by licensed doctors affiliated with the platform. The questions typically describe symptoms, health concerns, or treatment-related issues, while the doctor responses provide medical explanations, diagnoses, or treatment recommendations. Alodokter contains 1,348 health-related topic pages covering a wide range of medical domains, such as measles vaccination, dental calculus, and blindness. Since this work focuses specifically on women's sexual and reproductive health (SRH), we manually filter these topic pages and retain only topics relevant to women's SRH. This filtering step yields 49 candidate topics, corresponding to approximately 3.6% of all available topics.

We then clean and further filter the candidate topics. Topics with no retrievable QA instances are removed, such as polycystic ovary syndrome and breast swelling. This process results in 42 final SRH topics comprising 21,327 Indonesian QA instances. Before sampling and annotation, we anonymize personally identifiable information (PII) from both patient questions and doctor responses. We apply an automatic string-replacement procedure to mask sensitive entities, followed by manual verification. Identifiable names and similar sensitive strings are partially obfuscated, for example, "John" → "J***". From the cleaned corpus, we sample 1,050 QA instances, corresponding to ap-

[1] https://www.alodokter.com/komunitas/topik

proximately 5% of the available data. To preserve topic coverage, we use stratified sampling over the 42 SRH topics that ensures proportional allocation while enforcing the constraint that every topic contributes at least one instance.

The sampled Indonesian patient questions are then translated into six Indonesian regional languages: Javanese, Sundanese, Acehnese, Buginese, Minangkabau, and Betawi. The original Indonesian doctor responses are retained as reference answers for evaluation. Each translation is performed independently by a native speaker of the corresponding language. Before participating in the full translation task, each translator completes a qualification phase consisting of 10 sample translations, which are manually reviewed for linguistic accuracy and suitability. During the three-week translation process, we conduct weekly quality checks. As an additional integrity check, we compare human translations against machine-generated translations to detect possible reliance on machine translation rather than as the sole measure of translation quality. Low similarity to machine outputs suggests that the translations are not directly copied from automated systems (see Appendix B), while the qualification review and weekly audits provide the main quality-control mechanism.

Both the original Indonesian questions and the translated regional-language questions are then passed to the euphemism annotation stage. Annotation is conducted using Label Studio[2] (see Appendix G). For each language, we recruit two women annotators who are native speakers and active users of the corresponding language, except for Buginese, where only one annotator is available due to resource constraints. Annotator profiles, including language-use background and regional exposure, are provided in Appendix A.

Annotators are first asked: *"Does the question contain any euphemism related to women's health?"* If no euphemism is present, annotators select **No** and proceed to the next item. If euphemism is present, annotators select **Yes**, highlight all euphemistic spans in the question, and provide the corresponding literal meaning for each span in the same language as the question. For example, in Indonesian, *"datang bulan"* is annotated with the literal meaning *"menstruasi"*, and *"begituan"* with *"berhubungan seksual"*.[3]

[2] https://labelstud.io/

[3] A literal English translation of *"datang bulan"* is *"the moon has come"*, while *"menstruasi"* means *"menstruation"*. Similarly, *"begituan"* roughly corresponds to *"doing it"*, while *"berhubungan seksual"* means *"having sex"*.

Lang.	#Agr.	#Dis.	Sim.	α	κ
Acehnese	949	101	90.4	0.765	0.765
Betawi	942	108	89.7	0.683	0.684
Javanese	892	158	85.0	0.600	0.606
Minangkabau	804	246	76.6	0.511	0.521
Sundanese	819	231	78.0	0.555	0.557
Indonesian	1003	47	95.5	0.901	0.901

Table 1: Inter-annotator agreement for euphemism annotation across languages. **#Agr.** and **#Dis.** denote the number of agreement and disagreement cases between annotators, respectively. **Sim.** represents similarity percentage, while α and κ correspond to Krippendorff's alpha and Cohen's kappa.

For languages with two annotators, we compute inter-annotator agreement on the binary euphemism decision, as shown in Table 1. In the benchmark set, we retain only instances where both annotators agree on the binary euphemism label. For agreed-positive instances, the annotated spans and literal meanings are used to construct the de-euphemized input variant. Cases with binary label disagreement are excluded from the main benchmark analysis to ensure annotation reliability. Because Buginese is annotated by a single annotator, we do not compute inter-annotator agreement for that language.

Overall, the final benchmark contains 1,050 aligned QA instances per language across Indonesian and six regional languages. Each instance is associated with its original Indonesian QA pair, translated question, topic label, euphemism label, euphemistic span(s) when applicable, and literal meaning(s). The agreement results indicate at least moderate reliability across languages with two annotators, with substantial agreement for Acehnese, Betawi, and Javanese, and near-perfect agreement for Indonesian.

3.2 Dataset Statistics

Table 2 summarizes euphemism use across the seven languages in EUPHEMISM-SRH. We analyze four complementary dimensions: overall prevalence, coverage across questions, lexical diversity, and euphemism density within euphemistic questions.

Prevalence. The **#Euph. Occ.** column reports the total number of annotated euphemism occurrences. Acehnese (3,993), Buginese (2,910), and Indonesian (1,597) have the highest counts, indicating that euphemistic expressions appear most

Language	#Euph. Occ.	#Euph. Qs	#Euph. Form	Avg. Freq.	Avg. Euph./Q
Acehnese	3993	763	120	33.28	3.42
Betawi	598	260	42	14.24	1.87
Buginese	2910	595	148	19.66	3.28
Javanese	526	169	139	3.78	2.51
Minangkabau	770	438	41	18.78	1.42
Sundanese	285	125	31	9.19	1.80
Indonesian	1597	366	109	14.65	2.93

Table 2: Dataset statistics across languages. **#Euph. Occ.** denotes the total number of euphemism occurrences per language. **#Euph. Qs** indicates the number of question instances containing at least one euphemism. **#Euph. Form** refers to the number of distinct euphemistic spans from the annotation. **Avg. Freq.** represent the average frequency of euphemisms. **Avg. Euph./Q** indicates the average number of euphemisms per euphemistic question.

frequently in these language versions of the dataset. These patterns may reflect sociocultural norms around indirectness, politeness, and taboo avoidance, particularly in sensitive domains such as women's SRH, though further sociolinguistic validation is needed. Prior work has documented the role of indirectness and taboo avoidance in Acehnese (Helmi and Liana), as well as politeness, honorifics, and implicit communication in Buginese and related Indonesian sociolinguistic contexts (Agus, 2017; Andini, 2021; Fatimah et al., 2025; Pawito et al., 2020; Widayati, 2020; Farhansyah et al., 2025).

Coverage across questions. The **#Euph. Qs** column reports the number of questions containing at least one euphemism. Acehnese (763), Buginese (595), and Minangkabau (438) show the broadest coverage, suggesting that euphemisms are distributed across many questions rather than appearing only in a small subset. Indonesian has fewer euphemistic questions (366) despite having the third-highest number of total euphemism occurrences. This suggests that, in Indonesian, euphemisms may be more concentrated in particular questions or topics rather than evenly distributed across the dataset.

Lexical diversity. The **#Euph. Form** column reports the number of distinct euphemistic expressions identified by annotators. Buginese has the largest number of distinct forms (148), followed by Javanese (139), Acehnese (120), and Indonesian (109). This indicates substantial variation in how euphemistic meaning is expressed across languages. By contrast, Sundanese has the fewest distinct forms (31). This lower count may reflect dataset-specific effects, translation choices, or broader patterns of language use; prior work has discussed concerns about Sundanese language attrition in some contexts (Sobarna, 2007; Marlia, 2020), but further validation is needed before drawing strong sociolinguistic conclusions.

Repetition and density. The **Avg. Freq.** column measures the average number of occurrences per distinct euphemistic form. Acehnese (33.28), Buginese (19.66), and Minangkabau (18.78) have the highest values, suggesting that certain euphemistic forms recur frequently in these language versions. Finally, **Avg. Euph./Q** measures the average number of euphemisms per euphemistic question.[4] Acehnese (3.42) and Buginese (3.28) have the highest values, indicating that when euphemisms appear in these languages, they often occur multiple times within the same question. In contrast, Minangkabau has the lowest value (1.42), suggesting that euphemistic questions in this language version tend to contain fewer euphemistic expressions per question.

Overall, the results show that euphemism usage in EUPHEMISM-SRH varies not only in frequency, but also in distribution, lexical diversity, and density. The consistent presence of euphemistic questions across all seven languages further suggests that euphemism is a systematic characteristic of women's SRH communication rather than a rare edge case, motivating language-sensitive evaluation of LLM robustness.

4 Task Setup

4.1 Benchmark Setup

For each question in EUPHEMISM-SRH, we prompt an LLM to generate a patient-facing medical response, which is then evaluated against the corresponding doctor-authored reference answer. The benchmark evaluates whether LLMs can maintain medical factuality when a patient's intended meaning is expressed indirectly through euphemistic language.

Prompt construction. We construct prompts from three components: *Main*, *Description*, and

[4] Computed only among questions containing at least one euphemism.

Language Forcing. The *Main* component instructs the model to answer clearly and politely, as if speaking directly to a patient, with the goal of eliciting answers similar to those a medical professional might provide in a question–answer exchange. The *Description* component specifies that the questioner is a woman from the relevant regional background, such as a Javanese woman for Javanese questions. This component is intended to examine whether geographic and cultural identity cues influence the responses generated by the LLM. The *Language Forcing* component instructs the model to respond in the same language as the input question. Its purpose is to assess the LLM's ability to answer patient queries in specific regional languages.

The preliminary experiment considers four prompt variants: **A**: *Main*; **B**: *Main + Description*; **C**: *Main + Language Forcing*; and **D**: *Main + Description + Language Forcing*. In the main benchmark, we use Prompt D because it provides both the relevant patient background and the target-language constraint. Full prompt components and templates are provided in Appendix H and Appendix I.

Model selection. The preliminary selection experiment compares three proprietary models, GPT-4o, Gemini-2.5-Flash, and Gemini-2.5-Pro, and one open-source model, SahabatAI-Llama3-8B.[5] The main benchmark uses Gemini-2.5-Flash, which achieved the best overall trade-off between factual precision and target-language adherence in the preliminary experiment. Details of the compared models and selection results are reported in Appendix C.

4.2 Evaluation Metrics

We evaluate model responses using a fact-based LLM-as-a-judge framework (Ben Abacha et al., 2023). Specifically, we use GPT-5-mini to decompose both the model-generated answer and the doctor-authored reference answer into atomic factual statements. The judge then compares the model facts against the reference facts to identify supported facts, unsupported facts, hallucinated facts, and critical omissions.

Based on these outputs, we compute four metrics:

- **Factual Precision:** $\frac{\#\textit{Correct Facts}}{\#\textit{System Facts}}$, measuring the proportion of model-generated facts supported by the reference answer.
- **Factual Recall:** $\frac{\#\textit{Correct Facts}}{\#\textit{Reference Facts}}$, measuring the proportion of reference facts covered by the model answer.
- **Hallucination Rate:** $\frac{\#\textit{Hallucinated Facts}}{\#\textit{System Facts}}$, measuring the proportion of model-generated facts that are unsupported by the reference answer and potentially harmful.
- **Omission Rate:** $\frac{\#\textit{Critical Omissions}}{\#\textit{Reference Facts}}$, measuring the proportion of essential reference facts missing from the model answer.

To assess the reliability of the LLM-as-a-judge framework, we compare its outputs against human annotation on a sample of 50 Indonesian-language instances. The results show moderate-to-strong agreement across the evaluated components, suggesting that the judge provides a useful proxy for reference-based factual evaluation, while leaving multilingual judge validation to future work (see Appendix D for detailed results).

5 Benchmarking LLMs on Euphemistic Medical QA

We use Euphemism-SRH as a benchmark for evaluating whether LLMs can preserve medical factuality when patients express sensitive health concerns indirectly. The experiments test two aspects of the benchmark gap: whether euphemistic phrasing changes model behavior, and whether explicit literalization improves responses.

Benchmark tasks. We define three evaluation settings. First, in *Raw Euphemistic QA*, the model answers the patient's original question when it contains one or more euphemistic expressions. Second, in *Non-euphemistic QA*, the model answers questions from the same dataset that do not contain annotated euphemisms. Third, in *De-euphemized QA*, each annotated euphemistic span is replaced with its annotator-provided literal meaning before inference. These settings allow us to test both naturally occurring differences between euphemistic and non-euphemistic questions, and a paired intervention in which only the euphemistic wording is made explicit. Moreover, we report uncorrected significance values and therefore interpret language-specific significance markers as descriptive rather than confirmatory across all language–metric comparisons.

Based on initial experiment results (see Ap-

[5] https://huggingface.co/Sahabat-AI

Lang.	#Instances		Avg. Factual Precision (↑)		Avg. Factual Recall (↑)		Avg. Hallucination Rate (↓)		Avg. Omission Rate (↓)	
	E	N	E	N	E	N	E	N	E	N
Acehnese	763	287	**0.358**	0.343	**0.318**	0.283	0.200	**0.192**	0.465	**0.462**
Betawi	260	790	0.304**	**0.339****	**0.273**	0.272	**0.200**	0.210	**0.437**	0.463
Buginese	595	455	**0.377**	0.372	**0.293**	0.246	**0.237**	**0.237**	0.419**	**0.373****
Javanese	169	881	0.334**	**0.368****	**0.323**	0.312	0.215	**0.192**	**0.455**	0.466
Minangkabau	438	612	0.298	**0.309**	0.304	**0.306**	0.212	**0.203**	0.519	**0.496**
Sundanese	125	925	0.270**	**0.301****	0.268**	**0.320****	0.194	**0.173**	**0.460****	0.507**
Indonesian	366	684	0.239	**0.253**	0.303	**0.324**	0.172**	**0.133****	**0.554**	0.579

Table 3: Fact-based evaluation results (LLM-as-a-Judge) for euphemistic (**E**) and non-euphemistic (**N**) questions. **#Instances** denotes the number of instances. Bold values indicate better performance within each language. Values marked with ** are statistically significant under Welch's t-test ($p < 0.05$).

pendix C), we use GEMINI-2.5-FLASH to conduct the full experiments. For each language, we construct prompts using variant D (Section 4.1), submit the prompt and question to GEMINI-2.5-FLASH to obtain a system-generated answer, translate the system response into Indonesian when necessary using GPT-4.1, and evaluate it against the doctor-authored reference using factual precision, factual recall, hallucination rate, and omission rate with our LLM-as-a-Judge framework (Section 4.2). The full prompt templates used in the experiments are provided in Appendix J.

5.1 RQ1: Do LLMs produce different factual, hallucinated, or incomplete answers when patient questions are euphemistic rather than non-euphemistic?

Table 3 reports LLM-as-a-Judge results for euphemistic and non-euphemistic questions across all languages. Higher factual precision and recall indicate better factual coverage, while lower hallucination and omission rates indicate fewer unsupported or missing facts. Bold values indicate the best score within each language, and ** marks statistically significant differences under Welch's t-test ($p < 0.05$).

The results show that the effect of euphemism is language-dependent. In some languages, non-euphemistic questions are associated with better factuality or lower hallucination rates, suggesting that explicit phrasing can help the model recover medically relevant information. For example, Betawi and Minangkabau show higher factual precision for non-euphemistic questions, and Indonesian shows a significantly lower hallucination rate under non-euphemistic phrasing.

However, euphemistic questions do not uniformly degrade performance. In several languages, euphemistic questions perform similarly to, or better than, non-euphemistic questions on factual precision or recall. This may occur because euphemistic phrasing reflects common local usage, making such expressions more natural or contextually informative in some language communities. Omission results are also mixed: some languages show lower omission rates for non-euphemistic questions, while others show the opposite pattern.

Taken together, these findings suggest that raw euphemism presence alone is not a uniform predictor of model failure. Euphemism can affect factuality, hallucination, and completeness, but its impact varies across languages and metrics. This motivates a more controlled intervention, where the same euphemistic questions are evaluated before and after replacing euphemistic spans with their literal meanings.

5.2 RQ2: Does replacing euphemistic expressions with literal meanings improve medical QA factuality?

To isolate the effect of euphemistic wording, we conduct a paired de-euphemization experiment. For each euphemistic question, we replace the annotated euphemistic span with its annotator-provided literal meaning before inference, while keeping the rest of the question unchanged. This setting directly tests whether making the intended medical meaning explicit improves the model's response.

Table 4 shows that de-euphemization (✓) consistently improves factual precision and factual recall across languages. It also reduces hallucination and omission rates in most settings. Because each de-euphemized question is paired with its original euphemistic version, we use the Wilcoxon signed-rank test to assess significance. The improvements are statistically significant across the evaluated metrics.

These results provide stronger evidence than the raw euphemistic versus non-euphemistic compar-

Lang.	Avg. Factual Precision (↑)		Avg. Factual Recall (↑)		Avg. Hallucination Rate (↓)		Avg. Omission Rate (↓)	
	✓	✗	✓	✗	✓	✗	✓	✗
Acehnese	**0.571****	0.358**	**0.550****	0.318**	**0.143****	0.200**	**0.295****	0.465**
Betawi	**0.650****	0.304**	**0.595****	0.273**	**0.094****	0.200**	**0.307****	0.437**
Buginese	**0.466****	0.377**	**0.383****	0.293**	**0.185****	0.237**	**0.230****	0.419**
Javanese	**0.628****	0.336**	**0.623****	0.323**	**0.103****	0.215**	**0.331****	0.455**
Minangkabau	**0.659****	0.298**	**0.677****	0.304**	**0.099****	0.212**	**0.371****	0.519**
Sundanese	**0.621****	0.270**	**0.649****	0.268**	**0.106****	0.194**	**0.370****	0.460**
Indonesian	**0.697****	0.239**	**0.878****	0.303**	**0.061****	0.172**	**0.554****	0.554**

Table 4: Fact-based evaluation results using an LLM-as-a-Judge for euphemistic questions with (✓) and without (✗) de-euphemization. Values marked with ** are statistically significant under a Wilcoxon signed-rank test ($p < 0.05$).

ison in RQ1. While RQ1 shows that naturally occurring euphemism effects are language-dependent, RQ2 shows that explicitly literalizing euphemistic expressions consistently improves model factuality. This suggests that at least part of the model's difficulty comes from recovering the intended clinical meaning of indirect or culturally grounded expressions, rather than from medical reasoning alone.

Qualitative error analysis. To better understand the improvements after de-euphemization, we conduct a paired qualitative audit of cases with the largest differences between euphemistic and de-euphemized inputs. We consider an error euphemism-mediated only when the incorrect, omitted, or unsupported fact is directly tied to an annotated euphemistic span, and replacing that span with its literal meaning resolves or mitigates the error in the paired output. The audit shows that some failures are plausibly euphemism-mediated. These include contradictions between model-generated facts and the doctor-authored reference, incorrect diagnoses that lead to misaligned reasoning, and plausible but unsupported responses. Other errors appear to reflect broader limitations of medical QA or reference-based evaluation, including syntactic mismatch despite semantic overlap, repetition of unsupported patient details, and unclear doctor responses. We therefore treat this qualitative analysis as diagnostic rather than standalone causal evidence, and provide representative examples in Appendix E.

LLM-based euphemism normalization. We additionally evaluate whether LLMs can identify euphemistic spans and generate their literal meanings. Current models remain weakly aligned with human annotators, and human-curated de-euphemization produces substantially stronger downstream QA performance than LLM-generated de-euphemization. Full results, including span-level detection, overprediction analysis, and qualitative examples, are provided in Appendix F.

6 Conclusion & Future Work

We present EUPHEMISM-SRH, a culturally grounded multilingual QA dataset for studying euphemistic language in women's sexual and reproductive health (SRH) consultations. The dataset contains patient–doctor QA instances across seven languages: Indonesian and six regional Indonesian languages, namely Javanese, Sundanese, Minangkabau, Acehnese, Betawi, and Buginese. Through native-speaker translation and annotation, EUPHEMISM-SRH captures euphemism presence, spans, and literal meanings, enabling systematic evaluation of how culturally situated indirect expressions affect medical QA. Our analysis reveals substantial cross-lingual variation in euphemism frequency, diversity, and density, reflecting differences in local communicative norms, politeness strategies, and sociocultural sensitivities surrounding women's SRH. These findings suggest that euphemism is not a uniform linguistic phenomenon, but varies considerably across languages and topics. Benchmark experiments further show that euphemism constitutes a meaningful challenge for medical QA. While the impact of euphemistic versus non-euphemistic phrasing is language-dependent, replacing euphemistic expressions with literal meanings consistently improves factual precision and recall while reducing hallucination and omission rates. This suggests that current LLMs often struggle to recover intended medical meaning from euphemistic language, and that evaluating only explicit patient questions may overestimate model readiness for real-world SRH communication. Overall, our findings highlight the need for more robust and culturally aware medical NLP systems capable of reasoning over euphemistic language in sensitive healthcare settings.

Limitations

Dataset Size. Our dataset comprises 1,050 QA instances spanning 42 categories of women's sexual and reproductive health (SRH), across seven languages, and includes annotations for both euphemistic expressions and their literal counterparts. Empirically, this scale is sufficient to reveal consistent behavioral patterns in LLMs when responding to patient queries in a clinically grounded manner. Nevertheless, scaling the dataset could further enhance the robustness of the observed trends and improve generalizability.

Annotation Coverage. For Buginese, euphemism annotation was conducted by a single native-speaker annotator due to resource constraints, so we cannot compute inter-annotator agreement for this language. Although the annotator satisfied the same language-qualification criteria as annotators for the other languages, Buginese-specific euphemism statistics and de-euphemization results should be interpreted with this annotation-coverage limitation in mind. Future work will add additional Buginese annotators to validate span-level annotations and literal meanings.

Scope. Our investigation is intentionally scoped to a relatively underexplored domain–euphemistic language in women's SRH–motivated by its linguistic and societal sensitivity. While this focused setting enables a controlled analysis, extending the scope to adjacent health-related domains (e.g., euphemisms in mental health discourse or general medical complaints) would allow for a broader characterization of how euphemistic phenomena influence LLM responses. In addition, the benchmark captures culturally specific settings in several Indonesian regions. Future work should test whether similar euphemism-driven failures occur across other languages and regions.

Metric Evaluation. We employ evaluation metrics that quantify Factual Precision, Factual Recall, Hallucination Rate, and Omission Rate by comparing LLM outputs against physician-provided references for identical patient queries. These metrics provide a reliable estimate of performance gaps and exhibit strong alignment with human judgment. However, future work should incorporate metrics that explicitly account for the patient's original utterance, enabling a more holistic and context-sensitive evaluation framework.

Models. Due to computational constraints, our experiments are limited to models with moderate parameter scales. These models were selected to ensure adequate multilingual coverage aligned with the study's scope, including Indonesian regional languages (e.g., Javanese and Sundanese). Exploring larger-scale open-source models with comparable linguistic coverage remains a promising direction, with the potential to yield additional insights and further substantiate the findings presented in this work.

Ethics and IRB applicability. This work uses publicly available, non-identifiable doctor–patient QA content. We do not collect private information from patients, attempt to contact them, or attempt to re-identify any individual. The released dataset contains only anonymized/non-identifiable text and derived annotations. Human language annotators were engaged as paid contractors to provide translation and euphemism annotation services; they were not the subjects of the study, and the research does not analyze their personal characteristics, behavior, or health information. Accordingly, we treat this work as not involving human-subjects research requiring IRB/HREC review; authors will follow institutional guidance where formal determination is required. We nevertheless applied safeguards appropriate for sensitive health-language data, including data minimization, non-reidentification, anonymized release, and non-commercial research licensing.

Potential Risks

This work may pose several potential risks. Since the dataset focuses on euphemistic expressions in women's sexual and reproductive health, LLMs may generate inaccurate medical interpretations when failing to understand culturally grounded or indirect language, potentially leading to misinformation. There are also potential fairness and representation risks, as the dataset is limited to seven languages and culturally specific Indonesian contexts. Although the dataset uses anonymized public data, the sensitive nature of sexual and reproductive health discussions still raises potential privacy and misuse concerns. To mitigate these risks, the released portion dataset is released only in anonymized form and under a non-commercial research license.

Generative AI Usage Statement

All authors confirm that generative AI tools were used in a limited and supportive capacity during the manuscript preparation process. Specifically, ChatGPT (version 5.2) was employed to assist with grammar correction, stylistic refinement, sentence restructuring for clarity, and overall language polishing. The tool was also used to suggest alternative phrasings to improve fluency and coherence. Importantly, generative AI was not used to generate scientific content, experimental results, analyses, interpretations, or conclusions. All ideas, methods, experimental designs, results, and discussions presented in this paper are the original work of the authors. The use of generative AI did not replace human authorship, intellectual contribution, or scholarly judgment. All authors take full responsibility for the originality, accuracy, integrity, and ethical compliance of the manuscript.

Acknowledgement

According to the Alodokter platform's Terms and Conditions (last updated 30/10/2025), reproduction and related uses are permitted for non-commercial purposes. Accordingly, the anonymized portion of our dataset will be released upon acceptance under a non-commercial research license (e.g., CC BY-NC 4.0 or equivalent).

References

Nuraidar Agus. 2017. Bentuk eufemisme dalam pertuturan bahasa bugis. *Sawerigading*, 18(1):25–34.

Frank B Agyei and Doreen K Kaura. 2023. A systematic review of effective parent-adolescent sexual and reproductive health information communication in lower-and middle-income countries. *Health SA Gesondheid (Online)*, 28:1–9.

Citra Andini. 2021. The use of honorifics in english and buginese with special reference to bone language: A comparative study. *International Journal of Innovative Science and Research Technology*, 6(7):873–877.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Stella Bullo and Annalise Weckesser. 2021. Addressing challenges in endometriosis pain communication between patients and doctors: the role of language. *Frontiers in Global Women's Health*, 2:764693.

Mia Cate. 2025. Improving communication between healthcare providers and patients -focusing on transparency, empathy, and accessibility of medical information.

Ayesha Izhar Chaudhri, Tazanfal Tehseem, and Barirah Nazir. 2022. A sociolinguistic study of taboos and euphemisms surrounding pakistani females' daily issues. *Global Social Sciences Review*, 7(2):34–45.

Wen-ying Sylvia Chou, Lauren M Hamel, Chan L Thai, David Debono, Robert A Chapman, Terrance L Albrecht, Louis A Penner, and Susan Eggly. 2017. Discussing prognosis and treatment goals with patients with advanced cancer: A qualitative analysis of oncologists' language. *Health Expectations*, 20(5):1073–1080.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Kristin N Dew, Anne M Turner, Yong K Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, 85:56–67.

Toochukwu John Ezeugo and Ephraim A Chukwu. 2024. Euphemism as a discursive strategy in doctor-patient encounter. *AWKA JOURNAL OF ENGLISH LANGUAGE AND LITERARY STUDIES*, 11(2):148–169.

Mohammad Rifqi Farhansyah, Iwan Darmawan, Adryan Kusumawardhana, Genta Indra Winata, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. Do language models understand honorific systems in Javanese? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26732–26754, Vienna, Austria. Association for Computational Linguistics.

Ainun Fatimah, Andi Wardatul Wahidah Lufini, and Kamsinah Kamsinah. 2025. Eksistensi eufemisme dalam bahasa bugis untuk diksi sensitif sebagai upaya menjaga harmoni: Perspektif sosiolinguistik. *Kongres Internasional Masyarakat Linguistik Indonesia*, 2025.

Todd Firsich and Anthony Rios. 2024. Can GPT4 detect euphemisms across multiple languages? In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 65–72, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Rahil Helmi and Putri Liana. Bahasa tabu dalam bahasa aceh.

Ameemah Humayun, Bushra Zubair, and Maryam Mustafa. 2025. Between myths and metaphors: Rethinking llms for srh in conservative contexts. *arXiv preprint arXiv:2511.01907*.

Eglantina Hysi. 2011. Aspects of taboos and euphemisms in women's language. *Mediterranean Journal of Social Sciences*, 2(3):379–383.

Denis Jamet. 2018. The neological functions of disease euphemisms in english and french: Verbal hygiene or speech pathology? *Lexis. Journal in English Lexicology*, (12).

Otto Jespersen. 2013. *Language: its nature and development*. Routledge.

Anna Kapron-King and Yang Xu. 2021. A diachronic evaluation of gender asymmetry in euphemism. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 28–38, Online. Association for Computational Linguistics.

Shoira T Khaydarova. 2024. Euphemisms: Their scope, functions, and applications in everyday communication. *Mental Enlightenment Scientific-Methodological Journal*, 5(06):64–75.

Venera Khisamova and Liliia Abdullina. 2022. The problem of euphemisms in the medical terminology of the english and tatar languages. *Journal of Educational and Social Research*, 12(1):14–23.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2022. Adversarial perturbations augmented language models for euphemism identification. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 154–159, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Robin Lakoff. 1977. *6. You say what you are: Acceptability and Gender-related Language*, pages 73–86. De Gruyter Mouton, Berlin, Boston.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

Marlia Marlia. 2020. Penggunaan bahasa sunda di lingkungan keluarga sebagai upaya konservasi budaya. *Artikulasi: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 2(2):156–166.

Akin Odebunmi. 2010. Code selection at first meetings: A pragmatic analysis of doctor-client conversations in nigeria. *InLiSt*, 48:1–41.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Pawito Pawito, Widodo Muktiyo, and Hamid Arifin. 2020. Nilai budaya dan gaya komunikasi warga minangkabau, jawa, dan bugis. *Jurnal Ilmu Komunikasi*, 17(3):249–261.

Jan Rodríguez-Miret, Eulàlia Farré-Maduell, Salvador Lima-López, Laura Vigil, Vicent Briva-Iglesias, and Martin Krallinger. 2024. Exploring the potential of neural machine translation for cross-language clinical natural language processing (nlp) resource generation through annotation projection. *Information*, 15(10):585.

Julia Sammartino, Libby Barak, Jing Peng, and Anna Feldman. 2025. When does language transfer help? sequential fine-tuning for cross-lingual euphemism detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 1058–1065.

Anke Samulowitz, Ida Gremyr, Erik Eriksson, and Gunnel Hensing. 2018. "brave men" and "emotional women": A theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain. *Pain research and management*, 2018(1):6358624.

Cece Sobarna. 2007. Bahasa sunda sudah di ambang pintu kematiankah? *Makara Human Behavior Studies in Asia*, 11(1):13–17.

Michael Tayler and Jane Ogden. 2005. Doctors' use of euphemisms and their impact on patients' beliefs about health: an experimental study of heart failure. *Patient education and counseling*, 57(3):321–326.

Amita Tuteja, Lena Sanci, Lester Mascarenhas, Elisha May Riggs, Dianne Van-Vliet, Katrina Sangster, Kimberley McGuinness, and Meredith Temple-Smith. 2022. Negotiating language in providing sexual and reproductive health care to refugees from burma: experiences of service providers from australia. *Women's Reproductive Health*, 9(1):57–79.

Mukti Widayati. 2020. The uses of honorifics by president candidate in the campaign news text on national newspapers. *International Journal of English Linguistics, Literature, and Education (IJELLE)*, 2(2):116–130.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, and 1 others. 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.

Zelin Yan, Jingwen Liu, Yihong Fan, Shiyuan Lu, Dingting Xu, Yun Yang, Honggang Wang, Jie Mao, Hou-Chiang Tseng, Tao-Hsing Chang, and 1 others. 2025. Ability of chatgpt to replace doctors in patient education: cross-sectional comparative analysis of inflammatory bowel disease. *Journal of Medical Internet Research*, 27:e62857.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA network open*, 7(3):e240357–e240357.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. A survey of large language models in medicine: Progress, application, and challenge. *ArXiv*, abs/2311.05112.

A Annotators' Profile

Table 5 summarizes all language contractors involved in translation and/or euphemism annotation. For Buginese, two contractors were involved overall, but only one completed the euphemism annotation stage; therefore, inter-annotator agreement is not reported for Buginese (Table1). We report only language-relevant eligibility information to document annotator expertise. All annotators were provided with fair and standardized compensation, aligned with prevailing wage standards in Indonesia.

Language	# Annotators	Min. active use
Acehnese	2	25 years
Betawi	2	15 years
Buginese	2	29 years
Javanese	2	25 years
Minangkabau	2	26 years
Sundanese	2	16 years
Indonesian	2	34 years

Table 5: Aggregate language qualifications of language annotators. We report only the minimum active-use experience for each language group, rather than individual demographic profiles, because annotators were contracted to provide language services and are not the subjects of analysis in this study.

B Verification of Annotator Translation Results

Table 6 presents similarity scores based on several evaluation metrics between human annotators' translations and machine translation outputs for each week (Weeks 1-3). Specifically, we use Google Translate as a reference for Javanese, Sundanese, Minangkabau, Buginese, and Acehnese, while GPT-4 is used for Betawi due to limited support in the first platform. The relatively low similarity scores–in every weeks' result– suggest that annotators were unlikely to rely on machine translation during the translation process. Therefore, the resulting translations can be considered more reliable.

C Initial Experiment

Given the scale of our full annotated dataset (1,050 questions across seven languages) and the combinatorial space induced by four prompt variants and four LLMs, we first run a smaller set of experiments to improve computational and resource efficiency before conducting deeper analyses of euphemism effects. Specifically, we compute *Factual*

Lang.	Week 1				Week 2				Week 3			
	BLEU	METEOR	R-1	R-L	BLEU	METEOR	R-1	R-L	BLEU	METEOR	R-1	R-L
Acehnese	0.1464	0.4456	0.5430	0.5033	0.1421	0.4403	0.5287	0.5033	0.1459	0.4417	0.5417	0.5033
Betawi	0.0632	0.2857	0.3216	0.3088	0.0608	0.2843	0.3228	0.3114	0.0622	0.2932	0.3320	0.3182
Buginese	0.0240	0.1551	0.2369	0.2034	0.0256	0.1649	0.2364	0.2066	0.0250	0.1585	0.2330	0.2035
Javanese	0.0682	0.3121	0.3807	0.3604	0.0785	0.3295	0.3801	0.3616	0.0727	0.3209	0.3751	0.3557
Minangkabau	0.1257	0.4346	0.5163	0.4879	0.1238	0.4289	0.5147	0.4867	0.1335	0.4415	0.5281	0.5012
Sundanese	0.0502	0.2654	0.3468	0.3264	0.0438	0.2393	0.3220	0.2996	0.0477	0.2565	0.3342	0.3111

Table 6: Similarity scores between human and machine translations across Weeks 1–3 using BLEU, METEOR, ROUGE-1 (R-1), and ROUGE-L (R-L). Consistently low scores suggest limited reliance on machine translation and more reliable human outputs.

Prompt Type	Model			
	GEMINI-2.5-FLASH	GEMINI-2.5-PRO	GPT-4O	SAHABATAI-LLAMA3
A	0.402	0.356	0.229	0.151
B	0.396	0.365	0.243	0.147
C	0.405	0.347	0.230	0.150
D	0.394	0.362	0.238	0.161

Table 7: Average factual precision scores from the initial experiments across all model–prompt combinations.

Precision to identify the most effective prompt–model combination using a subset of the data. We use an Indonesian question subset of 215 instances, sampled via stratified sampling to preserve the category distribution.

Table 7 reports results for all model–prompt combinations, where color intensity is proportional to the magnitude of the average Factual Precision score of the system's answers to the 215 questions. Overall, GEMINI-2.5-FLASH achieves the highest performance among the evaluated models. Prompt D also exhibits the most stable performance across models, yielding consistently higher factual precision. Although Prompt A and Prompt C produce slightly higher Factual Precision for GEMINI-2.5-FLASH, these differences are not statistically significant. Moreover, Prompt D outperforms Prompt A and Prompt C in its ability to produce responses in the same regional language as the input question (Figure 2).

Based on these initial experiments, we select GEMINI-2.5-FLASH as the best-performing model and Prompt D as the preferred prompt configuration for subsequent large-scale multilingual experiments.

D LLM-as-a-Judge Reliability

We compare the evaluations produced by the LLM-as-a-Judge framework with those of human annotators by sampling 50 Indonesian-language instances (approximately ±5% of the dataset). Table 8 reports the correlation between the two evaluation sources. The results indicate a relatively strong degree of similarity, suggesting that the LLM-as-a-Judge evaluator demonstrates high reliability and exhibits a substantial linear correlation with human preferences.

Component	Similarity (%)	Agreement	
		Cohen's κ	Krippendorff's α
Reference Facts	76.28	–	–
Common Facts	84.46	0.602	0.603
Omission Rate	90.44	0.807	0.806
Hallucination Rate	95.13	0.881	0.881

Table 8: Similarity scores alongside inter-annotator agreement metrics (Cohen's κ and Krippendorff's α) between LLM-as-a-Judge and a Human Annotator. Higher values indicate stronger alignment with human judgments.

E Qualitative Analysis of Errors in Responses to Euphemistic Questions.

Several recurring error patterns are observed in model responses to euphemistic queries. These errors contribute to degraded performance across evaluation components, leading to a significant drop in overall scores compared to de-euphemistic inputs. To further investigate this phenomenon, we compare the bottom-50 instances for factual precision and recall, as well as the top-50 instances for hallucination and omission rates from euphemistic queries, against the top-50 instances for factual precision and recall and the bottom-50 instances for hallucination and omission rates from de-euphemized queries (from Indonesian subset). Based on this analysis, we identify several notable and recurring error patterns in LLM responses to euphemistic questions, which provide deeper insight into the models' limitations in handling indi-

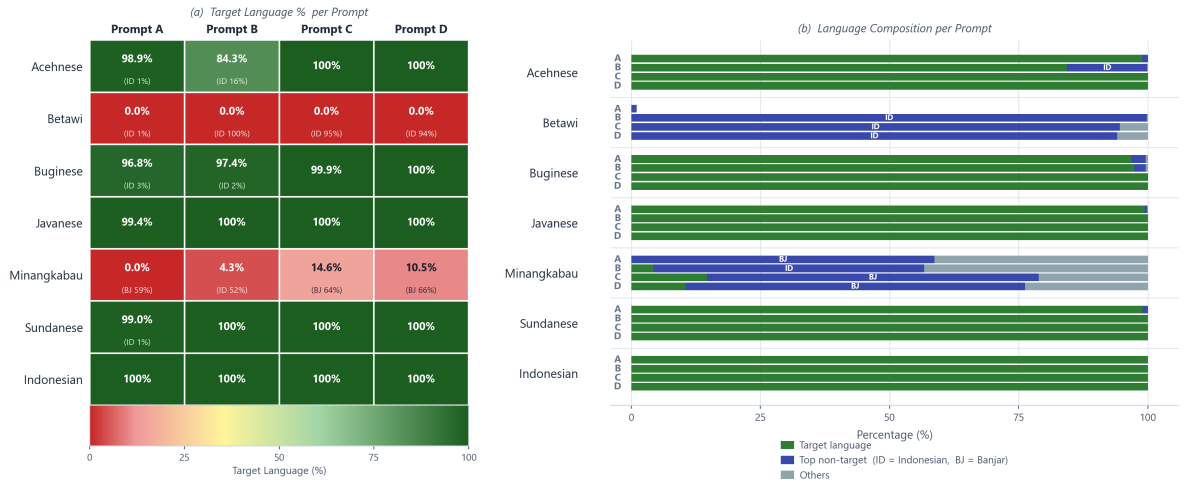

Figure 2: LID Result with FastText for GEMINI-2.5-FLASH's answers.

rect and culturally nuanced expressions. Subsections E.1, E.2, E.3, E.4, E.5, and E.6 present representative examples of these error patterns.

E.1 Contradiction Between LLM-Generated Facts and Reference Response

Figure 3 demonstrates a clear contradiction between the doctor's statement and the facts generated by the LLM. In the doctor's response, it is stated that, *"...Abdominal pain during sexual intercourse, as you are experiencing, is unlikely to be related to medication use or a fertility program you may be undergoing..."*. In contrast, the LLM-generated fact asserts that, *"The abdominal cramping you are experiencing may be a side effect of the fertility medication you are currently taking."*. Thus, while the LLM attributes the symptom to fertility medication, the doctor explicitly indicates that the two are not related.

E.2 Incorrect Diagnosis Leading to Misaligned Reasoning

Figure 4 shows that the LLM prematurely narrows the diagnosis by attributing the patient's *dysmenorrhea* directly to endometriosis. In contrast, the doctor's response indicates that lower abdominal pain may be associated with a range of possible conditions. This overly specific diagnostic inference by the LLM, made without adequately considering alternative causes, leads to a disproportionately focused discussion on endometriosis. As a result, the generated facts deviate from the broader and more balanced explanation provided in the doctor's response.

E.3 Plausible but Unsupported Response

In Figure 5, the LLM generates a response discussing ectopic pregnancy and urinary tract infection (UTI) as potential causes of the patient's abdominal pain. While these explanations may be medically valid, the discussion extends beyond the scope of the doctor's original statement. As a result, the evaluator considers this response to be less appropriate.

E.4 Syntactic Form Mismatch Despite Semantic Presence

In Figure 6, the doctor explains blackish menstrual blood with the statement, *"... Blackish menstrual blood can result from older blood that is expelled more slowly or contains a higher proportion of dead cells..."*. Meanwhile, the LLM-generated facts include a semantically similar statement with a slightly different surface realization, *"Brown menstrual blood indicates older blood or a slower flow, allowing it to oxidize"*. Despite their close semantic alignment but with different syntactic form, these are treated as distinct statements, leading the fact to be categorized as an unsupported fact.

E.5 LLM Repeats Unsupported Patient Details

Figure 7 shows that the LLM introduces facts that are not present in the doctor's response. These facts originate from the patient's description of their condition and are subsequently treated by the model as important information to be reiterated in its answer. However, the evaluator considers this behavior inappropriate, as the doctor's reference response does not include these details. This observation further

highlights the need for more reliable evaluation metrics in future work–ones that can jointly account for the patient's query, the doctor's response, and the LLM-generated answer.

E.6 Unclear Doctor's Response

Figure 8 shows that the doctor's response simply refers to a previous post, indicating that the answer has already been provided elsewhere. As a result, none of the LLM-generated responses can be directly supported by the doctor's statement. However, this case occurs in only three instances–(1) Vagina with ID 303, (2) Vagina with ID 671, and (3) Vaginitis with ID 800–and therefore is unlikely to substantially affect the overall performance measured by the evaluation metrics used.

F LLM-based Euphemism Annotation

F.1 Impromptu Euphemism Detection

Table 12 presents the performance comparison of all evaluated models in the impromptu euphemism detection task across Indonesian + 6 regional languages. The evaluation employs span-based Precision, Recall, and F1-score metrics to measure the correctness of euphemism span production generated by each model. Higher values in these metrics indicate that the produced euphemism spans are more aligned with the human-annotated references. In addition, the BLEU score is used to evaluate the correctness of direct term generation corresponding to the detected euphemisms. A higher BLEU score indicates that the generated direct terms are semantically and lexically closer to the reference annotations provided by human annotators.

To further analyze disagreement patterns between human annotations and model predictions, False Positive (FP) Rate and False Negative (FN) Rate are also reported. The FP Rate indicates the proportion of generated euphemism spans that do not correspond to the annotated references, while the FN Rate represents the proportion of annotated euphemism spans that were not detected by the model. Within each model group, underlined values indicate the highest score among all evaluated languages for a particular metric. Meanwhile, bold values represent the overall best performance across all languages and all evaluated models for the corresponding metric.

Additionally, the evaluation also considers tautology occurrences. In this context, tautology refers to cases where a euphemistic expression is assigned a direct term that is identical to the euphemism itself, resulting in no semantic transformation between the euphemism and its intended direct meaning. The presence of tautological outputs may indicate two possible issues: (1) the model does not sufficiently understand the conceptual distinction between euphemistic and direct expressions, or (2) the model struggles to infer the appropriate direct meaning of the euphemism. Therefore, lower tautology counts indicate better semantic understanding and euphemism interpretation capability.

Based on the overall span-based evaluation metrics, FP and FN disagreement rates, and tautology analysis, GEMINI-2.5-PRO demonstrates the most balanced and reliable performance for the impromptu euphemism detection task. The model consistently achieves the highest overall Precision, Recall, F1-score, and BLEU scores while simultaneously producing zero tautological outputs across all evaluated languages, indicating superior understanding of euphemistic semantics and direct-term generation.

F.2 Agreement with Human Annotators

Lang.	Agree		Disagree		Agreement		
	TP	TN	FP	FN	Rate (%)	Cohen κ	Krippendorf α
Acehnese	705	166	86	93	83.0	0.537	0.537
Betawi	257	151	631	11	38.9	0.086	-0.235
Buginese	629	93	217	111	68.8	0.167	0.156
Javanese	295	163	545	47	43.6	0.067	-0.145
Minangkabau	501	171	339	39	64.0	0.267	0.202
Sundanese	549	135	219	147	65.1	0.179	0.175
Indonesian	347	382	285	36	69.4	0.421	0.388

Table 9: Agreement between LLM-based and Human-curated euphemism detection.

Tabel 9 presents the agreement scores between LLM-based (GEMINI-2.5-PRO) and human-curated euphemism detection. True Positive (TP) denotes instances where both the LLM and human annotators agree that a question is euphemistic, whereas True Negative (TN) denotes instances where both agree that a question is non-euphemistic. In contrast, False Positive (FP) refers to instances identified as euphemistic by the LLM but not by human annotators, while False Negative (FN) refers to instances identified as non-euphemistic by the LLM despite being labeled as euphemistic by human annotators. The results indicate that most language settings—Betawi, Buginese, Javanese, and Sundanese—exhibit only *slight agreement* between LLM-based and human-curated annotations. Minangkabau demonstrates *fair agreement*, while Acehnese and Indonesian

achieve *moderate agreement*. Overall, these findings suggest that current LLM-based euphemism detection systems remain insufficiently aligned with human judgment and interpretation of euphemistic expressions.

F.3 Euphemism Overprediction

Tabel 13 presents a comparative analysis of inference results between instances labeled as non-euphemistic by human annotators and the subset of those instances incorrectly identified as euphemistic by the LLM (GEMINI-2.5-PRO), which we refer to as the *Overpredict* phenomenon. The results reveal a consistent trend in which overprediction tends to increase factual precision while simultaneously decreasing factual recall. In addition, overprediction generally leads to higher hallucination rates and lower omission rates. Most metric differences are also statistically significant under Wilcoxon's signed-rank test.

F.4 Overprediction Yields Shorter Responses

Lang.	Non-Euph Qs.	Overpredict	Delta (%)
Acehnese	18	12	-28.64
Betawi	18	12	-28.50
Buginese	15	10	-27.81
Javanese	18	13	-27.01
Minangkabau	22	15	-30.07
Sundanese	21	15	-29.59
Indonesian	26	18	-28.49

Table 10: Comparison of the average number of generated #SystemFacts between human-labeled non-euphemistic questions (**Non-Euph Qs.**) and LLM-overpredicted euphemistic questions (**Overpredict**).

Tabel 10 shows that LLM inference outputs consistently exhibit a reduction in the average number of generated #SystemFacts across all language settings when the inputs correspond to overpredicted questions produced by GEMINI-2.5-PRO. This behavior suggests that the generated responses tend to be substantially shorter compared to those produced from the original non-euphemistic questions. Nevertheless, despite being shorter, the responses are generally more targeted and concise, as evidenced by the consistent improvement in Factual Precision (Appendix F.3).

F.5 Behavioral Effects of Euphemism Overprediction

Overprediction affects LLM inference behavior by producing responses that are consistently shorter yet more on-point (see Appendix F.3 and Appendix F.4). Through qualitative inspection, we identify two distinctive phenomena induced by overprediction, which we hypothesize contribute to this conciseness–precision trade-off.

Overreach. Terms not annotated as euphemistic by human annotators are incorrectly interpreted as euphemisms, leading the model to generate literal substitutions based on unsupported assumptions.

- *Example (ID: 14, Indonesian, topic: hormone):* The phrase "orang gemuk" (people with higher body weight) is not annotated as a euphemism by human annotators, but is identified as euphemistic by GEMINI-2.5-PRO. The model maps it to "obesitas" (obesity), instead of the more neutral "kelebihan berat badan" (overweight). This appears to be driven by contextual inference from the dialogue (e.g., "...berat badan sy (120kg)..."), which biases the model toward a clinical classification.

Medicalization. Terms not considered euphemistic by human annotators are nevertheless interpreted as euphemisms by the LLM, after which the generated direct term is replaced with a medically formalized or clinical expression.

- *Example 1 (ID: 24, Indonesian, topic: pregnant mother):* The term "cantengan" (ingrown toenail) is not labeled as euphemistic by human annotators but is interpreted as euphemistic by GEMINI-2.5-PRO, which maps it to "paronikia".
- *Example 2 (ID: 135, Indonesian, topic: breast cancer):* The term "penyebarannya" (its spread) is not annotated as euphemistic by humans, yet is treated as euphemistic by GEMINI-2.5-PRO, which substitutes it with "metastasis".

F.6 Can LLMs detect euphemisms and their direct terms in a human-aligned manner?

Based on the impromptu euphemism detection evaluation (Appendix F.1), GEMINI-2.5-PRO achieved the strongest overall performance. Accordingly, its euphemism annotations are used as the primary point of comparison against human annotators. Table 11 presents the comparative performance of de-euphemization inference's results derived from human and LLM annotations. Because the set of euphemistic questions identified by human annotators

Language	De-Euph. by	Avg. Factual Precision (↑)	Avg. Factual Recall (↑)	Avg. Hallucination Rate (↓)	Avg. Omission Rate (↓)
Acehnese	HUMAN	**0.571****	**0.550****	**0.143****	**0.295****
Acehnese	LLM	0.376**	0.228**	0.217**	0.323**
Betawi	HUMAN	**0.650****	**0.595****	**0.094****	0.307
Betawi	LLM	0.330**	0.212**	0.226**	**0.294**
Buginese	HUMAN	**0.466****	**0.383****	**0.185****	**0.230****
Buginese	LLM	0.372**	0.213**	0.227**	0.269**
Javanese	HUMAN	**0.628****	**0.623****	**0.103****	**0.331****
Javanese	LLM	0.352**	0.261**	0.190**	0.358**
Minangkabau	HUMAN	**0.659****	**0.677****	**0.099****	**0.371**
Minangkabau	LLM	0.328**	0.229**	0.252**	0.374
Sundanese	HUMAN	**0.621****	**0.649****	**0.106****	**0.370**
Sundanese	LLM	0.315**	0.225**	0.198**	0.395
Indonesian	HUMAN	**0.697****	**0.878****	**0.061****	**0.554**
Indonesian	LLM	0.286**	0.263**	0.250**	0.493

Table 11: Fact-based evaluation results using an LLM-as-a-Judge for euphemistic questions with human (HUMAN) vs. LLMs (LLM) de-euphemization on questions labeled as euphemistic by human annotators. If LLM failed to identify a human-labeled euphemistic question, the original plain question was used without de-euphemization. Values marked with ** are statistically significant under Wilcoxon's signed-rank test ($p < 0.05$).

differs from that identified by GEMINI-2.5-PRO (Appendix F.2), let A denote the set of instances labeled as euphemistic by human annotators, and let B denote the set of instances labeled as euphemistic by GEMINI-2.5-PRO. The evaluation reported in Table 11 is computed over the left-join intersection between A and B. Specifically, for instances $a \in A$ where $a \notin B$, the corresponding performance value is replaced with the inference result obtained from the original question without de-euphemization. The results show that human-curated de-euphemization consistently and significantly outperforms LLM-based de-euphemization across nearly all evaluation metrics and language settings.

As appeared in Appendix F.3, we also analyze the subset of instances identified as euphemistic by GEMINI-2.5-PRO but not by human annotators, which we refer to as *Overpredict* cases. Let C denote the set of instances labeled as non-euphemistic by human annotators. We then evaluate the subset $C \cap B$. The results indicate that overprediction tends to improve factual precision and reduce hallucination rates, while simultaneously lowering factual recall and increasing omission rates. We hypothesize that this behavior arises because overprediction encourages the model to produce shorter yet more targeted responses, as reflected by the reduced average number of generated System-Facts (Appendix F.4). Further qualitative inspection reveals two distinctive phenomena associated with overprediction (Appendix F.5), Overreach and Medicalization. Collectively, these findings suggest that current LLMs still struggle to detect and interpret euphemistic expressions in a manner that is consistently aligned with human judgment.

G Euphemism Annotation Platform

On the euphemism annotation platform, annotators may encounter four possible interface states during the annotation process:

Initial View. Before selecting either option in the binary decision ("Yes" / "No"), annotators are presented with the initial interface as shown in Figure 9.

Negative Selection ("No"). If the annotator selects **"No"**, indicating that no euphemism is present, no additional interface elements are displayed. The annotator can directly proceed to the next question by clicking the *Submit* button (Figure 10).

Positive Selection ("Yes"). If the annotator selects **"Yes"**, indicating the presence of a euphemism, an additional interface panel is displayed to enable span selection and annotation (Figure 11).

Euphemism Span Highlighting and Literal Interpretation. Annotators are then required to highlight all spans corresponding to euphemistic expressions and provide their literal meanings (Figure 12).

H Prompt's Component

Table 14 reports the prompt components used for each language, including the *Description*, *Lan-*

guage Forcing, *Main* instruction, and the *Question Indicator*.

I Prompt Variation Template

Based on the prompt construction rules described in Section 4.1, Figure 13 illustrates the template for Prompt A, while Figures 14, 15, and 16 present the templates for Prompts B, C, and D, respectively.

J Experiment's Prompt

For the experiments described in Section 5, we employ four prompt variations corresponding to different stages of the evaluation pipeline:

Translation. This prompt is used to instruct GPT-4.1 to translate the responses generated by GEMINI-2.5-FLASH into Indonesian, ensuring alignment with the Indonesian-language doctor references (Figure 17).

Common-Facts Extractor. This prompt is used to extract *correct* and *incorrect* facts from the LLM-generated responses by comparing them against the corresponding doctor answers, which serve as reference facts (Figure 18).

Hallucinated-Facts Extractor. This prompt is designed to identify and extract hallucinated facts, i.e., incorrect facts that are semantically deviated from the reference or contain potentially harmful information (Figure 19).

Critical-Omission Extractor. This prompt is used to extract critical information present in the doctor answers that is omitted in the model-generated responses, capturing clinically important content that the model fails to mention (Figure 20).

Impromptu Euphemism Detection. This prompt is used to extract euphemisms and its direct terms for each question of the dataset (Figure 21).

K Computational Resources

We utilized four A40 40GB GPUs for the experiments, particularly for running the open-source model (SAHABATAI-LLAMA3-8B). Overall, however, the majority of our evaluations were conducted using proprietary models. Computational resources were also allocated for post-processing and analyzing the outputs generated by these proprietary systems.

Model	Lang.	Precision	Recall	F1	BLEU	FP (%)	FN (%)	#Tautology
GPT-4o	Acehnese	0.485	0.421	0.410	0.041	85.02	1.44	11
	Betawi	0.156	0.200	0.161	0.069	92.41	0.38	0
	Buginese	0.348	0.292	0.284	0.091	79.12	6.55	4
	Indonesia	0.498	0.470	0.465	0.291	60.67	3.83	1
	Javanese	0.072	0.091	0.075	0.052	93.87	1.78	0
	Minangkabau	0.244	0.395	0.278	0.056	90.36	0.46	3
	Sundanese	0.087	0.090	0.085	0.059	93.30	2.40	0
GEMINI-2.5-FLASH	Acehnese	0.642	0.364	0.424	0.105	61.67	6.42	4
	Betawi	0.131	0.117	0.117	0.041	95.70	1.92	0
	Buginese	0.418	0.278	0.313	0.134	69.01	19.66	4
	Indonesia	0.536	0.462	0.479	0.322	54.97	5.19	4
	Javanese	0.081	0.074	0.075	0.063	92.51	4.73	0
	Minangkabau	0.265	0.289	0.266	0.094	83.82	1.60	0
	Sundanese	0.139	0.124	0.128	0.104	88.22	4.80	0
GEMINI-2.5-PRO	Acehnese	**0.697**	0.481	0.532	0.176	35.54	9.44	0
	Betawi	0.236	0.290	0.246	0.148	81.01	3.08	0
	Buginese	0.388	0.324	0.334	0.103	76.26	15.97	0
	Indonesia	0.587	**0.587**	**0.572**	**0.396**	**43.86**	8.47	0
	Javanese	0.201	0.208	0.200	0.172	79.57	8.28	0
	Minangkabau	0.354	0.491	0.387	0.172	70.59	5.48	0
	Sundanese	0.278	0.278	0.273	0.235	73.30	16.00	0
SAHABATAI-LLAMA3	Acehnese	0.229	0.309	0.221	0.000	100	0.13	273
	Betawi	0.028	0.058	0.035	0.001	100	0.38	110
	Buginese	0.132	0.238	0.141	0.000	100	0.17	144
	Indonesia	0.152	0.203	0.159	0.020	100	0.27	273
	Javanese	0.024	0.053	0.029	0.000	100	0.59	194
	Minangkabau	0.081	0.230	0.108	0.000	100	0.23	133
	Sundanese	0.023	0.048	0.028	0.003	99.78	**0.00**	119

Table 12: Performance comparison of all evaluated models on the impromptu euphemism detection task across Indonesian and six regional languages. Span-based **Precision**, **Recall**, and **F1-score** measure the correctness of euphemism span generation, while **BLEU** evaluates the correctness of direct term generation corresponding to each euphemism. **FP** and **FN** rates indicate disagreement levels between model predictions and human annotations. The **#Tautology** column reports the number of cases where the generated direct term is identical to the euphemistic expression itself. Underlined values denote the best performance among languages within the same model, whereas bold values indicate the best overall performance across all models and languages.

Language	Qs. Type	Avg. Factual Precision (↑)	Avg. Factual Recall (↑)	Avg. Hallucination Rate (↓)	Avg. Omission Rate (↓)
Acehnese	NON-EUPH	0.332	**0.278****	**0.207**	0.483**
Acehnese	OVERPREDICT	**0.354**	0.201**	0.214	**0.330****
Betawi	NON-EUPH	0.336**	**0.273****	**0.213****	0.470**
Betawi	OVERPREDICT	**0.353****	0.213**	0.242**	**0.297****
Buginese	NON-EUPH	0.367	**0.237****	**0.217**	0.358**
Buginese	OVERPREDICT	**0.386**	0.163**	0.229	**0.194****
Javanese	NON-EUPH	**0.379**	**0.338****	**0.184****	0.463**
Javanese	OVERPREDICT	0.376	0.237**	0.220**	**0.311****
Minangkabau	NON-EUPH	0.320**	**0.325****	**0.192****	0.509**
Minangkabau	OVERPREDICT	**0.354****	0.266**	0.243**	**0.371****
Sundanese	NON-EUPH	0.324**	**0.391****	**0.168****	0.483**
Sundanese	OVERPREDICT	**0.342****	0.275**	0.237**	**0.341****
Indonesian	NON-EUPH	0.257**	**0.341****	**0.136****	0.580**
Indonesian	OVERPREDICT	**0.299****	0.262**	0.271**	**0.481****

Table 13: Fact-based evaluation results comparing human-labeled non-euphemistic questions (NON-EUPH) against instances falsely identified as euphemistic by the LLM (OVERPREDICT). Values marked with ** are statistically significant under Wilcoxon's signed-rank test ($p < 0.05$).

Category : Endometriosis, ID : 30

Doctor's Answer (English Translation)

Hi S******, Abdominal pain during sexual intercourse, as you are experiencing, is unlikely to be related to medication use or a fertility program you may be undergoing. In general, pain during intercourse can be caused by various medical conditions, including endometriosis, uterine fibroids, ovarian cysts, sexually transmitted infections, pelvic inflammatory disease, irritable bowel syndrome, hemorrhoids, and others. To determine the exact cause of your condition, further examination is necessary, and it is advisable to follow up with your gynecologist. Additional tests such as blood tests, laparoscopy, or a CT scan may be recommended if needed to investigate the underlying cause, allowing appropriate treatment to be provided based on your specific condition. You are also advised to engage in sexual activity slowly and without rushing, communicate with your partner about comfortable and uncomfortable positions, try different positions during intercourse, and postpone sexual activity if the pain becomes bothersome. You may also find it helpful to read about enjoying intercourse without pain. Hope this helps.

LLM's Facts (SAMPLE)

...
9. The abdominal cramping you are experiencing may be a side effect of the fertility medication you are currently taking.
...

Figure 3: Example of contradictive response

Category : Menstruation, ID : 6625

Doctor's Answer (English Translation)

Hello Y***, Dysmenorrhea refers to pain in the lower abdomen that occurs before or during menstruation. It is caused by contractions of the uterine muscles, which compress blood vessels and reduce oxygen supply to the uterus, resulting in pain. Typically, dysmenorrhea subsides within 2 to 4 days, and as age increases, the pain often diminishes on its own. However, in certain conditions, the pain can be more severe, such as in pelvic inflammation, endometriosis, adenomyosis (the growth of the uterine lining into surrounding areas), or cervical stenosis (narrowing of the cervix). The medication you mentioned contains progestin as its active ingredient. This hormonal medication is commonly used to help relieve pain in individuals with endometriosis. It is recommended that you follow your doctor's advice and take the medication as prescribed. However, if the pain persists or becomes severe despite treatment, you should return to your gynecologist for further evaluation of your condition and treatment plan. You may also apply warm compresses to help alleviate the pain. Additionally, adopting a healthy lifestyle–such as regular exercise and maintaining a balanced diet–is recommended. Avoid consuming alcohol and refrain from smoking. Hope this information is helpful.

LLM's Facts (SAMPLE)

...
4. Endometrial tissue responds to the body's hormones in a manner similar to the uterine lining.
5. During menstruation, endometrial tissue also undergoes bleeding and inflammation.
6. Endometriosis that adheres to or irritates nerves can lead to severe inflammation and intense pain.
...

Figure 4: Example of incorrect diagnosis

Category : Endometriosis, ID : 65

Doctor's Answer (English Translation)

Hello, Abdominal pain is a fairly common condition. In many cases, it resolves on its own, but in certain situations, it may require medical attention. The location of the pain can help determine its underlying cause. Some possible causes of left-sided abdominal pain include ovulation, ovarian cysts, infections of the fallopian tubes or ovaries, cervical issues, endometriosis, injury, stomach problems, and disorders of the urinary or digestive system, among others. Ovulation pain is generally mild and does not usually cause severe discomfort. However, mild pain may also be associated with stomach issues or other reproductive organ conditions. If the pain has subsided and does not recur, there is usually no need for concern. However, if the pain returns and is accompanied by other symptoms such as nausea, vomiting, or others, it may indicate a gastrointestinal or other underlying condition. In such cases, it is advisable to consult an internal medicine specialist. You may also refer to articles on ovulation for further information. Hope this helps.

LLM's Facts (SAMPLE)

...
16. A ruptured ectopic pregnancy can lead to sudden severe abdominal pain, abnormal vaginal bleeding, dizziness, weakness, or fainting.
17. A urinary tract infection (UTI) is a bacterial infection of the urinary system.
18. Urinary tract stones are stones that obstruct the flow of urine in the kidneys or ureters.
...

Figure 5: Example of plausible but unsupported response error.

Category : Menstruation, ID : 5785

Doctor's Answer (English Translation)

Dear R***, Menstruation is considered normal if it has the following characteristics: it lasts between two to seven days per cycle; the cycle repeats every 21–35 days; and the total volume of menstrual blood per cycle ranges from 10–80 mL. If blood clots are present, their volume should not exceed one quarter of the total menstrual blood. The color of menstrual blood can vary, including red, brownish, or blackish tones. Blood that is predominantly red contains more red blood cells, while brownish blood may contain more mucus. Blackish menstrual blood can result from older blood that is expelled more slowly or contains a higher proportion of dead cells. If you have concerns regarding menstruation or reproductive health, you are advised to consult a gynecologist. Hope this helps.

LLM's Facts (SAMPLE)

...
9. Brown menstrual blood indicates older blood or a slower flow, allowing it to oxidize.
...

Figure 6: Example of mismatch syntactic information

Category : Menstruation, ID : 7048

Doctor's Answer (English Translation)

Hello, thank you for your question on the Alodokter website. I will help answer your question. Menstruation normally occurs in cycles of 21–35 days and lasts for 3–7 days. Based on the dates you provided, your cycle is 24 days, which is still within the normal range. In young adult women, it is quite common for menstrual cycles to be somewhat irregular. However, as long as the cycle remains within the normal range, this is generally not a concern. If your menstrual cycle falls outside the normal range, it is advisable to consult a doctor for further evaluation to determine the underlying cause. Some factors that can cause pain during menstruation include hormonal imbalance, stress, and endometriosis (inflammation of the uterine lining). At this point, regularly checking your condition with a doctor or consulting an obstetrician-gynecologist would be the most appropriate step. Avoid frequently taking pain medications without medical advice, as they may cause irritation or injury to your stomach. I hope this answer is helpful.

LLM's Facts (SAMPLE)

...
3. You took the medication on August 27.
4. You had sexual intercourse on September 4.
5. You experienced symptoms on September 18.
...

Figure 7: Example of unsupported patient details' repetation

Category : Vagina, ID : 303

Doctor's Answer (English Translation)

Hi, this has already been answered in the previous post. Thank you.

LLM's Facts (SAMPLE)

...
2. The lump in the genital area is slightly painful at first.
3. The lump releases pus when pressed.
4. The lump decreases in size after the pus is discharged.
...

Figure 8: Example of unclear doctor's response

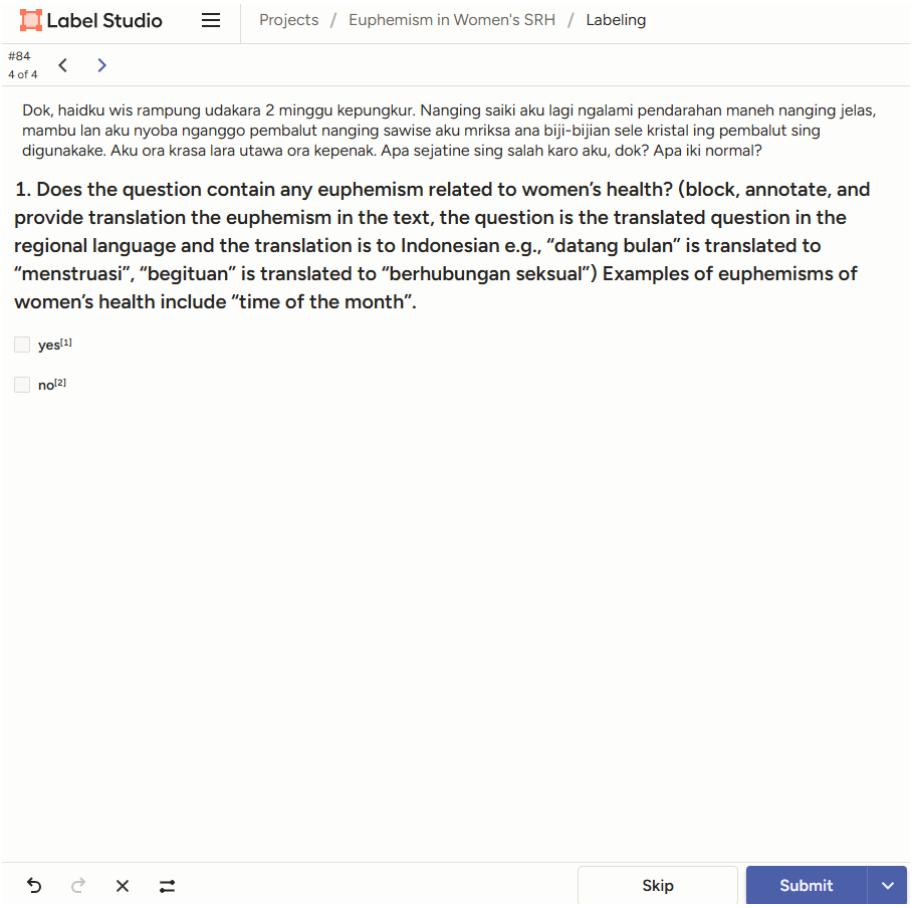

Figure 9: Euphemism Annotation Platform (Part-1).

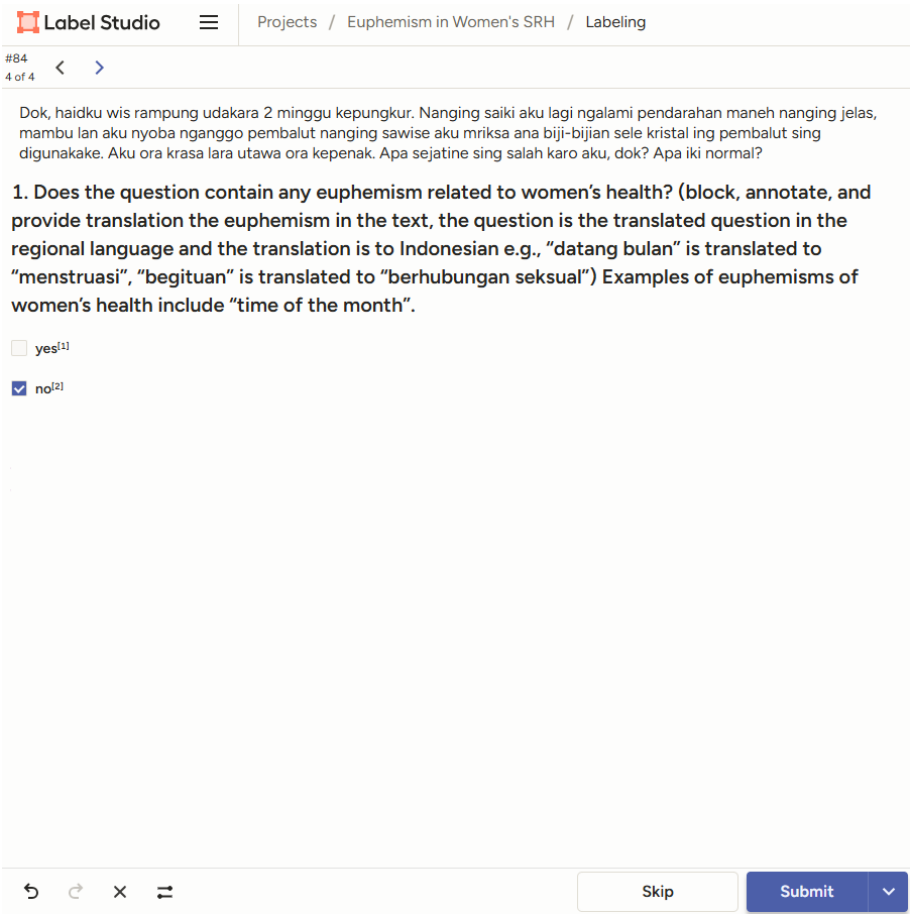

Figure 10: Euphemism Annotation Platform (Part-2).

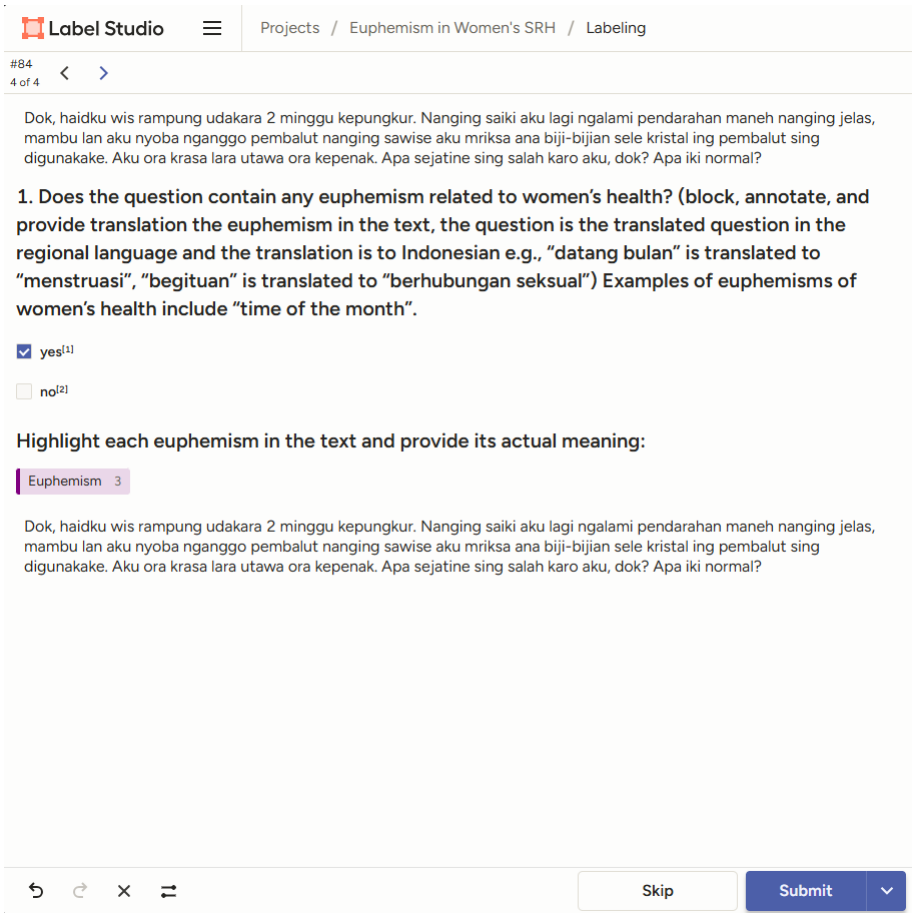

Figure 11: Euphemism Annotation Platform (Part-3).

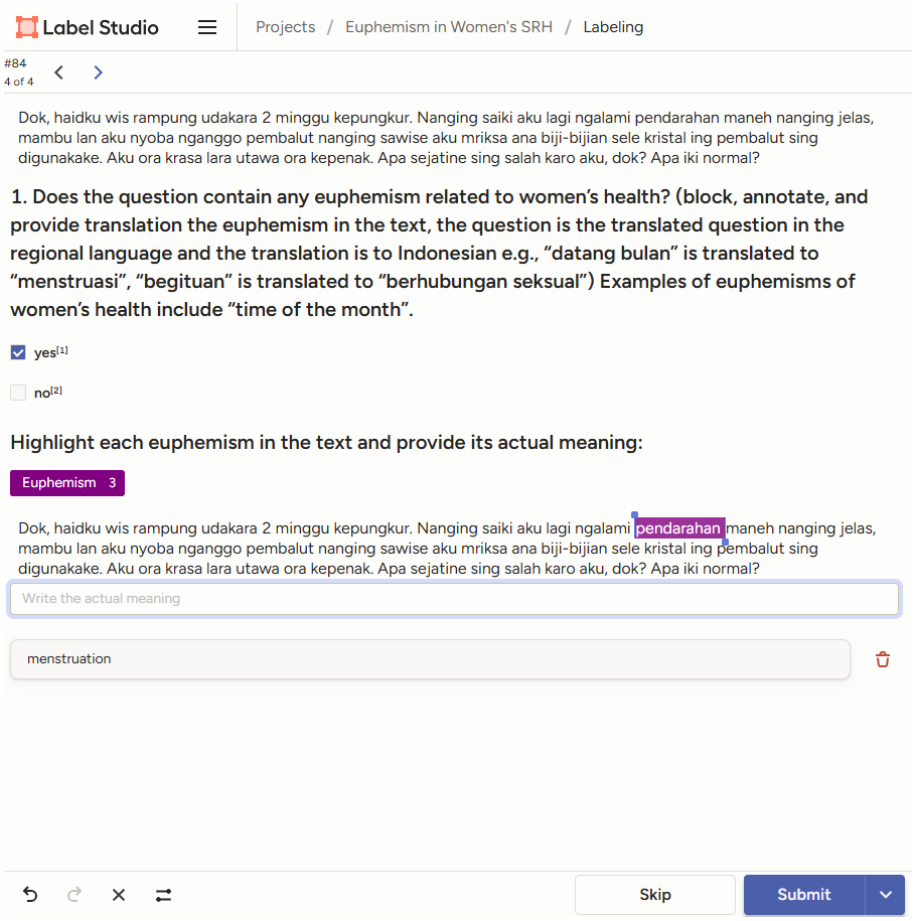

Figure 12: Euphemism Annotation Platform (Part-4).

Prompt example for Variation A : *Main*

Wangsulana pitakoné kanthi basa sing sopan lan gampang dimangertèni, kaya déné sampeyan ngomong langsung karo pasiené.

PITAKONAN:
<Patient's Question>

Figure 13: Prompt variation *A* for Task Setup (Javanese Language).

Prompt example for Variation B : *Main + Description*

Wangsulana pitakoné kanthi basa sing sopan lan gampang dimangertèni, kaya déné sampeyan ngomong langsung karo pasiené.

Ana pasien wadon saka Jawa takon marang sampeyan soal kesehatan seksualé.

PITAKONAN:
<Patient's Question>

Figure 14: Prompt variation *B* for Task Setup (Javanese Language).

Language	Description	Language Forcing	Main	Question Indicator
Javanese	Ana pasien wadon saka *Jawa* takon marang sampeyan soal kesehatan seksualé.	Sampeyan kudu njawab nganggo basa *Jawa*.	Wangsulana pitakoné kanthi basa sing sopan lan gampang dimangertèni, kaya déné sampeyan ngomong langsung karo pasiené.	PITAKONAN
Sundanese	Aya hiji pasien awéwé asal *Sunda* nu nanyakeun hal ngeunaan kaséhatan seksualna ka anjeun.	Anjeun kedah ngajawab nganggo basa *Sunda*.	Mangga jawabanana dijelaskeun ku basa nu sopan tur gampang kaharti, saolah-olah anjeun keur ngajelaskeun langsung ka pasienna.	PANYAWALAN
Buginese	Adae' makkunrai ri *Bugis* mappoji' pertanyaan rilakkoe nasaba' kesehatanna.	Iye' wajib majawab pake basa *Bugis*.	To jawabbakae' pertanyaan e' nasaba' ade'na, maddécéngngi nasaba' engka makkita langsung ri pasienna.	PERRANYE
Acehnese	Gopnyan ureuëng inong dari *Aceh* lon tanyakan meunan masalah kesehatan seksual.	Gob lé wajib menjawab nyan ngon bahasa *Aceh*.	Jawab lon pertanyaan nyan ngon bahasa sopan dan mudah pahah, seolah gob lé neujak droe-droe ka pasien.	SOALON
Minangkabau	Urang awak perempuan dari *Minang* baratanyo ka ang mengenai kesehatan seksualnyo.	Ang harus manjawo bana jo baso *Minang*.	Jawablah pertanyaan ko dengan sopan jo jelas, seolah-olah ang sedang babicaro langsung jo pasiennyo.	PERTANYOAN
Betawi	Ada cewek dari *Betawi* nanya soal kesehatan seksualnya ke lu.	Lu mesti jawabnya pake bahasa *Betawi*.	Jawab pertanyaannya dengan gaya yang sopan dan gampang dimengerti, kayak lagi ngobrol langsung sama pasiennya.	PERTANYAANNYE NIH
English	A female patient from *United States* is asking you a question regarding her sexual health.	You must answer in *English*.	Please answer the following question clearly and politely, as if you are speaking directly to the patient.	QUESTION

Table 14: Prompt components (*Description*, *Language Forcing*, and *Main*) + *Question Indicator* for six Indonesian regional languages (Javanese, Sundanese, Buginese, Acehnese, Minangkabau, and Betawi). Words highlighted in ■ indicate the region referenced in the *Description* component, while words highlighted in ■ denote the target language specified in the *Language Forcing* component.

Prompt example for Variation C : *Main + Language Forcing*

Wangsulana pitakoné kanthi basa sing sopan lan gampang dimangertèni, kaya déné sampeyan ngomong langsung karo pasiené.

Sampeyan kudu njawab nganggo basa Jawa.

PITAKONAN:
<Patient's Question>

Figure 15: Prompt variation *C* for Task Setup (Javanese Language).

Prompt example for Variation D : *Main + Description + Language Forcing*

Wangsulana pitakoné kanthi basa sing sopan lan gampang dimangertèni, kaya déné sampeyan ngomong langsung karo pasiené.

Ana pasien wadon saka Jawa takon marang sampeyan soal kesehatan seksualé.

Sampeyan kudu njawab nganggo basa Jawa.

PITAKONAN:
<Patient's Question>

Figure 16: Prompt variation *D* for Task Setup (Javanese Language).

Prompt template for Translation phase

You are a professional translator.
Translate the following text into natural, fluent Indonesian.
Preserve meaning exactly. Do not add explanations.

INPUT:
<Model's Answer>

Figure 17: Prompt template for Translation phase.

Prompt template for Common-Facts extraction phase

Consider the INPUT below.
INPUT:
<Doctor's Answer>

1. Is the fact <Fact 1> stated or supported by the INPUT?
2. Is the fact <Fact 2> stated or supported by the INPUT?
3. Is the fact <Fact 3> stated or supported by the INPUT?
...

Write your answers by iterating only "Yes" or "No" for each question, using the format below, without adding any other output:
1. Yes
2. No
3. Yes
...

OUTPUT:

Figure 18: Prompt template for the Common-Facts (Correct and Incorrect Facts) extraction phase.

Prompt template for Hallucinated-Facts extraction phase

Consider the INPUT below.
INPUT:
<Doctor's Answer>

1. Does the claim <Fact 1> deviate from the INPUT or contain harmful information or contradict the INPUT?
2. Does the claim <Fact 2> deviate from the INPUT or contain harmful information or contradict the INPUT?
3. Does the claim <Fact 3> deviate from the INPUT or contain harmful information or contradict the INPUT?
...

Write your answers by iterating only "Yes" or "No" for each question, using the format below, without adding any other output:
1. Yes
2. No
3. Yes
...

OUTPUT:

Figure 19: Prompt template for the Hallucinated-Facts extraction phase.

Prompt template for Critical-Omission extraction phase

Consider the following INPUT FACTS and OUTPUT.
INPUT FACTS (Medical Facts from Doctor):
<1. Doctor's Fact 1>
<2. Doctor's Fact 2>
<3. Doctor's Fact 3>
...
OUTPUT (LLM Answer):
<LLM's Answer>

For each numbered medical fact above, answer whether the fact is CRITICAL INFORMATION but NOT supported by the OUTPUT.
Write your answers by iterating only "Yes" or "No" for each question, using the format below, without adding any other output:
1. Yes
2. No
3. Yes
...

OUTPUT:

Figure 20: Prompt template for the Critical-Omission extraction phase.

Prompt template for Impromptu Euphemism Detection

You are a linguistic expert specializing in medical communication.
Your task is to identify euphemistic expressions in a medical question and provide their literal equivalents.

A euphemism is an indirect or mild word/phrase used instead of a more direct clinical term. Examples:
- "kandungan" (womb, used colloquially for pregnancy) -> "kehamilan" (pregnancy)
- "terlambat bulan" (late month) -> "tidak menstruasi" (not menstruating)
- "berhubungan" (to associate/relate) -> "berhubungan seks" (to have sex)

Rules:
- If NO euphemistic expressions are present, respond with NOTHING (completely empty response).
- If euphemistic expressions are found, list ONLY them in this exact format:
1. <euphemism term> : <literal medical equivalent>
2. <euphemism term> : <literal medical equivalent>
- Use the exact euphemism term as it appears in the question (preserve original casing/spelling).
- The literal medical equivalent (direct term) MUST be written in the same language as the original euphemism.
- Do NOT translate the euphemism or the literal equivalent into another language.
- Do NOT add explanations, headers, footnotes, or any extra text.

INPUT:
<Patient's Question>

Figure 21: Prompt template for the Impromptu Euphemism Detection phase.