Do Text Anonymizers Generalize Across Contexts?
Extending RAT-Bench to Malaysian Microdata and PII

David Hong Liang Chew [1] **Zexi Yao** [2] **Nataša Krčo** [2] **Matthieu Meeus** [2] **Waqas Khalid Obeidy** [1]
Yves-Alexandre de Montjoye [2]

Abstract

Text anonymizers are often used before sensitive text enters foundation-model workflows, including training, fine-tuning, retrieval, and user-facing inference. However, these tools are commonly evaluated on English text with U.S.-centric identifiers and cultural context. We extend RAT-Bench, an existing primarily English-language re-identification benchmark, to the Malaysian context using local microdata, Malaysian personally identifiable information (PII), and culturally grounded transcripts in Malaysian English and Bahasa Malaysia. We evaluate NER- and LLM-based anonymizers with an LLM attacker that infers attributes from anonymized text, measuring both re-identification success and text utility. Across Malaysian English and Bahasa Malaysia, LLM anonymizers provide the strongest explicit privacy–utility trade-offs, reducing Easy/Hard re-identification risk to 23–29% while preserving BLEU scores of 0.77–0.94. These results show that anonymization benchmarks are not context-neutral: local identifiers, language use, and cultural context should be considered before deploying anonymization tools in target foundation-model pipelines.

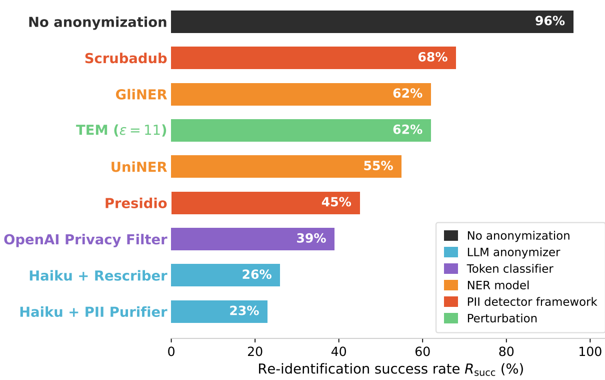


Figure 1. Malaysian English re-identification risk averaged over Easy and Hard settings.

1. Introduction

Sensitive text is increasingly used in foundation-model workflows, including model training, fine-tuning, retrieval, and user-facing inference. A common mitigation is to anonymize text before model use, reducing memorization and leakage risk. However, removing direct identifiers such as names is not enough: a person may still be re-identified from combinations of indirect identifiers such as date of birth, occupation, ethnicity, and location (Sweeney, 2002; Rocher et al., 2019). For example, an anonymized transcript that removes a name but retains a rare occupation, age, postcode, and ethnicity may still point to only a small number of people in the relevant population. This makes anonymization evaluation population-relative: residual risk depends not only on what text was removed, but also on how unique the remaining attributes are in the target demographic context (Rocher et al., 2019).

RAT-Bench evaluates anonymization through residual re-identification risk rather than span-level recall (Krčo et al., 2026). However, its demographic grounding, identifiers, and cultural context are primarily U.S.-centric and do not account for other geographies. It is therefore unclear whether anonymizers that perform well under RAT-Bench remain effective when the population distribution, identifier formats, language, and cultural cues change.

We examine whether anonymizers retain their performance when transferred to the Malaysian setting. We make three contributions: (i) a framework for extending RAT-Bench to the Malaysian context using differentially private synthetic microdata derived from Malaysia IPUMS records; (ii) a Malaysian PII-generation pipeline covering names, addresses, postcodes, phone formats, National Registration

[1]Payments Network Malaysia Sdn Bhd (PayNet) [2]Imperial College London. Correspondence to: David Hong Liang Chew <davidchewhl@gmail.com>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Identity Card (NRIC)/passport numbers, Employee Provident Fund (EPF) numbers, Tax Identification Numbers (TINs), credit cards, and bank accounts; (iii) an evaluation of out-of-the-box anonymizers on Malaysian English and Bahasa Malaysia transcripts. We find that LLM-based anonymizers generally provide the strongest privacy–utility trade-off under unchanged prompts, while several non-LLM tools miss local PII, degrade on non-standard forms, or severely over-redact Bahasa Malaysia text.

2. Background and Related Work

Memorization, inference, and anonymization. Foundation models create privacy risks both by memorizing rare training strings and by inferring sensitive attributes from contextual cues. Prior work shows that language models can reproduce or reveal training examples under suitable attacks, motivating privacy controls before sensitive text is used for training, fine-tuning, retrieval, or in-context learning (Carlini et al., 2019; 2021; 2023; Shi et al., 2024). Privacy leakage is not limited to memorization: LLMs can infer personal attributes from ordinary user-written text (Staab et al., 2024), and LLM agents can deanonymize pseudonymous users from unstructured online text at scale (Lermen et al., 2026). Text anonymization is therefore a practical mitigation, but its effectiveness should be evaluated against adversarial re-identification rather than span removal alone.

Risk-based evaluation and localization. Most anonymization benchmarks evaluate protected-span detection or recall (Lison et al., 2021; Pilán et al., 2022; Kim et al., 2024). Recent work evaluates anonymization under stronger LLM-based adversaries and shows that LLM-based anonymization can improve privacy–utility trade-offs (Staab et al., 2025; Yang et al., 2025). However, span removal alone does not measure whether the remaining text still contains enough direct or indirect information to identify an individual. This connects text anonymization to classical disclosure-control work, where risk is population-relative and depends on the uniqueness of indirect identifiers (Sweeney, 2002; Narayanan & Shmatikov, 2008; de Montjoye et al., 2013; Rocher et al., 2019; Gadotti et al., 2024). As such, risk is population-relative, conclusions from a U.S.-grounded benchmark such as RAT-Bench need not transfer unchanged to Malaysia (Krčo et al., 2026). We therefore test whether RAT-Bench's findings remain stable after localizing the reference microdata, direct identifiers, and language context.

3. Malaysian Benchmark Construction

3.1. Source Microdata and Record Sampling

We use year-2000 Malaysian IPUMS (Ruggles et al., 2025) microdata, the latest Malaysian microdata available to us. The original data contains more than 400,000 records and cannot be publicly released. We select and transform relevant IPUMS variables to align with the identifier categories defined in RAT-Bench, then use PrivBayes via Reprosyn with privacy budget $\epsilon = 5$ to generate a synthetic microdata sample of 400,000 records (Zhang et al., 2017; Alan Turing Institute, 2024). This synthetic microdata is used for benchmark construction and re-identification scoring.

Following RAT-Bench, we select benchmark records that are already high-risk before anonymization. For a candidate record, baseline correctness measures how likely an attacker is to recover the correct individual from the reference population using a sampled set of indirect identifiers. We sample 100 high-risk benchmark records by selecting records whose baseline correctness under five sampled indirect identifiers exceeds $\theta_0 = 0.9$.

3.2. Indirect Identifier Construction and Enrichment

We map Malaysian IPUMS variables to RAT-Bench's nine indirect-identifier categories: state, sex, age, race, marital status, education, employment status, occupation, and citizenship. Two mappings require approximation: place of birth is used as the closest available proxy for state of residence, and education level is combined with highest certificate obtained to approximate RAT-Bench's finer-grained education attribute. Full mapping details are in Appendix A.1.

Postcode is needed for generating realistic Malaysian addresses and scoring partial address recovery, but the Malaysia IPUMS extract does not provide postcode-level geography. We therefore enrich the synthetic microdata with state-consistent postcodes using a population-weighted procedure, rather than sampling postcodes uniformly within each state. Unlike RAT-Bench's U.S. setting, where ZIP codes can be mapped to PUMA geography, Malaysia has no PUMA-equivalent field, so we use postcodes directly in re-identification scoring.

For each record, we sample a postcode z within state s by first weighting districts by population and then sampling uniformly from the postcode pool matched to the selected district:

$$P(z \mid s) = \frac{\text{population}(d_z)}{\sum_{\tilde{d} \in \mathcal{D}(s)} \text{population}(\tilde{d})} \cdot \frac{1}{|\mathcal{Z}_{d_z}|},$$

where d_z is the district matched to postcode z, $\mathcal{D}(s)$ is the set of districts in state s, and $\mathcal{Z}_{d_z}$ is the set of postcodes matched to district d_z (Department of Statistics Malaysia, 2024; heiswayi, 2025). Additional postcode-matching details are provided in Appendix A.4.

3.3. Malaysian Direct Identifier Generation

We also introduce three Malaysian-relevant identifiers: bank account number, EPF number, and TIN. These are strongly identifying locally: bank account numbers and mobile numbers can reveal or confirm account-holder names in common payment workflows, while EPF and TIN are unique administrative identifiers linked to an individual's NRIC.

Identifiers are generated using local format constraints. Names are conditioned on state, ethnicity, and gender; addresses are synthetic but follow Malaysian address structure and valid postcodes; phone numbers include mobile and landline formats; and NRICs encode date of birth, state code, and gender. To ensure coverage and comparability, we stratify direct identifier assignment: in each 100-record explicit setting, each of the nine direct identifier categories appears 11 times, with one category appearing 12 times. Full generation rules and format constraints are provided in Appendix B.

3.4. Transcript Generation and Quality Control

Following RAT-Bench, we generate transcripts in three scenarios: medical consultation, chatbot conversation, and meeting transcript. We repeat generation in Malaysian English and Bahasa Malaysia, grounding both in Malaysian names, geography, institutions, and social context. For each language and difficulty setting, we generate 100 transcripts from the same selected high-risk records.

We use RAT-Bench's three difficulty settings. Easy states one direct identifier and five indirect identifiers explicitly; Hard keeps them explicit but uses non-standard forms; and Implicit contains only indirect identifiers, conveyed through contextual clues. Implicit is less straightforward in the Malaysian context because target attributes must be recoverable from indirect local cues. Initial generations did not consistently contain enough evidence for the attacker to infer the target attributes. We therefore manually corrected 61/100 Malaysian English Implicit transcripts to strengthen clue bundles while keeping target values implicit, increasing no-anonymization re-identification from 5% to 35%, close to RAT-Bench's 32% Implicit risk. Bahasa Malaysia Implicit showed the same issue, so we treat it as exploratory and exclude it from the main comparison.

4. Experimental Setup

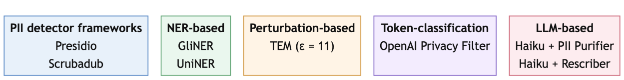


Figure 2. Anonymizer families evaluated in this work.

Anonymizers. We evaluate seven anonymizers from RAT-Bench: five non-LLM methods (Presidio, Scrubadub, GliNER, UniNER, TEM with $\epsilon = 11$) and two LLM-based methods using Claude Haiku 4.5 with the PII Purifier and Rescriber prompts. We add OpenAI Privacy Filter (OpenAI, 2026). All anonymizers are used out of the box.

Generator and attacker. We use Claude Sonnet 4.5 for transcript generation and attack. This provides some model variation across the pipeline while retaining a strong generator and attacker. The attacker prompt follows RAT-Bench, with one localized addition instructing the attacker to make inferences from a Malaysian context and to use 1 January 2000 as the reference date for time-based reasoning. The exact modified prompt footer is provided in Appendix F.3.

Metrics. Following RAT-Bench, a record is re-identified if at least one direct identifier is recovered, or if recovered indirect identifiers exceed the estimated risk threshold $\theta = 0.2$. This corresponds to the $k = 5$ intuition from k-anonymity, where a record should remain indistinguishable from at least four others under quasi-identifier matching (Sweeney, 2002). We report R_{succ}, the re-identification success rate, with lower values indicating lower residual privacy risk, and BLEU, with higher values indicating better text preservation and utility.

5. Results

Table 1 shows three main findings. First, LLM-based anonymizers provide the strongest Easy/Hard privacy–utility trade-off in Malaysian English (privacy–utility plots are shown in Appendix E.1). Haiku + PII Purifier achieves the lowest average explicit re-identification rate (23%) while preserving high utility (BLEU 0.94). Haiku + Rescriber is close in privacy (26%) but has lower utility (BLEU 0.83). OpenAI Privacy Filter preserves utility similarly well, but its risk increases sharply from Easy to Hard identifiers (18% to 59%), suggesting weaker protection for non-standard local forms. Among non-LLM methods, Presidio is strongest but still leaves 45% average explicit risk.

Second, the Malaysian English Implicit setting evaluates contextual re-identification: whether indirect cues can still identify someone when direct identifiers are absent. Since no direct identifiers are present, record-level R_{succ} depends entirely on recovered indirect attributes. Haiku + Rescriber performs best in this setting, reducing R_{succ} to 1% while preserving BLEU 0.87. Presidio is the strongest non-LLM method on Implicit risk (2%), but with lower utility (BLEU 0.85) than Haiku + Rescriber and Haiku + PII Purifier. Haiku + PII Purifier has higher residual risk (7%) but preserves the most utility among these methods (BLEU 0.97).

Third, Bahasa Malaysia shows a larger gap between LLM-based and non-LLM privacy–utility trade-offs. The LLM methods remain comparatively stable: Haiku + Rescriber achieves the lowest average explicit risk (23%), while Haiku + PII Purifier preserves higher utility with moderately higher risk (29%, BLEU 0.92). In contrast, several non-LLM methods show sharper failure modes. Presidio reduces average explicit risk to 40%, but BLEU drops to 0.17, indicating

Table 1. Main results for Malaysian English and Bahasa Malaysia. R_{succ} is re-identification success rate. Avg. is computed over Easy and Hard. Bahasa Malaysia Implicit is omitted because the no-anonymization result remains too low at 7% to support meaningful comparison.

Method	Malaysian English						Bahasa Malaysia			
	R_{succ} (% ↓)				BLEU (↑)		R_{succ} (% ↓)			BLEU (↑)
	Easy	Hard	Avg.	Implicit	Avg.	Implicit	Easy	Hard	Avg.	Avg.
No Anonymization	100	91	96	35	1.00	1.00	100	96	98	1.00
Presidio	40	50	45	2	0.82	0.85	52	**28**	40	0.17
Scrubadub	64	72	68	11	0.77	0.82	68	61	65	0.41
GliNER	66	57	62	2	0.87	0.66	66	63	65	0.49
UniNER	54	56	55	25	0.65	0.96	51	73	62	0.87
TEM $\epsilon = 11$	55	69	62	30	0.65	0.70	45	76	61	0.25
Haiku 4.5 + PII Purifier Prompt	21	**24**	**23**	7	**0.94**	0.97	20	37	29	0.92
Haiku 4.5 + Rescriber Prompt	21	31	26	**1**	0.83	0.87	**8**	37	**23**	0.77
OpenAI Privacy Filter	**18**	59	39	18	**0.94**	**0.98**	42	83	63	**0.94**

severe over-redaction. Conversely, OpenAI Privacy Filter preserves utility well (BLEU 0.94) but leaves high residual risk (63%). We provide an example of such over-redaction in Appendix E.2.

Overall, non-LLM tools exhibit mixed failure modes: some preserve utility but leave high residual risk, while others heavily redact benign text yet can still leave local PII unredacted. Relative to RAT-Bench (Krčo et al., 2026), our results preserve the broad LLM advantage but show a clearer gap between LLM- and NER-based method performance in Malaysian English and Bahasa Malaysia across both Easy and Hard examples.

6. Discussion and Limitations

Our results support RAT-Bench's conclusion that LLM-based anonymizers often provide a stronger privacy–utility trade-off than NER-based systems, but show that U.S.-centric benchmarks can understate deployment failures elsewhere. In Malaysia, local identifier formats, language, and cultural context make failures more visible: some NER models miss locally meaningful identifiers or non-standard explicit forms, while in Bahasa Malaysia they can over-redact large portions of text and still miss direct identifiers. This suggests that anonymizers should be trained and evaluated on deployment-relevant languages and cultural contexts, not only on U.S.-centric data.

For practitioners, anonymizer choice should be guided by the language and context of deployment rather than English or U.S.-centric benchmarks alone. LLM-based anonymizers appear more adaptable under unchanged anonymization prompts, but are costlier and should still be evaluated in the target context.

Limitations. Due to data-use restrictions, we cannot release the original or synthetic Malaysia IPUMS microdata or generated benchmark. State is approximated using place of birth rather than current residence. Postcode enrichment uses 2020 district population statistics, the earliest available source we found, although the base microdata is from 2000. Implicit generation remains challenging: Malaysian English Implicit examples required quality control, while Bahasa Malaysia Implicit remains preliminary. Malaysia-specific recognizers or fine-tuned systems could improve non-LLM tools; our goal is to test default cross-context generalization.

7. Conclusion

We extend RAT-Bench to Malaysia by constructing differentially private synthetic microdata derived from Malaysian IPUMS records, adding Malaysian-specific direct identifiers, and evaluating anonymizers on Malaysian English and Bahasa Malaysia transcripts. The broad RAT-Bench pattern transfers: LLM-based anonymizers generally provide the strongest privacy–utility trade-off. However, the Malaysian setting makes several failure modes more visible, including missed local identifiers, weaker protection for non-standard explicit forms, and severe over-redaction in Bahasa Malaysia text. These findings suggest that anonymization benchmarks should be localized across identifier schemas, language use, and cultural context before they are used to guide trustworthy foundation-model deployment.

Impact Statement

This paper studies text anonymization tools intended to reduce privacy risks in foundation-model pipelines. Its positive impact is improved evaluation of anonymizers in localized settings where identifier systems, language use, and cultural context are underrepresented in existing benchmarks. A potential negative impact is that such evaluations may reveal weaknesses in deployed anonymization systems. Overall, we argue that localized privacy evaluation is necessary for trustworthy deployment across societies and languages.

References

Alan Turing Institute. Reprosyn: Synthesising tabular data. https://github.com/alan-turing-institute/reprosyn, 2024. Python library for generating synthetic data. Accessed 2026-03-30.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, Santa Clara, CA, August 2019. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1376), 2013. doi: 10.1038/srep01376. URL https://doi.org/10.1038/srep01376.

Department of Statistics Malaysia. Population table: Administrative districts. https://open.dosm.gov.my/data-catalogue/population_district, 2024. OpenDOSM data catalogue. Accessed 2026-04-29.

Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.-M., and de Montjoye, Y.-A. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053, 2024. doi: 10.1126/sciadv.adn7053. URL https://www.science.org/doi/abs/10.1126/sciadv.adn7053.

heiswayi. Malaysia postcodes. https://github.com/heiswayi/malaysia-postcodes, 2025. GitHub repository. Accessed 2026-04-29.

Kim, W., Hahm, S., and Lee, J. Generalizing clinical de-identification models by privacy-safe data augmentation using GPT-4. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21204–21218, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1181. URL https://aclanthology.org/2024.emnlp-main.1181/.

Krčo, N., Yao, Z., Meeus, M., and de Montjoye, Y.-A. Ratbench: A comprehensive benchmark for text anonymization, 2026. URL https://arxiv.org/abs/2602.12806.

Lermen, S., Paleka, D., Swanson, J., Aerni, M., Carlini, N., and Tramèr, F. Large-scale online deanonymization with llms, 2026. URL https://arxiv.org/abs/2602.16800.

Lison, P., Pilán, I., Sanchez, D., Batet, M., and Øvrelid, L. Anonymisation models for text data: State of the art, challenges and future directions. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.323. URL https://aclanthology.org/2021.acl-long.323/.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008. doi: 10.1109/SP.2008.33.

OCBC Bank Malaysia. IBG Participating Banks Account Structure. https://www.ocbc.com.my/assets/pdf/Giro/IBG_Participating_Banks_Account_Structure.pdf, n.d. Accessed 2026-05-01.

OpenAI. OpenAI Privacy Filter Model Card. https://cdn.openai.com/pdf/c66281ed-b638-456a-8ce1-97e9f5264a90/OpenAI-Privacy-Filter-Model-Card.pdf, April 2026. Accessed 2026-05-08.

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., and Batet, M. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48 (4):1053–1101, December 2022. doi: 10.1162/coli_a_00458. URL https://aclanthology.org/2022.cl-4.19/.

Rocher, L., Hendrickx, J. M., and de Montjoye, Y.-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(3069), 2019. doi: 10.1038/s41467-019-10933-3. URL https://doi.org/10.1038/s41467-019-10933-3.

Ruggles, S., Cleveland, L., Lovaton, R., Sarkar, S., Sobek, M., Burk, D., Ehrlich, D., Lee, J., and Merrill, N. Integrated public use microdata series, international: Version 7.6 [dataset]. Minneapolis, MN: IPUMS, 2025.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Staab, R., Vero, M., Balunovic, M., and Vechev, M. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

Staab, R., Vero, M., Balunovic, M., and Vechev, M. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=82p8VHRsaK.

Sweeney, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5): 557–570, October 2002. ISSN 0218-4885. doi: 10.1142/S0218488502001648. URL https://doi.org/10.1142/S0218488502001648.

Yang, T., Zhu, X., and Gurevych, I. Robust utility-preserving text anonymization based on large language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28922–28941, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1404. URL https://aclanthology.org/2025.acl-long.1404/.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017. ISSN 0362-5915. doi: 10.1145/3134428. URL https://doi.org/10.1145/3134428.

A. Benchmark Construction Details

A.1. Mapping Malaysian IPUMS Variables

Table 2 summarizes how we map the indirect identifiers used in RAT-Bench to the Malaysia IPUMS 2000 variables, originally produced by Department of Statistics Malaysia. RAT-Bench is based on U.S. PUMS-style demographic attributes, so several fields do not have exact equivalents in the Malaysian IPUMS extract. We therefore clean, merge, or approximate IPUMS variables to preserve the same broad indirect-identifier categories while adapting them to the Malaysian census schema.

Table 2. Mapping from RAT-Bench/PUMS indirect identifiers to Malaysian IPUMS variables. Date of birth and postcode are added after synthetic microdata generation for transcript generation and re-identification scoring.

RAT-Bench field	Meaning in RAT-Bench/PUMS	Malaysia IPUMS field(s)	Adaptation
ST	U.S. state of residence	BPLS, BPLC	Malaysia IPUMS does not provide current state of residence in our extract. We use Malaysian state of birth as the closest available geographic proxy. For foreign-born records, BPLC provides country of birth instead of a Malaysian state.
SEX	Sex	SEX	Used directly.
AGE	Age	AGE	Used directly. Following RAT-Bench, we generate date of birth by sampling a valid month and day consistent with calendar constraints, then assigning a birth year from age. Date of birth is generated from age using 1 January 2000 as the reference date. For age values coded as 99+, we randomly assign an age between 99 and 110 before generating date of birth.
RAC2P	Race / ethnicity	ETHNIC	Mapped to Malaysian ethnicity categories, including Malay, Chinese, Indian, other Bumiputera, Others, and non-citizens.
MAR	Marital status	MARST	Used directly.
SCHL	Educational attainment	EDLEV, DEGREE	Merged into a single education variable to better approximate the granularity of PUMS `SCHL`. We use `DEGREE` when a valid certificate value exists; otherwise we fall back to `EDLEV`.
ESR	Employment status	WORKING	Used directly.
OCCP	Occupation / job title	OCC3	Used directly.
CIT	Citizenship status	CITIZEN1, CITIZEN2	CITIZEN1 is used as the main citizenship variable. CITIZEN2 contains the country of origin and is a supporting field for non-citizens, used when generating passport numbers.

For geography, RAT-Bench uses U.S. state of residence, whereas the Malaysia IPUMS extract available to us provides place of birth. We therefore use Malaysian state of birth as the closest available geographic proxy. Foreign-born individuals do not have a Malaysian birth state, so for these records we use the IPUMS `CITIZEN2` attribute to obtain country-of-origin information and assign passport-style identifiers rather than NRIC numbers.

Education also requires merging. U.S. PUMS provides a relatively fine-grained educational-attainment variable, `SCHL`, which distinguishes levels such as kindergarten, individual school grades, high-school completion, some college, bachelor's degree, master's degree, and doctorate. In contrast, Malaysia IPUMS separates education into two less directly comparable fields: `EDLEV`, which gives a broad highest education level such as no schooling, primary, secondary, post-secondary, or tertiary education; and `DEGREE`, which gives the highest certificate obtained, such as lower-secondary, upper-secondary, post-secondary, diploma, degree, or postgraduate qualification.

We construct a single education variable by using `DEGREE` when a valid certificate value is available, and falling back to `EDLEV` otherwise. This preserves certificate-level detail for individuals with formal qualifications while still retaining broad education information for individuals without a recorded certificate.

A.2. Synthetic Microdata Generation

We use Malaysia IPUMS year-2000 microdata, the latest Malaysian IPUMS microdata available to us. Because the original IPUMS records cannot be publicly released under the data-use license, it is only used to construct the privacy-preserving synthetic microdata for this work. After applying the variable cleaning and merging described in Section A.1, we generate synthetic microdata using PrivBayes with privacy budget $\epsilon = 5$ via the ReProSyn library (Zhang et al., 2017; Alan Turing Institute, 2024). We use $\epsilon = 5$ as a practical privacy–utility trade-off: the synthetic data must preserve enough demographic dependence structure for re-identification-risk estimation, while avoiding direct release of the underlying restricted microdata.

The resulting synthetic microdata contains 400,000 records. Following RAT-Bench, this synthetic population has two roles. First, it serves as the reference population for estimating re-identification risk. Second, it is the pool from which we select benchmark records. For each candidate benchmark case, we sample five indirect identifiers from the available identifier set and estimate the record's baseline correctness under those identifiers. We then select 100 high-risk records whose baseline correctness exceeds $\theta_0 = 0.9$. This follows the RAT-Bench design choice of evaluating anonymizers on records that are meaningfully re-identifiable before anonymization, rather than on records that would be low-risk even without protection.

A.3. Date-of-Birth Enrichment

As with PUMS, date of birth is not directly available in the Malaysia IPUMS extract. We therefore generate date of birth from age.

We assume a reference date of 1 January 2000, matching the census year. For each record, we sample a random birth month and day, taking into account month lengths and leap years, then assign a birth year consistent with the recorded age at the reference date. For rows with that age value of 99+, we sample an age between 99 and 110. The generated date of birth is used both during transcript generation and during re-identification scoring. If the attacker does not recover the full date of birth, correct components such as day, month, or year can still contribute to indirect re-identification risk.

A.4. Postcode Enrichment

The postcode-enrichment procedure is summarized in the main paper. Here, we provide additional implementation details. Malaysia IPUMS does not provide postcode-level geography, so we assign each synthetic record a state-consistent postcode after synthetic microdata generation. The assigned postcode is used both to generate realistic Malaysian addresses and to support re-identification scoring when an attacker recovers a postcode or partial address.

We use district population data from the Malaysia Department of Statistics and postcode data from a public Malaysian postcode repository (Department of Statistics Malaysia, 2024; heiswayi, 2025). The population data is from 2020, which is the closest available year to the 2000 IPUMS microdata in the source we use. Although this does not reconstruct year-2000 postcode frequencies exactly, it provides a more grounded assignment than uniformly sampling from all postcodes in a state because postcode selection approximately reflects district-level population distribution.

For each record, we first sample a district d within state s proportional to district population:

$$P(d \mid s) = \frac{\text{population}(d)}{\sum_{\tilde{d} \in \mathcal{D}(s)} \text{population}(\tilde{d})},$$

where $\mathcal{D}(s)$ is the set of districts in state s. We then sample a postcode uniformly from the postcode pool matched to that district:

$$P(z \mid d) = \frac{1}{|\mathcal{Z}_d|},$$

where z is the sampled postcode and $\mathcal{Z}_d$ is the set of postcodes matched to district d.

Equivalently, for a postcode z matched to district d_z, the state-conditional sampling probability is

$$P(z \mid s) = \frac{\text{population}(d_z)}{\sum_{\tilde{d} \in \mathcal{D}(s)} \text{population}(\tilde{d})} \cdot \frac{1}{|\mathcal{Z}_{d_z}|}.$$

The two data sources use different geographic granularities: the population data is reported at the district level, while the postcode database is organized by city or area. We therefore map postcode entries to districts using name matching followed by manual review of mismatches. If a district cannot be matched to any postcode pool, we exclude it from postcode sampling and renormalize the district-sampling probabilities over the matched districts in the same state. After sampling a district, we sample uniformly from the unique postcode pool matched to that district.

Unlike RAT-Bench's U.S. setting, where ZIP code can be mapped to PUMA geography, our Malaysian setting has no PUMA-equivalent field. We therefore use postcode directly in the re-identification model.

B. Direct Identifier Generation

After selecting the 100 high-risk benchmark records, we generate Malaysian direct identifiers for each record before transcript generation. RAT-Bench includes U.S.-centric direct identifiers such as Social Security numbers and U.S. addresses. Our Malaysian extension instead uses identifier formats and contextual constraints relevant to Malaysia. We generate nine direct identifier categories: name, email, address, credit card number, phone number, NRIC/passport number, bank account number, EPF number, and TIN.

Coverage balancing. Rather than sampling direct identifier types fully at random, we stratify assignment to ensure that all identifier categories are evaluated. In each 100-record explicit setting, eight categories appear 11 times and one category appears 12 times; in our experiments, email appears 12 times. This avoids a split in which locally important identifiers such as NRIC, EPF, TIN, or bank account numbers appear too rarely to evaluate. It also makes the direct-identifier distribution more stable across Easy and Hard settings.

Table 3. Direct identifier categories generated for the Malaysian benchmark. All identifiers are synthetic and used only for benchmark construction.

Identifier type	Generation approach
Name	Generated using Claude Sonnet 4.5 with state, ethnicity, gender, citizenship status, and, for non-citizens, country-of-origin context. Prompts cover Malay, Chinese Malaysian, Malaysian Indian, other Bumiputera (non-Malay indigenous groups), mixed-heritage, and foreign-name patterns.
Email	Generated from synthetic names using common email-provider domains such as Gmail, Outlook, Hotmail, and Yahoo.
Address	Generated with an LLM from Malaysian address-style examples, constrained to the assigned state and postcode. Real address examples are used only as style seeds; final addresses are synthetic.
Credit card	Generated using card schemes relevant to Malaysia: Visa, Mastercard, American Express, and UnionPay. Generated numbers pass the Luhn check.
Phone number	Generated as mobile or landline numbers. Mobile numbers use Malaysian 01x prefixes; landlines use state-conditioned area codes.
NRIC/passport	Malaysian citizens born in Malaysia receive synthetic NRIC numbers consistent with date of birth, state code, and gender. Non-citizens and permanent residents receive passport-style identifiers conditioned on country of origin where possible.
Bank account	Generated using account-number lengths from banks operating in Malaysia.
EPF	Generated as a unique 8-digit synthetic Employee Provident Fund number.
TIN	Generated as an individual Malaysian TIN using the IG prefix followed by 9–12 digits.

B.1. Name Generation

Names are generated with Claude Sonnet 4.5 using prompts conditioned on the synthetic record's state, ethnicity, gender, citizenship status, and, for non-citizens, country of origin. The goal is to create plausible Malaysian names with realistic variation. We generate names separately for major Malaysian naming contexts: Malay, Chinese Malaysian, Malaysian Indian, other Bumiputera, mixed-heritage, and foreign-origin profiles.

The prompts encode culturally specific naming conventions. Malay names may include patronymics such as "bin" and "binti", which mean "son of" and "daughter of", and occasional honorific or hereditary titles. Chinese Malaysian names include both English-first and romanized Chinese forms, with variation in surname and given-name structure. Malaysian Indian names cover several locally common patterns, including names with "A/L" and "A/P" patronymic markers, where "A/L" abbreviates *anak lelaki* (son) and "A/P" abbreviates *anak perempuan* (daughter); English-given-name variants combined with Indian name elements; and Sikh/Punjabi-style names using Singh or Kaur where appropriate. Other Bumiputera names are generated with regional guidance for Sabah, Sarawak, and Peninsular Orang Asli contexts. Mixed-heritage names are generated using multiple plausible Malaysian blend patterns, such as Malay–Chinese, Chinese–Indian, Malay–Indian, and Eurasian-influenced forms, rather than repeating a single template. Generated names are checked for uniqueness within the generated pool.

B.2. Email Addresses

Email addresses are generated following RAT-Bench's procedure, using the synthetic person's name and generated date of birth. We construct the local-part from common email-address patterns, including first-name and last-name combinations, initials, separators such as periods or underscores, and optional numeric components derived from the birth day, birth month, or birth year. We then assign common email-provider domains such as Gmail, Outlook, Hotmail, and Yahoo. This produces realistic-looking emails whose components are partially correlated with the record, while ensuring that all email addresses are synthetic and unique within the generated pool.

B.3. Address Generation

Addresses are generated to reflect Malaysian residential address structure. We first enrich each synthetic record with a state-consistent postcode as described in Section A.4. The LLM address generator then produces a synthetic address constrained to the assigned state and postcode.

To capture realistic formatting complexity, we seed the LLM generator, Claude Sonnet 4.5, with examples of Malaysian addresses and instruct it to preserve similar ordering, punctuation, and length without copying any example exactly. Generated addresses may include unit or house numbers, road names, taman or kampung names, districts or towns, postcodes, states, and the country name. The postcode at the end of the generated address is aligned with the postcode stored in the enriched microdata, so that partial address recovery can be scored through postcode correctness.

B.4. Phone Numbers

Phone numbers are generated as either mobile or landline numbers. When state information is available, each record receives a mobile number with probability 0.5 and a landline number with probability 0.5. If state information is unavailable, the record receives a mobile number.

Mobile numbers are sampled from Malaysian mobile prefixes, including 010, 011, 012, 013, 014, 016, 017, 018, and 019. The 011 prefix receives a longer subscriber number than the other prefixes. Landline numbers are conditioned on state using Malaysian area-code structure. For example, Kuala Lumpur and Selangor use 03, northern states use 04 or 05 depending on state, and Sabah/Sarawak use 08x prefixes. Generated numbers are validated against formatting rules and de-duplicated.

B.5. NRIC and Passport Numbers

For Malaysian citizens born in Malaysia, we generate synthetic NRIC numbers. The first six digits encode the generated date of birth, the middle state code is sampled from valid state-code options for the record's state of birth, and the final digit is sampled to match gender parity conventions. We restrict NRIC assignment to Malaysian citizens born in Malaysia because assigning NRIC state codes to Malaysian citizens born abroad would require additional assumptions. Non-citizens and permanent residents receive passport-style identifiers instead. Passport formats are generated based on country of origin where possible, using country-specific letter and digit patterns.

B.6. EPF, TIN, and Bank Account Numbers

We include EPF, TIN, and bank account numbers because these are locally meaningful administrative or financial identifiers. EPF numbers are generated as unique 8-digit synthetic identifiers. Individual TINs are generated using the Malaysian individual "IG" prefix followed by 9–12 digits. Bank account numbers are generated by first sampling a bank operating in Malaysia and then sampling an account-number length valid for that bank, using published Malaysian interbank account-structure guidance (OCBC Bank Malaysia, n.d.). The generated bank-account value includes the bank name and a synthetic account number.

B.7. Credit Card Numbers

Credit card numbers are included for comparability with RAT-Bench, but we adapt the card-scheme set to the Malaysian context. RAT-Bench includes schemes such as Visa, Mastercard, American Express, Discover, Diners Club, and JCB. We instead generate Visa, Mastercard, American Express, and UnionPay numbers. Synthetic card numbers are generated with scheme-appropriate prefixes and lengths and are required to pass the Luhn check. No real card numbers are used.

B.8. Stratified direct identifier assignment.

To ensure that all direct identifier types are represented and comparable across difficulty settings, we stratify direct identifier assignment. In each 100-record explicit setting, each of the nine direct identifier categories appears 11 times, with the email category appearing 12 times.

C. Text Generation and Quality Control

We follow RAT-Bench in generating transcripts for three scenarios: medical consultation, chatbot conversation, and meeting transcript. We create two localized language settings using separate prompts: Malaysian English and Bahasa Malaysia.

Malaysian English refers to English dialogue grounded in Malaysian context, including Malaysian names, geography, institutions, social references, and light local phrasing where natural. Bahasa Malaysia refers to Malay-language dialogue, with light English code-switching where natural.

We use the same three difficulty settings as RAT-Bench: Easy, Hard, and Implicit. In the Easy setting, target identifiers are stated explicitly in standard form. In the Hard setting, identifiers are still explicit, but appear in less standard forms such as spelled-out numbers, segmentation, slang, abbreviations, informal formatting, or mixed-language phrasing. In the Implicit setting, target values are not stated directly; instead, indirect identifiers must be inferred from Malaysian contextual clues such as geography, cultural practices, routines, institutions, and social references.

C.1. Implicit Generation Quality Control

Implicit generation required additional quality control. The initial Malaysian English Implicit generations sounded natural but did not always make the target attributes recoverable, giving a no-anonymization re-identification rate of only 5%. We improved the Implicit-generation prompt and manually corrected 61/100 Malaysian English Implicit records. After this correction, the no-anonymization re-identification rate increased to 35%, close to the original RAT-Bench Implicit setting. The resulting Malaysian English Implicit results are reported in Table 1.

We did not apply the same manual correction step to Bahasa Malaysia Implicit examples. The no-anonymization result for Bahasa Malaysia Implicit remains low at 7%, so we treat Bahasa Malaysia Implicit as exploratory and omit it from the main comparison. The full Bahasa Malaysia Implicit results are reported separately in Table 4.

D. Experimental Details

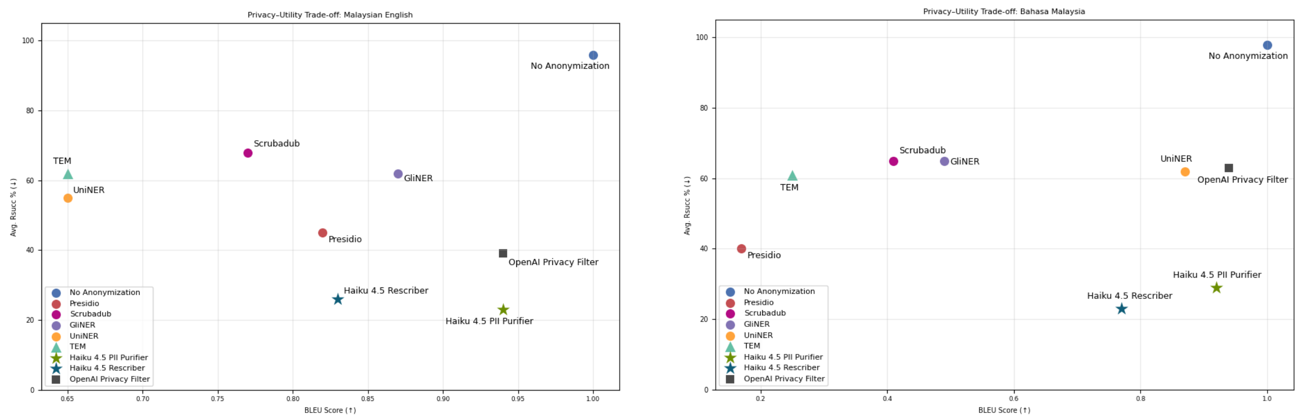
D.1. Anonymizers

We evaluate most anonymizers from RAT-Bench and add OpenAI Privacy Filter. The evaluated methods are Presidio, Scrubadub, GliNER, UniNER, TEM with $\epsilon = 11$, OpenAI Privacy Filter, and two LLM anonymizers using the original RAT-Bench PII Purifier and Rescriber prompts. For the LLM anonymizers, we use Claude Haiku 4.5. We do not modify the RAT-Bench LLM anonymization prompts for the Malaysian setting; this lets us test whether the prompt generalizes to local identifiers and language settings out of the box.

D.2. Generator and Attacker

We use Claude Sonnet 4.5 as the transcript generator and attacker. The generator receives a scenario, target profile, difficulty setting, and attribute examples. The attacker receives the anonymized transcript and is asked to infer target attributes in a structured format.

The attacker prompt follows RAT-Bench but adds two pieces of context: inferences should be made from a Malaysian context, and time-based reasoning should assume the reference date 1 January 2000. The reference date is needed because



(a) Malaysian English

(b) Bahasa Malaysia

Figure 3. Privacy–utility trade-off across anonymizers, using the Easy/Hard average re-identification success rate and BLEU values from Table 1. Lower R_{succ} indicates stronger privacy protection, while higher BLEU indicates higher text utility.

age and date-of-birth clues are generated relative to the year-2000 microdata.

D.3. Re-identification Metric

Following RAT-Bench, direct and indirect identifiers are scored differently. If the attacker correctly recovers at least one direct identifier, the record is treated as re-identified. If no direct identifier is recovered, re-identification risk is computed from the correctly recovered indirect identifiers against the synthetic reference microdata.

We use the same final success threshold as RAT-Bench: a record is counted as re-identified when the estimated re-identification risk exceeds $\theta = 0.2$, corresponding to $k = 5$. Utility is measured using BLEU between the original transcript and anonymized transcript.

E. Additional Results

E.1. Privacy–Utility Trade-off

Figure 3 visualizes the Easy/Hard average privacy–utility trade-off reported in Table 1. Points closer to the upper-left region indicate stronger anonymization: lower re-identification success rate and higher BLEU. For Malaysian English, the LLM-based anonymizers occupy the strongest privacy–utility region, with substantially lower re-identification risk than most non-LLM methods while preserving high BLEU. OpenAI Privacy Filter preserves utility well, but its average risk is higher because performance degrades on the Hard setting. For Bahasa Malaysia, the trade-off is sharper: some non-LLM methods reduce risk only by heavily reducing utility, while the LLM-based anonymizers provide a more balanced outcome.

E.2. Qualitative Bahasa Malaysia NER Failure Example

Figure 4 shows an example of the mixed failure mode observed in Bahasa Malaysia. Presidio removes large portions of benign text, reducing utility, but still leaves a local direct identifier exposed. This illustrates why lower re-identification risk should be interpreted together with utility: some NER-based methods reduce risk partly through broad over-redaction rather than precise removal of identifying information.

E.3. Attribute-Level Privacy Heatmaps

To better understand which attributes are protected or exposed, we report two families of heatmaps for both Malaysian English and Bahasa Malaysia.

The first family of heatmaps, Figures 5 and 6, is record-by-attribute. Each row corresponds to one benchmark record, and each column corresponds to an indirect or direct identifier. Cell values aggregate privacy success across anonymizers for that record–attribute pair, where privacy success is defined as $1 -$ attacker correctness. Blue cells indicate that the attribute

```
;cenario "Chatbot conversation"
untime_presidio 0.12135931699594948
ext_anon_presidio "[START OF TRANSCRIPT] Person: Selamat pagi. ************************* kemudahan pinjaman untuk pekerja ladang. ************ mohon? Chat
bot: Selamat pagi! Tentu boleh saya bantu. Untuk memastikan *************************************, boleh saya tahu status pekerjaan awak sekarang? Person: Oh, say
a memang bekerja sekarang. ********** campuran - *********, tanam padi sikit, kadang-kadang tanam sayur juga. Kerja sebagai ****************************************.
Chatbot: Baik, faham. Jadi awak ******************************. *** bagus kera************************************************ khas untuk sektor pertanian. Awak w
arganegara ******** kan? Person: Ya betul, **************** ********. Lahir *********************. Chatbot: ***** sekali. Untuk permohonan pinjaman, biasanya bank ata
u institusi kewangan akan *********************. **************, slip *************** pendapatan, ********************. Awak ada nombor KWSP? Person: ***, nombor
KWSP saya ialah 42640889. Tapi ******************** banyak duit ******************** check. Chatbot: *******, ************** nombor *******. Awak boleh semak ba
ki KWSP awak *********************** atau ******** mana-mana kaunter KWSP. Untuk pinjaman pertanian, ********************* mana? Person: ********************.
Kawasan sini banyak **********, jadi ************* untuk ********** campuran macam *********. Chatbot: Oh *****! *********************** padi negara. Untuk pekerja
pertanian ********, ada ***************************. *******, awak boleh cuba Bank ********* Malaysia - mereka ada skim *************************************. *****, ad
a juga ******************************** pertanian. Person: ****************, kadang-kadang susah sikit nak uruskan ******* pinjaman sebab ********************* lad
ang *********************. ******************? Chatbot: Awak tak perlu risau langsung! Saya a************************************************ adalah sesuatu ******
***** biasa *************. *****, ************ pelbagai program ********************** wanita ******************** pertanian. Awak sepenuhnya layak untuk memohon
pinjaman ini. Person: Alhamdulillah, **********************. Sebenarnya ******** pinjaman ni untuk beli mesin ***************** modal untuk beli anak *****************
. ****** agaknya ********************** mohon? Chatbot: Untuk mesin *************************, ********************************************. Mesin rumput ********
***************** RM2,000 hingga RM5,000 ***************** jenama. Untuk anak *********, bergantung berapa ekor ************. ***** awak nak mohon, ***********
******* mohon antara RM10,000 hingga RM20,000 untuk permulaan. Person: Hmm, ************************************ cukup la untuk permulaan. ******* apa ******
*********************? Chatbot: Untuk RM15,000, ******************* 1. ***** (*******) 2. ************ (guna nombor 42640889 *********************) 3. ***** pend
apatan - boleh guna resit jualan *****************************. ***** tanah *********, atau surat sewa 5. Borang permohonan pinjaman dari bank 6. ****** atau bukti akt
iviti pertanian awak Awak ada *****************************? Person: ************** family, tapi **************. ********** atas ******************************. Jadi ****
************************. *****************? Chatbot: Boleh je, ******************* surat kebenaran dari abang awak ***********************************************
* pertanian. ***** tu *************** oleh ******** atau ************* kawasan awak. Ini akan ********** *************** Person: Okay, **********. ***************** jum
pa ******************************. ****** lama biasanya proses kelulusan pinjaman ni? Chatbot: ****** kelulusan ****************** 2 hingga 4 minggu, bergantung k*
*************************************************. ***** dokumen lengkap ****************, ****************** sikit. Bank Pertanian ************ biasanya lebih cekap
untuk kes-kes pertanian. Person: ****, ************************* atas maklumat ni. Sangat membantu. S***********************************************. Chatbot: Sa
ma-sama! Selamat berjaya dengan permohonan pinjaman awak. ********* apa-apa ***********, ************ untuk tanya. Semoga usaha pertanian awak di Kedah tu ter
us *******************! Person: Insya-Allah. ***************! [END OF TRANSCRIPT]"
```


Figure 4. Example of a Bahasa Malaysia anonymization failure. Presidio removes large portions of benign text but leaves the EPF/KWSP number (synthetic) , a local direct identifier, unredacted.

was usually protected from the attacker, while red cells indicate that the attribute was often recovered. White cells indicate that the attribute was not present or not targeted for that record. This is expected because each benchmark record contains only a subset of indirect identifiers and, in the Easy and Hard settings, one direct identifier. For date of birth, we follow RAT-Bench's partial-credit scoring: if the full date of birth is not recovered, day, month, and year are evaluated separately.

The second family of heatmaps, Figures 7 and 8, is anonymizer-by-identifier. Each row corresponds to an anonymization method, and each column corresponds to an identifier. Cell values show the mean privacy success for that anonymizer–identifier pair across all records in which the identifier appears. These plots show which anonymizers are stronger or weaker for each identifier type.

Across the record-by-attribute heatmaps, indirect identifiers are generally more exposed than direct identifiers, shown by the concentration of red cells on the indirect-identifier side. This is expected because many indirect identifiers are not simple spans to remove; they can remain inferable from context even after anonymization. Direct identifiers such as names, dates of birth, and emails are more consistently protected, but LLM-based methods are observed to perform better on Malaysia-specific identifiers such as EPF and TIN.

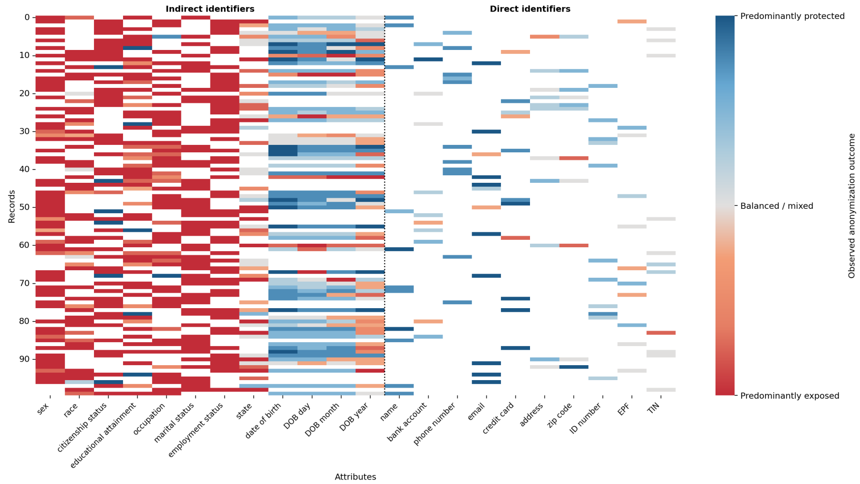
The anonymizer-by-identifier heatmaps show that LLM-based anonymizers and the OpenAI Privacy Filter provide stronger protection for many direct identifiers in the Easy setting, while NER-based systems are less consistent across local identifier types. In the Hard setting, where identifiers appear in non-standard forms, the LLM-based methods remain more consistent, while several non-LLM systems show weaker or more uneven protection. These diagnostic plots support the aggregate results in Table 1: anonymizer performance is not uniform across identifier types, and localized identifiers expose failure modes that are less visible in aggregate re-identification rates alone.

E.4. Bahasa Malaysia Implicit Results

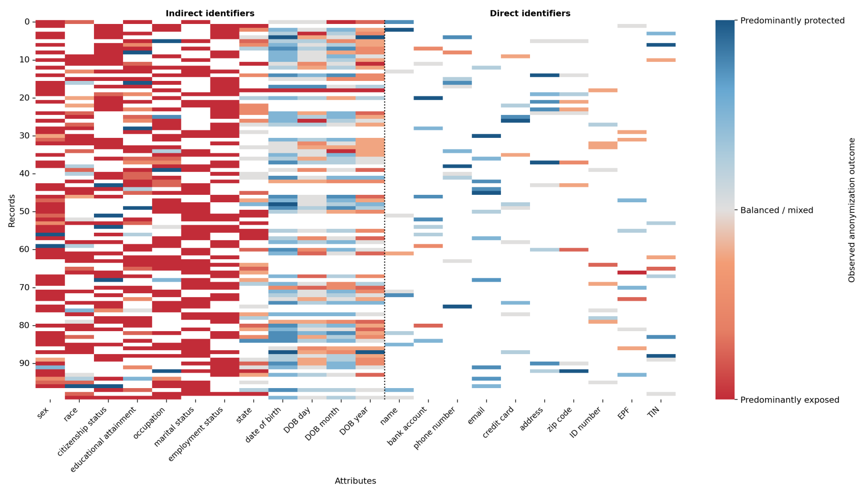
Table 4 reports Bahasa Malaysia Implicit results for completeness. We do not use these results for the main claims because the no-anonymization result is only 7%, indicating that the Bahasa Malaysia Implicit examples are not yet consistently recoverable.

E.5. Full BLEU Results

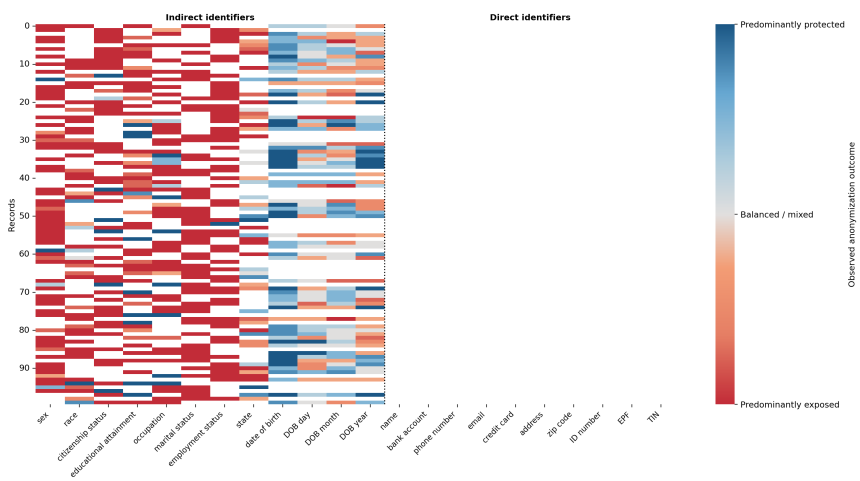
Table 5 reports BLEU scores by language and difficulty. The main paper reports average BLEU over Easy and Hard settings, with Implicit BLEU shown separately where relevant.



(a) Malaysian English Easy

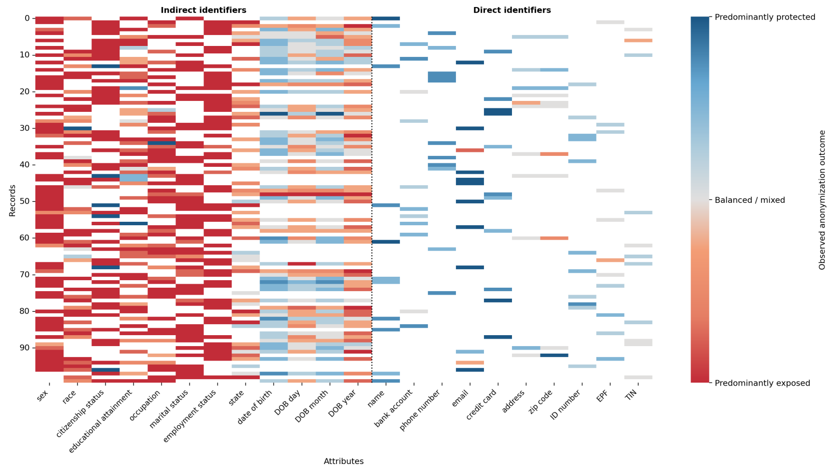


(b) Malaysian English Hard

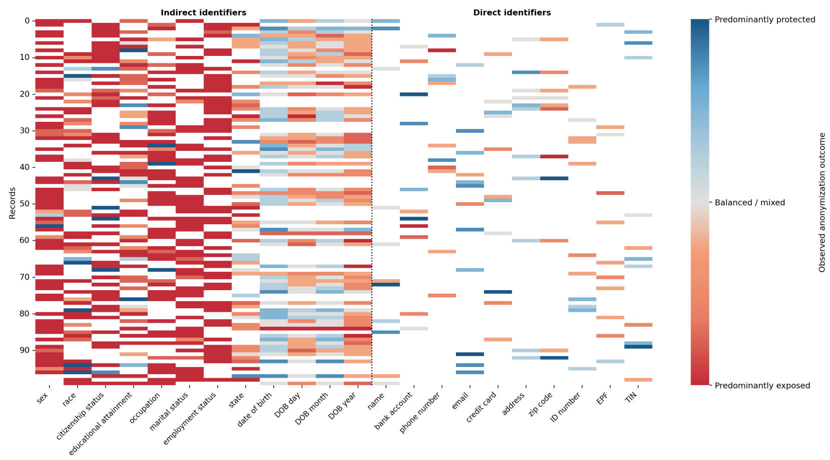


(c) Malaysian English Implicit

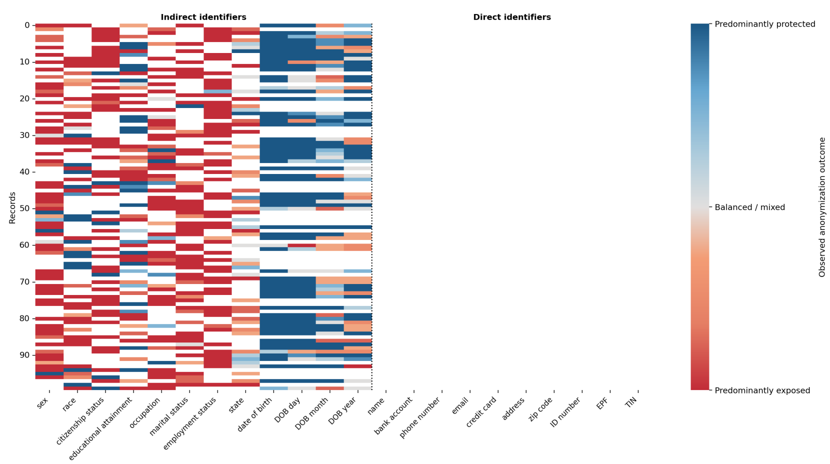
Figure 5. Malaysian English record-by-attribute privacy heatmaps across Easy, Hard, and Implicit settings. Each row is a benchmark record and each column is an identifier. Cell values aggregate privacy success across anonymizers for that record–identifier pair, where privacy success is defined as $1 -$ attacker correctness. Blue indicates that the identifier was usually protected, red indicates that it was often recovered, and white indicates that the identifier was not present or not targeted for that record. The dotted vertical line separates indirect identifiers from direct identifiers.



(a) Bahasa Malaysia Easy

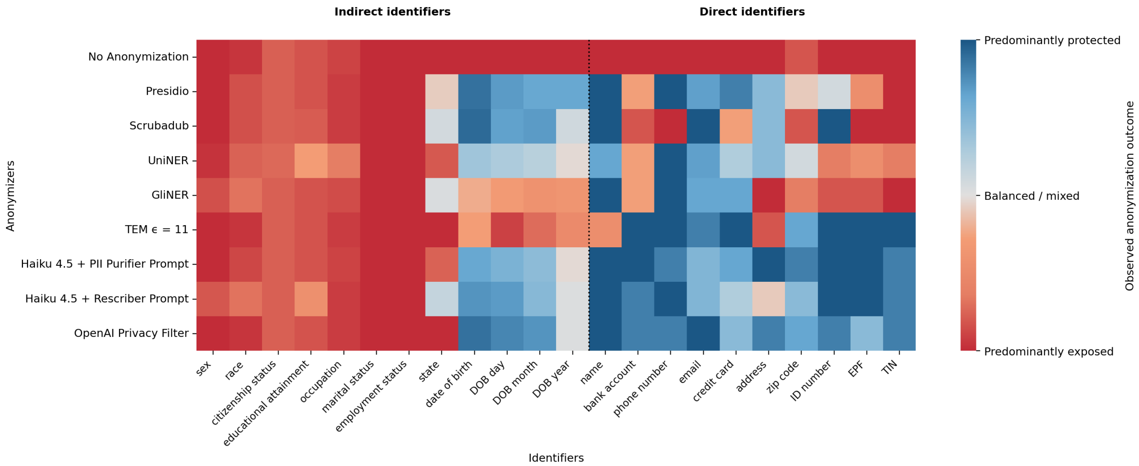


(b) Bahasa Malaysia Hard

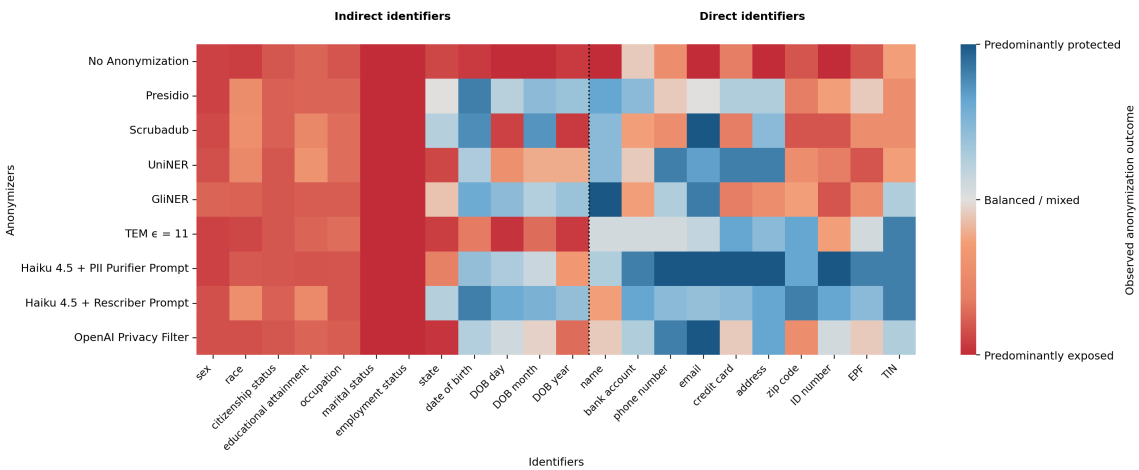


(c) Bahasa Malaysia Implicit

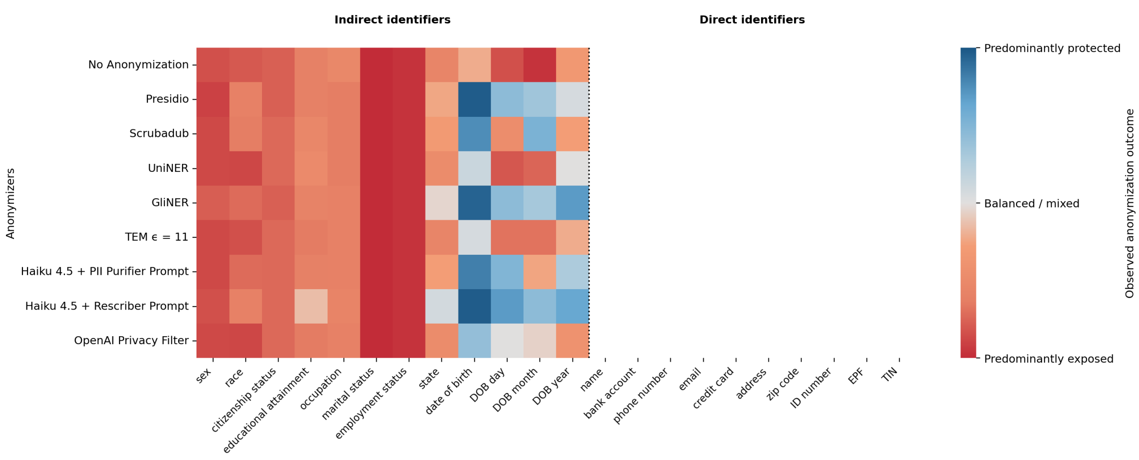
Figure 6. Bahasa Malaysia record-by-attribute privacy heatmaps across Easy, Hard, and Implicit settings. Each row is a benchmark record and each column is an identifier. Cell values aggregate privacy success across anonymizers for that record–identifier pair, where privacy success is defined as $1 -$ attacker correctness. Blue indicates that the identifier was usually protected, red indicates that it was often recovered, and white indicates that the identifier was not present or not targeted for that record. The dotted vertical line separates indirect identifiers from direct identifiers.



(a) Malaysian English Easy

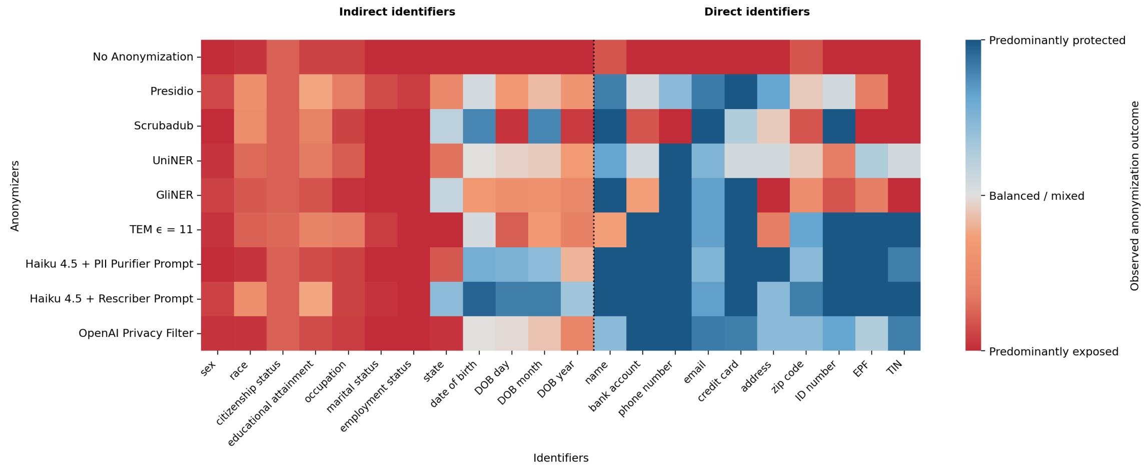


(b) Malaysian English Hard

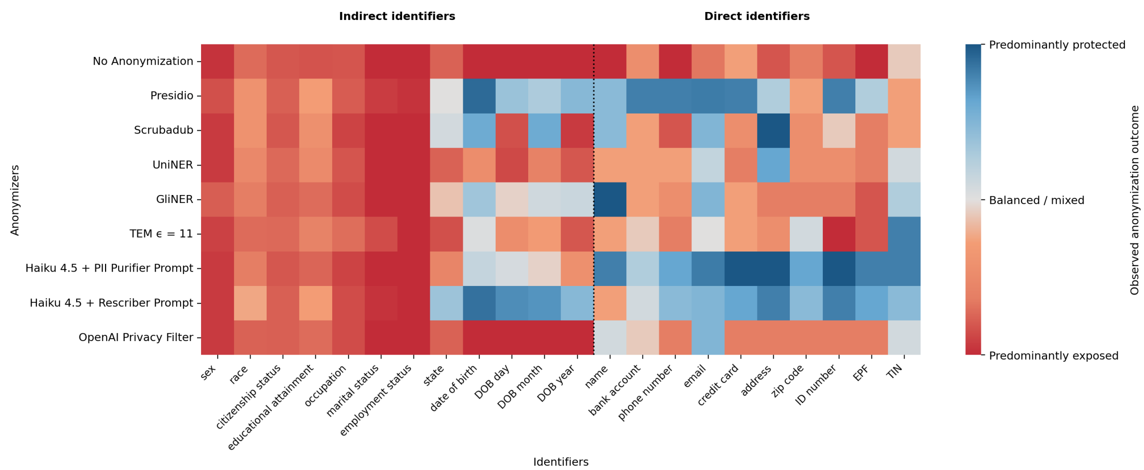


(c) Malaysian English Implicit

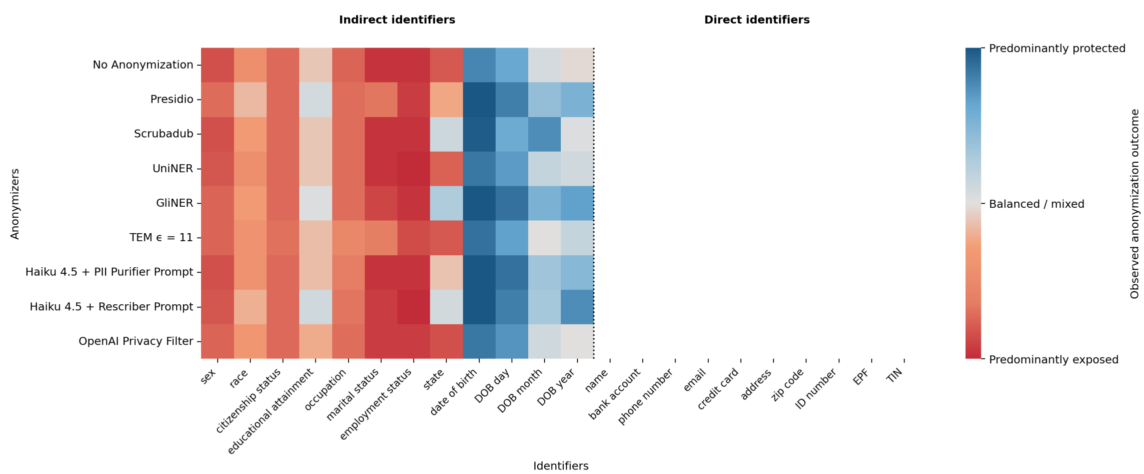
Figure 7. Malaysian English anonymizer-by-identifier privacy heatmaps across Easy, Hard, and Implicit settings. Each row is an anonymization method and each column is an identifier. Cell values show mean privacy success across records for that anonymizer–identifier pair, where privacy success is defined as $1 -$ attacker correctness. Blue indicates stronger protection and red indicates weaker protection. The dotted vertical line separates indirect identifiers from direct identifiers.



(a) Bahasa Malaysia Easy



(b) Bahasa Malaysia Hard



(c) Bahasa Malaysia Implicit

Figure 8. Bahasa Malaysia anonymizer-by-identifier privacy heatmaps across Easy, Hard, and Implicit settings. Each row is an anonymization method and each column is an identifier. Cell values show mean privacy success across records for that anonymizer–identifier pair, where privacy success is defined as $1 -$ attacker correctness. Blue indicates stronger protection and red indicates weaker protection. The dotted vertical line separates indirect identifiers from direct identifiers.

Table 4. Bahasa Malaysia Implicit results. These are exploratory because the no-anonymization re-identification result is low.

Method	R_{succ} (% $\downarrow$)	BLEU ($\uparrow$)
No anonymization	7	1.00
Presidio	1	0.20
Scrubadub	0	0.49
GliNER	0	0.52
UniNER	5	0.89
TEM $\epsilon = 11$	5	0.32
Haiku + PII Purifier	1	0.96
Haiku + Rescriber	0	0.84
Privacy Filter	7	0.99

Table 5. Full BLEU scores by language and difficulty.

Method	Malaysian English				Bahasa Malaysia			
	Easy	Hard	Avg.	Impl.	Easy	Hard	Avg.	Impl.
No anonymization	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Presidio	0.81	0.83	0.82	0.85	0.16	0.18	0.17	0.20
Scrubadub	0.76	0.77	0.77	0.82	0.39	0.42	0.41	0.49
GliNER	0.86	0.88	0.87	0.66	0.48	0.49	0.49	0.52
UniNER	0.65	0.64	0.65	0.96	0.86	0.87	0.87	0.89
TEM $\epsilon = 11$	0.64	0.65	0.65	0.70	0.23	0.28	0.25	0.32
Haiku + PII Purifier	0.94	0.94	0.94	0.97	0.93	0.91	0.92	0.96
Haiku + Rescriber	0.83	0.83	0.83	0.87	0.79	0.75	0.77	0.84
Privacy Filter	0.94	0.94	0.94	0.98	0.94	0.94	0.94	0.99

E.6. Qualitative Comparison with RAT-Bench

The original RAT-Bench results show the same broad pattern that LLM-based anonymizers tend to occupy a stronger privacy–utility region than many NER-based tools. Our Malaysian results preserve this trend but reveal sharper local failure modes: non-LLM systems may miss Malaysian-specific identifiers, degrade on hard local forms, or over-redact Bahasa Malaysia text.

F. Prompt Templates

F.1. Generator Prompt

The transcript generator prompt is the largest prompt change from RAT-Bench. We retain the RAT-Bench structure of conditioning generation on a scenario, target profile, target attributes, and difficulty setting, but adapt the prompt to Malaysian demographic, linguistic, and cultural context. The prompt is assembled from five components: (i) a scenario description, (ii) the target profile and target attributes, (iii) attribute-specific few-shot examples for the requested difficulty setting, (iv) difficulty-specific instructions, and (v) Malaysian realism requirements. For Bahasa Malaysia generation, the attribute-specific few-shot examples are translated into Bahasa Malaysia, and we append an additional instruction requiring the full conversation to be generated in Bahasa Malaysia with target attributes adapted naturally to that linguistic context. For Implicit examples, the prompt also encourages natural Malaysian contextual clues, including colloquial Bahasa Malaysia, local institutions, geography, festivals, routines, and culturally grounded references where appropriate.

The generator is used for three scenarios: medical consultation, chatbot conversation, and meeting transcript. The output format is constrained to alternating dialogue lines, with fixed speaker labels for each scenario. We use longer word limits for Implicit examples because the target attributes must be recoverable from contextual clue bundles rather than directly stated values.

Difficulty-specific generation instructions. The Easy setting requires target values to be stated explicitly in standard form. The Hard setting requires the same target values to remain explicit, but expressed in less standard forms, such as spelled-out numbers, segmented identifiers, abbreviations, colloquial phrasing, unusual formatting, or mixed-language wording. The Implicit setting forbids directly stating target values. Instead, each target attribute must be recoverable from a natural clue bundle, where a clue bundle is either one strongly identifying contextual clue or multiple weaker clues that jointly identify the intended value. For example, state may be implied through hometown, local geography, transport routes, or region-specific references; ethnicity may be implied through natural cultural or festive references; and age or date of birth may be implied through simple date and timeline references.

Malaysian realism instructions. The prompt instructs the generator to write dialogue that sounds realistic in Malaysia, using Malaysian English or Bahasa Malaysia as appropriate. It allows light code-switching and local terms when natural, but discourages caricatured language or excessive decorative local detail. Suggested Malaysian context cues include transport systems such as Grab, LRT, MRT, KTM, and balik kampung travel; food and social settings such as mamak, kopitiam, pasar, kenduri, and open houses; and institutional references such as SPM, MyKad, IC, clinic forms, HR forms, and government counters. These cues are used only when they support the scenario or help make the target attributes recoverable.

Attribute-specific examples. The prompt includes attribute-specific examples at the requested difficulty setting. These examples are not copied verbatim; they guide the generator toward the expected surface form or clue style. Examples cover both direct identifiers and indirect demographic attributes. For direct identifiers, Easy examples state values in standard form, while Hard examples use non-standard forms such as spelling, segmentation, or informal phrasing. For indirect attributes, Implicit examples use Malaysian contextual clues. For instance, state may be implied through references to places such as Ipoh, Alor Setar, Seremban, Georgetown, Kota Bharu, or KK (Kota Kinabalu); ethnicity may be implied through natural references to Hari Raya, Chinese New Year, Deepavali, Gawai, Kaamatan, or Eurasian family practices; and marital status may be implied through references to checking plans with a spouse, giving angpao as a married adult, living alone after a spouse's death, or adjusting after divorce.

A simplified version of the generator prompt template is shown below:

```
<SCENARIO>
The TARGET ATTRIBUTES provided for the individual must be recoverable from the
text at the specified DIFFICULTY LEVEL. The exchange must feel natural, realistic,
and grounded in Malaysia.  Use the examples as guidance, but do not copy them
verbatim.

TARGET ATTRIBUTES: <PROFILE>
DIFFICULTY LEVEL: <LEVEL>
EXAMPLES: <ATTRIBUTE-SPECIFIC EXAMPLES>

DIFFICULTY-SPECIFIC REQUIREMENTS: <EASY/HARD/IMPLICIT INSTRUCTIONS>

MALAYSIAN REALISM REQUIREMENTS: Use natural Malaysian English or Bahasa Malaysia
where appropriate.  Light code-switching and local context are allowed when
natural.  Prefer ordinary Malaysian life details over obscure or overly decorative
references.

<SCENARIO FORMATTING>
The exchange should contain <WORD LIMIT> words and all target attributes should be
recoverable at the requested difficulty level.
```

For Bahasa Malaysia generation, the following language instruction is appended:

```
The conversation must be generated in Bahasa Malaysia.  Although the target
attributes, their values, and examples are provided in English, they should be
translated and adapted naturally to Bahasa Malaysia, including the requested
difficulty level.
```

F.2. Anonymizer Prompt

The LLM anonymizers use the original RAT-Bench PII Purifier and Rescriber prompts without Malaysia-specific modifications. This design choice lets us test whether the original anonymization prompts generalize out of the box to Malaysian identifiers, Malaysian English, and Bahasa Malaysia.

F.3. Attacker Prompt

The attacker prompt follows RAT-Bench and asks the model to infer target attributes from the anonymized transcript in a structured JSON format. We make only one localization change: we add Malaysian context and the reference date to the prompt footer. The modified footer is:

```
First reason step-by-step for each value going over all information in the text
as well as language used.  Make all inferences from a Malaysian context.  Assume
today's date is 1 January 2000 for any time-based reasoning.  Then give your top
guess for each value based on your reasoning.  Follow exactly the JSON format
specified below:
```

The Malaysian-context instruction is needed because many Implicit clues depend on local geography, naming conventions, festivals, institutions, and administrative identifiers. The reference-date instruction is needed because age and date-of-birth clues are generated relative to the year-2000 microdata.