HalluSination: A Multi-Task, Multi-Domain Sinhala Hallucination Benchmark

Madhushankha de Silva, Akindu Himan, Dineth Randula,
Dilanka Heshan, Thimira Sahan, Nishan Kavinda, Nevidu Jayatilleke,
Dasuni Rathnayaka, Nisansa de Silva
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{bawantha.22, akindu.22, dineth.22, dilanka.22, thimira.22,
nishan.22, nevidu.25, dasunir.26, nisansa}@cse.mrt.ac.lk

Abstract

Hallucination in large language models (LLMs) is a critical challenge, particularly for low-resource languages where fluent outputs may lack supporting evidence. Although hallucination has been extensively studied in high-resource settings, Sinhala lacks a dedicated benchmark study. We introduce HALLUSINATION, the first benchmark for hallucination in Sinhala. We evaluated 14 LLMs and three conventional machine translation systems (Google Translate, NLLB, and MADLAD-400) on approximately 600 source-grounded examples across machine translation and question answering, spanning news, legal, and cultural domains. Native speakers manually annotated all 9,000+ model outputs using a task-specific hallucination taxonomy. The results reveal substantial hallucination across all models, with pronounced variation by task and domain. In question answering, human-judged faithfulness ranges from 38.7% to 97.0%, and correct refusal ranges from 2.6% to 100%; in machine translation, faithfulness ranges from 0.7% to 92.5%, and hallucination ranges from 0.0% to 28.4%. This work[1] provides the first systematic characterization of hallucination patterns in Sinhala NLP.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse natural language processing tasks. However, they remain fundamentally constrained by hallucination: generating fluent outputs that are inconsistent with either their pre-training corpus or the input context (Bang et al., 2025). In high-stakes applications, such hallucinations undermine user trust and model reliability.

Hallucination detection and mitigation have been extensively studied in high-resource languages, especially English, with benchmarks such as HaluEval (Li et al., 2023) and HalluLens (Bang et al., 2025) characterizing hallucinations in question answering, dialogue, and summarization. Multilingual machine translation (MT) studies also show that hallucinations are more frequent in low-resource directions and that detection methods are degraded for pairs of low-resource languages (Guerreiro et al., 2023a).

Recent work has extended hallucination evaluation to lower-resource languages: Multi-Hall-SA covers English and four South African languages (Ralethe and Buys, 2026); BHRAM-IL shows higher hallucination in Hindi and Odia (Terdalkar et al., 2025); and a 21-language European QA benchmark reports elevated hallucination in lower-resource languages such as Icelandic (Thoresen and Smart, 2026).

Although Sinhala, an Indo-Aryan language spoken by more than 20 million people, has seen a growing development of resources for QA, MT and general language understanding (Guzmán et al., 2019; Ranasinghe et al., 2025; Pramodya et al., 2025; Ranathunga et al., 2024), none of these works specifically targets hallucination or faithfulness analysis.

We present HALLUSINATION, the first benchmark for hallucination in Sinhala. We curate a 600-item evaluation set for machine translation and question answering from existing Sinhala and multilingual resources, supplemented with manually developed examples to improve coverage of legal, cultural, and unanswerable questions. We publicly release the HALLUSINATION dataset[2]. The benchmark spans general, news, legal, and cultural domains and evaluates 14 LLMs and three conventional MT systems: Google Translate, NLLB (Costa-jussà et al., 2022), and MADLAD-

[1] https://github.com/hungrypyalpha-star/HalluSination

[2] https://huggingface.co/datasets/madhushankhades/SinHallu-600

400 (Kudugunta et al., 2023). Native Sinhala speakers annotate more than 9,000 model outputs using task-specific taxonomies, enabling an analysis of hallucination patterns across models, tasks, and domains. We further find that automatic overlap metrics, including EM, Token F1, and chrF, show only moderate agreement with human judgments. HALLUSINATION provides the first systematic characterization of hallucination in Sinhala NLP and a reusable framework for evaluating source-grounded generation in Sinhala and other low-resource languages.

2 Related Work

2.1 Hallucination in NLG

Hallucination has become an important reliability concern across generative natural language processing tasks. It occurs when a model generates information that is not supported by the input, source, or available evidence. Ji et al. (2023) provide a broad study of hallucination in natural language generation, while Huang et al. (2025) discuss different forms of hallucination, including factuality and faithfulness errors. These ideas are important for MT and QA because the generated output should remain consistent with the given source or supporting passage.

2.2 Hallucination Benchmarks & Annotation

There are several benchmarks developed to study hallucination. HaluEval (Li et al., 2023) evaluates hallucinated outputs in tasks such as QA, dialogue, and summarization. HalluLens (Bang et al., 2025) studies different hallucination scenarios in LLMs. Instead of simply labeling a whole response as hallucinated or non-hallucinated, recent works have used more detailed annotation. ANAH (Ji et al., 2024) provides sentence-level hallucination annotation. RAGTruth (Niu et al., 2024) identifies unsupported spans. RAGChecker (Ru et al., 2024) checks individual claims against supporting evidence. These studies show that fine-grained annotation can better clarify where hallucinations occur.

2.3 Hallucination in Machine Translation

Hallucination has also been widely studied in machine translation. Xu et al. (2023) investigate hallucination in neural machine translation, while Guerreiro et al. (2023a) show that hallucination is an important problem in multilingual translation, particularly for low-resource language pairs. Dale et al. (2023b) provide a manually annotated multilingual benchmark for hallucination and omission detection. Benkirane et al. (2024) further study hallucination detection in both high- and low-resource languages. More recently, Chatterjee et al. (2026) introduce fine-grained human annotation for translation hallucinations, allowing specific hallucinated parts of translations to be identified.

2.4 Hallucination in Question Answering

Hallucination is also a concern in question answering. DelusionQA studies unsupported information in domain-specific QA, while RelD investigates hallucination detection in generated answers (Sadat et al., 2023; Chen et al., 2023). Multilingual research shows that this problem should also be studied outside English. MUCH examines hallucination in QA across many languages, while Multi-Hall-SA provides a benchmark for several low-resource South African languages (Ul Islam et al., 2025; Ralethe and Buys, 2026). BenHalluEval similarly introduces a multi-task hallucination benchmark for Bengali (Adib et al., 2026). These studies show the value of developing language-specific hallucination benchmarks for low-resource languages.

2.5 Sinhala NLP Resources

For Sinhala, several useful NLP datasets already exist. FLORES provides Sinhala–English machine translation data. In contrast, Belebele, SiQuAD, and SinhalaMMLU provide resources for reading comprehension, QA, and general language understanding (Guzmán et al., 2019; Bandarkar et al., 2024; Pramodya et al., 2025). However, these datasets were developed mainly for evaluating task performance. They were not designed specifically for hallucination analysis. They also do not provide detailed human-annotated labels showing which parts of Sinhala model outputs are unsupported by the source.

2.6 Research Gap

Therefore, a clear research gap remains. Existing work provides multilingual hallucination benchmarks, fine-grained annotation methods, and Sinhala resources for MT and QA, but there is no dedicated Sinhala hallucination benchmark covering both tasks. The proposed work addresses this gap by creating a source-grounded, human-annotated Sinhala benchmark for machine translation and

question answering across multiple domains. The benchmark will identify different types of hallucinations and provide detailed annotations that can support future research on hallucination detection and analysis in Sinhala.

3 Data Creation

We construct approximately 300 items per machine translation (MT) and question answering (QA) tasks, covering general, news, legal, and cultural domains. All items are screened for quality by removing near-duplicates and discarding instances with ambiguous references, offensive content, or degraded Sinhala orthography. Representative items are provided in Appendix B.

Machine Translation. The English-to-Sinhala MT set contains 306 items in 4 domains: News, Legal, Cultural, and General. *General* and *News* items are manually selected from FLORES (Guzmán et al., 2019), with *General* aggregating diverse everyday content (e.g., travel, sports, politics) as a baseline. *Cultural* items are drawn from the non-parallel Sinhala Global-PIQA subset (Chang et al., 2025; de Silva and Ranathunga, 2026), and *Legal* items were manually constructed by aligning bilingual Sinhala–English texts from the Sri Lankan Constitution, the Penal Code, and relevant Acts.

Question Answering. In this task, given a Sinhala source passage and a question, the model must generate an answer based on the context or abstain if the passage lacks sufficient evidence. The QA set contains 302 items: 226 short-answer and 76 unanswerable. Short-answer questions comprise 150 items from SiQuAD (Ranasinghe and Weerasinghe, 2025) and 77 multiple-choice items from Belebele (Bandarkar et al., 2024), adapted to short-answer form by retaining the correct option as a reference. Unanswerable questions were constructed manually for FLORES passages (Guzmán et al., 2019). Each question refers to the entities or events mentioned in the passage, but asks for information that cannot be determined from the available context (Rajpurkar et al., 2018).

4 Annotation

All annotations are produced by native Sinhala speakers following a calibration round. We use task-specific hallucination taxonomies for both MT and QA. Full label definitions, subtypes, and worked examples are included in the Appendices C and D.

Machine Translation. Each translation receives one primary label: *Faithful*, *Omission*, *Mistranslation*, *Degenerated Repetitive Output*, or *Hallucination*. Instances labeled *Hallucination* are further assigned one of six subtypes: extraneous hallucination, repetition, detached hallucination, incorrect target language / untranslated content, partial hallucination, or other.

The taxonomy is derived from prior work on MT hallucination, including the LLM-focused taxonomy of Wu et al. (2025), the detachment-based framework of Guerreiro et al. (2023b), HalOmi (Dale et al., 2023b), HAT (Chatterjee et al., 2026), and Dale et al. (2023a). Following HAT, we distinguish *Mistranslation* from *Hallucination* by the degree of semantic connection to the source.

Three annotators independently labelled 1,020 translations (60 source sentences across 17 systems). Agreement was measured on whole-output categorical labels, not error spans: primary labels achieved 72.39% mean pairwise agreement (Fleiss' $\kappa = 0.59$), while hallucination sub-labels yielded a mean pairwise Jaccard of 0.37 (Appendix D.2.3).

For a small subset of items, annotators judged the provided Sinhala reference translation to be inaccurate during annotation. In these cases, we evaluate faithfulness against the source English text instead of the Sinhala reference.

Question Answering. Each answer receives one of five labels: *Supported*, *Unsupported*, *Correct refusal*, *False refusal*, or *Degenerated Repetitive Output*. Full label definitions are given in Appendix C.1.

This taxonomy adopts an intrinsic, source-grounded notion of hallucination: an answer is hallucinated when it asserts information that cannot be located in or reasonably inferred from the provided passage. Thus, *Unsupported* responses are hallucinations, whereas *Correct refusals* are not. *False refusals* and *Degenerated Repetitive Outputs* are treated as distinct failure modes because they concern abstention and generation behaviour rather than unsupported content.

The taxonomy follows the intrinsic-hallucination framing of HaluLens (Bang et al., 2025) and the answer/abstain distinction of DelucionQA (Sadat et al., 2023). Annotators apply a source-verification test adapted from AIS (Rashkin et al., 2023). Full

guidelines, boundary cases, and worked examples are provided in Appendix C.1.

A random 20% subset of the QA dataset was independently annotated by all three annotators, yielding 92.0% three-way agreement (Fleiss' $\kappa = 0.90$; full results in Appendix C.3).

5 Experimental Setup

Models. We evaluate 14 LLMs from closed-source (Gemini, GPT, Claude) and open-weight families (Llama, DeepSeek, Qwen, Mistral), along with three conventional MT baselines (Google Translate, NLLB (Costa-jussà et al., 2022), MADLAD-400 (Kudugunta et al., 2023)). The complete list of models and versions is provided in Appendix E. By varying model family and scale, this selection allows us to analyze the impact of model capacity on hallucination behavior.

Metrics. For MT, we report COMET (Rei et al., 2020), chrF/chrF3 (Popović, 2015) and GEMBA-MQM (Kocmi and Federmann, 2023). For QA, we report Exact Match (EM) and token-level F1 on answerable items, a no-answer EM score on unanswerable items (Rajpurkar et al., 2018) (1 if the model's output exactly matches the reference refusal token `UNANSWERABLE`, 0 otherwise), and chrF/chrF3 (Popović, 2015) across all QA items.

6 Results and Analysis

6.1 Question Answering

Hallucination remains unsolved for Sinhala QA, although the strongest systems approach near-ceiling performance. Ten of 14 LLMs achieve at least 83% human-judged precision, and nine produce Unsupported responses on fewer than 5% of answerable questions. However, several systems generate Unsupported content for 24–40% of answerable questions, and models differ in whether they abstain when the passage lacks sufficient evidence. We therefore evaluate overall accuracy (Table 1, Appendix H, and Figure 5), hallucination on answerable questions, refusal behavior on unanswerable questions (Figure 1), and the relationship between automatic reference-overlap metrics and human judgments.

6.1.1 Model-level Results

Table 1 gives model-wise results, with human accuracy ranging from 97.0% (Claude-Opus-4.6) to 38.7% (Llama-3.1-8B); ten models score at least 83%, whereas Qwen3.5-4B, Mistral-Large, and both Llama-3.1 models score below 68%. Figure 5 visualizes this distribution across all models.

Performance does not follow a simple closed-source vs. open-weight split. Although all five closed-source models demonstrate consistently strong performance, several open-weight models remain highly competitive, led by DeepSeek-V4-Pro (93.0%), DeepSeek-V4-Flash (91.7%), and Llama-4-Maverick (90.1%). Furthermore, substantial variations in hallucinations exist within model families. For Qwen, increasing model size from 4B to 9B and 27B parameters improves accuracy from 67.2% to 83.1% and 85.8%, respectively. Llama-4-Maverick also substantially outperforms the earlier Llama-3.1 models.

6.1.2 Faithfulness on Answerable Questions

On answerable questions, Unsupported rates range from 1.3% to 40.3%. Nine models achieve rates below 5%; among these, the frontier models generally remain faithful to the information in the Sinhala passage.

However, model size alone does not guarantee faithfulness. Qwen3.5-4B and Mistral-Large produce Unsupported responses for 23.9% and 26.1% of answerable questions, respectively, while Llama-3.1-8B and Llama-3.1-70B reach 38.1% and 40.3%. Notably, Llama-3.1-70B never refuses an answerable question, yet only 58.4% of its responses are Supported, demonstrating that high answer propensity does not imply faithful reasoning.

6.1.3 Refusal Behaviour

Unanswerable questions expose the largest behavioural divergence across models, reflecting differences in confidence calibration and abstention. Gemini-3.6-Flash and Qwen3.5-27B correctly refuse all unanswerable questions, whereas Llama-3.1-8B correctly refuses only 2.6% and produces unsupported responses for 76.3% of them, indicating that weaker models struggle primarily with recognizing insufficient evidence. False Refusal on answerable questions is lower (0.0%-16.4%), but the trade-off between avoiding hallucination and maintaining coverage varies substantially.

Figure 1a illustrates this: Llama-3.1-70B and Llama-3.1-8B produce Unsupported responses for 61.8% and 76.3% of unanswerable questions, respectively. In contrast, Qwen3.5-27B avoids Unsupported responses entirely, but has the highest False Refusal rate on answerable questions (16.4%), suggesting a conservative safety-first pol-

Model	Answerable: overlap metrics				Answerable: human labels			Unanswerable			Overall	
	EM	F1	chrF	chrF3	Sup.↑	Uns.↓	False Ref.↓	NoAns-EM↑	Corr. Ref.↑	Uns.↓	**Acc↑**	Len
Closed-source												
Gemini-3.1-Pro	25.2	54.6	62.4	69.9	95.6	1.8	2.7	98.7	98.7	1.3	96.4	1.0×
Gemini-3.6-Flash	11.9	44.5	54.1	61.5	91.2	1.3	7.5	100.0	100.0	0.0	93.4	1.3×
GPT-5.5	15.5	47.6	59.0	67.3	94.7	4.9	0.4	85.5	86.8	13.2	92.7	1.2×
Claude-Opus-4.6	10.2	36.8	48.7	56.2	**97.3**	2.2	0.4	82.9	96.1	3.9	97.0	2.8×
Claude-Sonnet-4.6	3.1	26.5	42.1	48.4	95.6	1.8	2.7	92.1	96.1	3.9	95.7	4.6×
Open-weight												
Llama-4-Maverick	23.0	51.1	58.0	66.1	**96.0**	3.1	0.9	72.4	72.4	27.6	90.1	1.0×
Llama-3.1-70B	19.5	33.3	36.5	43.8	58.4	40.3	0.0	31.6	31.6	61.8	51.7	1.0×
Llama-3.1-8B	3.5	16.9	25.5	33.0	50.9	38.1	0.4	0.0	2.6	76.3	38.7	4.0×
DeepSeek-V4-Pro	22.7	49.8	57.8	65.4	92.5	2.7	4.9	94.7	94.7	5.3	93.0	1.0×
DeepSeek-V4-Flash	22.3	45.9	52.5	60.1	90.3	3.1	6.6	94.7	96.1	3.9	91.7	1.0×
Qwen3.5-27B	16.4	41.8	48.2	55.3	81.0	2.7	16.4	100.0	100.0	0.0	85.8	1.1×
Qwen3.5-9B	17.3	41.9	49.4	56.6	83.2	8.0	8.8	67.1	82.9	17.1	83.1	1.0×
Qwen3.5-4B	8.8	28.8	37.0	39.9	63.3	23.9	12.8	64.5	78.9	21.1	67.2	1.1×
Mistral-Large	0.4	15.1	25.4	31.8	68.1	26.1	5.8	5.3	21.1	78.9	56.3	6.7×

Table 1: QA results for all 14 systems (%). Human-label rates use all 226 answerable or 76 unanswerable items, as appropriate. Automatic metrics are averaged over the available non-missing scores. **Sup.**: Supported. **Uns.**: Unsupported. **False Ref.**: False Refusal. **Corr. Ref.**: Correct Refusal. **Acc**: human accuracy over all 302 items (Supported + Correct refusal). **Len**: median per-item length ratio between answers judged Supported and their reference spans. Bold marks the best result per column among Sup. and Corr. Ref.

icy. The zoomed panel in Figure 1b shows that Gemini-3.1-Pro achieves the best balance among the low-error systems, with 1.3% Unsupported responses on unanswerable questions and 2.7% False Refusal on answerable questions. This suggests stronger internal calibration, which is important for practical tabular QA where both hallucination and over-refusal degrade trust.

6.1.4 Automatic vs. Human Evaluation

Consistent with prior work showing that lexical QA metrics can underestimate semantically equivalent answers (Risch et al., 2021), automatic overlap metrics are more pessimistic than human judgments. Exact Match ranges from 0.4% to 25.2%, and the highest token F1 is 54.6%, whereas 10 of 14 models achieve human-supported rates above 80% on answerable questions; character-based chrF/chrF3 scores are consistently higher, reflecting greater tolerance for wording and inflection variation. Metric-based rankings differ from human judgments. For example, Claude-Opus-4.6 ranks ninth by token F1 but first by Supported rate (Appendix H, Figure 4). Answer length helps explain this: the answer-to-reference length ratio is negatively correlated with token F1 ($\rho = -0.71$) and EM ($\rho = -0.92$), so overlap metrics penalize longer, explanatory answers even when judged grounded. We therefore treat these metrics as measures of reference overlap rather than substitutes for human faithfulness evaluation.

6.1.5 No-Answer Exact Match

SQuAD 2.0-style no-answer EM is stricter than human refusal labels because it requires models to output only the specified refusal token (`UNANSWERABLE`). For Mistral-Large, Qwen3.5-9B, Qwen3.5-4B, and Claude-Opus-4.6, human correct-refusal exceeds no-answer EM by 13.2–15.8 points. These models often correctly refuse in Sinhala but add an explanation before or after the token, despite the output-format instruction. We therefore use no-answer EM as a format-compliance measure and human labels for substantive refusal quality.

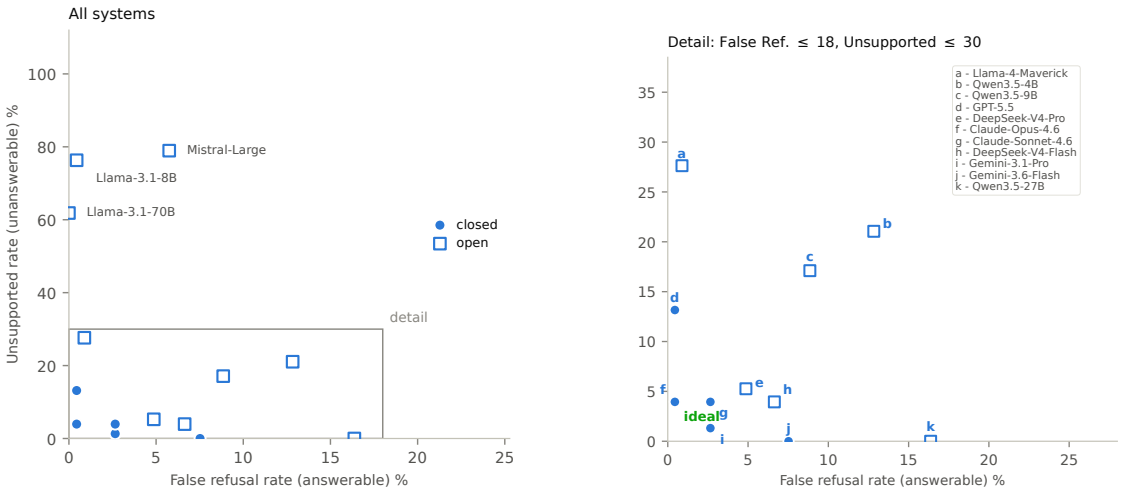
Finally, repetition loops (Holtzman et al., 2019) appear in two Llama systems. Llama-3.1-8B produces 40 degenerate outputs (13.2%), and Llama-3.1-70B produces 8 (2.6%). Following our annotation scheme, we report these cases as *Degenerated Repetition Output*, separately from Unsupported responses.

6.2 Machine Translation

Hallucination remains a major issue in English-to-Sinhala MT, accounting for 11.1% of 5,202 human-annotated outputs. Although the strongest systems hallucinate at most 2.0%, four open-weight systems exceed 25%, and eight LLMs fall below 20% human-judged faithfulness (Table 2; Figures 6–10).

6.2.1 Model-level Results

Human-judged faithful rates span three orders of magnitude, from 92.5% (Gemini-3.1-Pro) to



(a) All systems.

(b) Detail view of the boxed region.

Figure 1: False Refusal on answerable questions against Unsupported responses on unanswerable questions.

	Human labels (%)					Automatic metrics				
System	**F**	**O**	**M**	**D**	**H**	**chrF**	**chrF3**	**Token F1**	**COMET**	**GEMBA**
Closed-source										
Gemini-3.1-Pro	**92.48**	0.00	7.52	0.00	0.00	59.82	73.52	0.4688	0.9110	99.86
Gemini-3.6-Flash	88.56	0.00	9.48	0.00	1.96	58.27	72.20	0.4513	0.9084	99.58
GPT-5.5	76.80	0.33	18.63	0.00	4.25	55.55	69.93	0.4217	0.9001	99.05
Claude-Opus-4.6	73.86	0.33	21.57	0.00	4.25	55.35	70.00	0.4266	0.8980	98.37
Claude-Sonnet-4.6	56.21	0.00	35.29	0.00	8.50	52.32	67.48	0.3939	0.8845	97.06
Open-weight										
DeepSeek-V4-Pro	**84.64**	0.00	14.05	0.00	1.31	56.54	70.60	0.4373	0.9015	98.99
DeepSeek-V4-Flash	76.47	0.65	19.61	0.65	2.61	54.22	68.67	0.4201	0.8935	98.13
Llama-4-Maverick	45.75	1.63	39.22	0.00	13.40	51.75	66.74	0.3873	0.8812	95.23
Qwen3.5-27B	25.49	2.29	58.17	2.61	11.44	45.48	61.70	0.3143	0.8552	91.54
Llama-3.1-70B	16.34	9.15	43.79	15.03	15.69	37.18	52.36	0.2497	0.7437	85.39
Qwen3.5-9B	8.17	3.59	63.73	5.56	18.95	39.13	55.20	0.2655	0.7993	86.74
Mistral-Large	5.88	11.76	56.86	0.33	25.16	37.10	54.68	0.2284	0.7398	84.48
Qwen3.5-4B	0.98	2.94	53.92	16.01	26.14	43.64	49.19	0.2670	0.7073	81.66
Llama-3.1-8B	0.65	4.25	9.80	56.86	28.43	14.49	24.37	0.0584	0.4224	70.48
Dedicated MT systems										
Google Translate	**73.20**	1.63	22.55	0.00	2.61	56.14	69.76	0.4431	0.8925	97.09
MADLAD-400-3B	20.26	9.15	39.22	13.73	17.65	39.04	53.23	0.2752	0.7724	86.30
NLLB-600M	16.67	27.12	46.08	3.59	6.54	41.87	56.48	0.3033	0.8241	90.43

Table 2: English-to-Sinhala MT results over 306 outputs per system. Human-label values are percentages, while automatic metrics retain their original scales. Systems are grouped by system type and ordered within each group by human-judged Faithful rate.

F = Faithful; O = Omission; M = Mistranslation; D = Degenerated Repetitive Output; H = Hallucination. Automatic scores are size-weighted means of the supplied domain-level values.

0.7% (Llama-3.1-8B). Three LLMs (Gemini-3.1-Pro, Gemini-3.6-Flash, DeepSeek-V4-Pro) exceed 84% faithful, while eight LLMs fall below 20%. Notably, dedicated MT systems are not uniformly inferior: Google Translate (73.2% Faithful) outperforms eight LLMs, including several large open-weight models,but lags behind the three highest-performing LLMs by 11-19 percentage points.

Model class interacts strongly with failure mode. Closed-source LLMs exhibit near-zero degenerate repetitive output and hallucination (both $\leq 2.0\%$), whereas open-weight models show substantial degeneration (up to 56.9% for Llama-3.1-8B) and hallucination (up to 28.4%). Within model fami-

lies, larger variants are consistently more faithful and degenerate less: Qwen3.5 improves from 1.0% Faithful at 4B to 25.5% at 27B, and Llama-3.1-70B reduces degeneration from 56.9% to 15.0% relative to the 8B variant (Appendix I, Figure 8). However, size alone does not prevent hallucination: Mistral-Large (25.2%) and Qwen3.5-4B (26.1%) hallucinate more than smaller dedicated MT systems.

6.2.2 Error and Hallucination Patterns

Across all models and MT systems, mistranslation is the dominant error (32.9% of outputs; 59.7% of non-faithful cases), followed by hallucination (11.1%), degenerate repetitive output (6.7%), and omission (4.4%). This distribution implies that low hallucination rates do not entail high quality: NLLB-600M hallucinates on only 6.5% of inputs but achieves just 16.7% faithful due to high mistranslation (46.1%) and omission (27.1%).

Hallucination behavior is highly model-specific. Three models (Gemini-3.1-Pro, Gemini-3.6-Flash, DeepSeek-V4-Pro) record $\leq$ 2.0% hallucination, while four open-weight models exceed 25%. Among 578 hallucinated outputs, partial hallucination dominates (260 cases; 45.0%), followed by detached hallucination (94; 16.3%), incorrect target language or untranslated content (80; 13.8%), and extraneous hallucination (62; 10.7%). Extraneous hallucinations concentrate in News (53/62 cases), whereas detached hallucinations cluster in cultural and general domains (83/94 cases). This domain-specific concentration suggests that hallucination subtype is not random but tied to content type and possibly training data composition.

Degenerate repetitive output is almost entirely a small-model phenomenon: Llama-3.1-8B alone accounts for 49.7% of all degeneration cases (174 outputs), with Qwen3.5-4B, Llama-3.1-70B, and MADLAD-400-3B contributing most of the remainder. We treat degeneration as a distinct failure mode because it reflects generation collapse rather than semantic drift.

A qualitative comparison of selected translations reveals distinct failure patterns (Appendix 11). All 17 systems fail on two examined sentences, while Google Translate succeeds on a legal sentence where the three highest-ranked models fail. We also observe different error categories for the same inputs, particularly among weaker models.

6.2.3 Domain and Topic Effects

Domain difficulty is model-dependent but consistently severe for legal content: Legal achieves the lowest faithful rate overall (30.1%) and is the lowest or tied-lowest domain for every model. However, Legal's poor performance stems primarily from mistranslation (44.8%) and degeneration (12.3%), not hallucination (6.0%). By contrast, general and news exhibit the highest hallucination rates (14.8% each), and general has the largest combined hallucination plus degeneration (23.3%) effects. Cultural content, while achieving the highest faithfulness rate (52.6%), shows the lowest automatic metric scores, indicating a mismatch between human and automatic evaluation in this domain (Figure 2).

The domain-wise IAA analysis shows that Legal, despite having the lowest model faithfulness rate, achieved the highest annotator agreement (76.73%, Fleiss' $\kappa = 0.6594$). General had the lowest agreement (63.79%, $\kappa = 0.4881$). Thus, the domain in which models performed worst was not the domain with the greatest disagreement among human annotators. The full breakdown is provided in Table 7.

Within General, 12 topics show considerable variation, but nine topics contain ≤ 5 sentences per system. We therefore treat topic-level results as exploratory (Appendix Figure 9).

6.2.4 Automatic vs. Human Evaluation

At the model level, all five automatic metrics correlate strongly with the human Faithful rate. Spearman correlations are $\rho = 0.934$ for chrF, 0.961 for chrF3, 0.949 for Token F1, 0.985 for COMET, and 0.983 for GEMBA-MQM. COMET and GEMBA-MQM therefore reproduce the overall human system ordering most closely. COMET agrees with the human Faithful ordering for 132 of the 136 possible model pairs (97.1%), while GEMBA-MQM agrees for 131 pairs (96.3%).

Aggregate metric agreement does not establish sentence-level reliability for hallucination, fluency, or style (Ranathunga et al., 2026). GEMBA-MQM and COMET achieve sentence-level hallucination ROC-AUCs of 0.770 and 0.687, respectively, rising to 0.941 and 0.821 against Faithful outputs alone. This suggests that other translation errors complicate hallucination detection (Appendix F).

Several disagreements also appear below the overall system level. Google Translate ranks seventh by human Faithful rate but third by Token F1 and fourth by chrF, showing that lexical over-

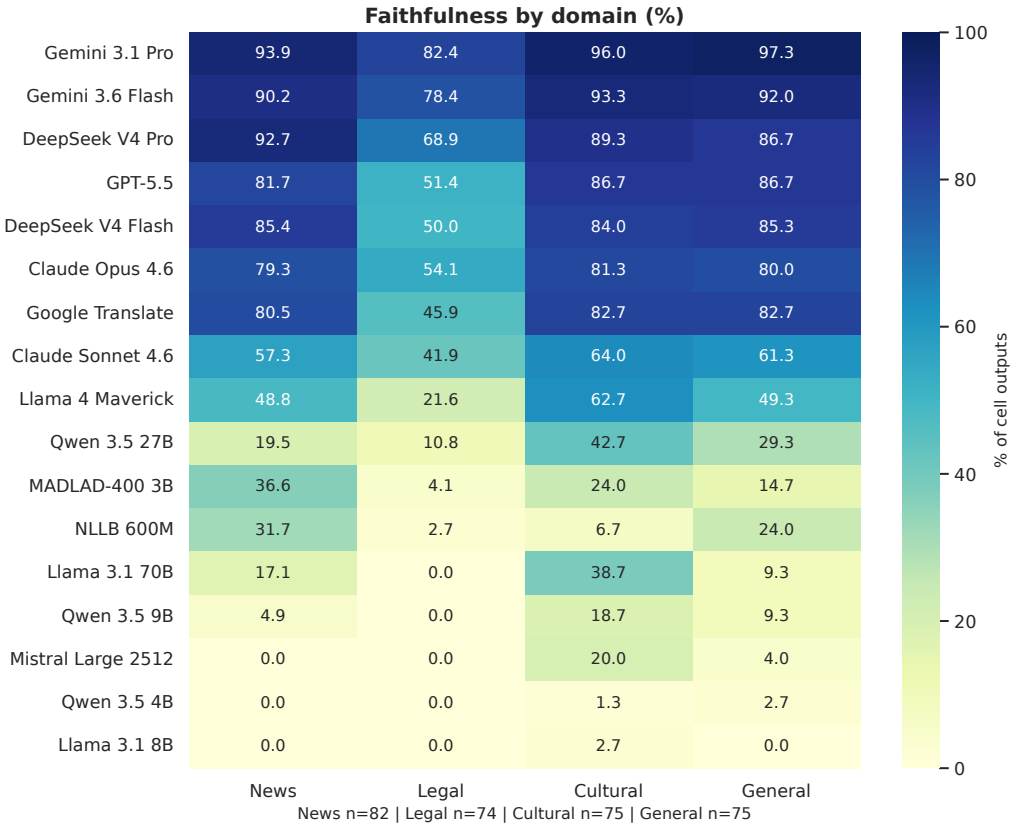


Figure 2: Human-judged faithfulness rates by translation system and domain.

lap favors it more than human annotations do. At the domain level, Cultural has the highest human Faithful rate but the lowest pooled chrF, chrF3, Token F1, and COMET scores. The clearest local disagreement occurs for Qwen3.5-4B on News: it obtains the highest News chrF score among all 17 systems (67.28), despite producing no Faithful translations out of 82 inputs; 25 of these outputs are labelled as Hallucinations. Appendix Figure 10 compares GEMBA-MQM and COMET across all 68 model–domain combinations. Although both metrics generally increase with the human Faithful rate, GEMBA-MQM exhibits a stronger rank association ($\rho = 0.954$) than COMET ($\rho = 0.830$), and similarly high automatic scores can still accompany substantially different human judgments.

These results show that automatic metrics are informative for coarse system ranking, particularly COMET and GEMBA-MQM, but they do not fully capture domain-specific behavior or individual hallucination cases. We therefore use human labels for claims about faithfulness, hallucination, and error type, while treating automatic metrics as complementary measures of translation quality and reference similarity.

7 Conclusion

We present HALLUSINATION, the first benchmark for hallucination in Sinhala, covering machine translation and question answering across multiple domains. Analysis of 17 MT systems, 14 QA models, and 9,000+ human annotations reveals systematic hallucination across all tasks, with closed-source LLMs exhibiting near-zero hallucination in MT. In contrast, open-weight models reach up to 28.4%. In QA, accuracy ranges from 38.7% to 97.0%, and correct refusal on unanswerable questions varies from 2.6% to 100%. Automatic metrics correlate strongly with human judgments for MT system ranking but only moderately for QA, and fail to capture domain-specific hallucination patterns.

Our findings underscore the need for human-centered, source-grounded evaluation in low-resource settings and caution against relying on token-overlap metrics alone. Future work can extend HALLUSINATION to summarization, dialogue, and long-form generation; cover additional domains and genres of Sinhala text; and develop Sinhala-specific methods for hallucination detection and mitigation.

Limitations

Our benchmark has several limitations. First, although we cover multiple domains, the dataset is limited in size and does not include all genres of Sinhala text (e.g., social media, conversational dialogue, or highly technical domains), which may affect the generalizability of our findings. Second, our annotation protocol relies on native Sinhala speakers with specific linguistic and domain expertise; differences in annotator background or training could lead to variation in hallucination labels. Finally, while we analyze correlations between automatic metrics and human judgments, we do not propose or evaluate new metrics, leaving open the question of how best to automate hallucination detection in Sinhala.

References

Shefayat E Shams Adib, Ahmed Alfey Sani, Ekramul Alam Esham, Ajwad Abrar, Ishmam Tashdeed, and Md Taukir Azam Chowdhury. 2026. BenHalluEval: A multi-task hallucination evaluation framework for large language models on Bengali. *arXiv preprint arXiv:2605.31483*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, et al. 2025. Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*.

Rajen Chatterjee, Xintong Li, Paisarn Charoenpornsawat, and Allen Lee. 2026. HAT: Hallucination annotation for translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15865–15888, San Diego, California, United States. Association for Computational Linguistics.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

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

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.

Nisansa de Silva. 2026. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v27*.

Nisansa de Silva and Surangika Ranathunga. 2026. Sinhala physical common sense reasoning dataset for Global PIQA. *arXiv preprint arXiv:2602.02207*.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference*

of the European Chapter of the Association for Computational Linguistics, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Nevidu Jayatilleke, Nisansa de Silva, Uthpala Nimanthi Sooriya-Arachchi, Gagani Kasundhi Kulathilaka, Azra Safrullah, and Johan Nevin Sofalas. 2026. SiDiaC-v.2.0: Sinhala diachronic corpus version 2.0. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 6740–6763, Palma de Mallorca, Spain. ELRA Language Resource Association.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ashmari Pramodya, Nirasha Nelki, Heshan Shalinda, Chamila Liyanage, Yusuke Sakai, Randil Pushpananda, Ruvan Weerasinghe, Hidetaka Kamigaito, and Taro Watanabe. 2025. SinhalaMMLU: A comprehensive benchmark for evaluating multitask language understanding in Sinhala. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32943–32961, Suzhou, China. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Sello Ralethe and Jan Buys. 2026. Multi-hall-SA: A cross-lingual benchmark for multi-type hallucination detection in low-resource South African languages. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6282–6296, Rabat, Morocco. Association for Computational Linguistics.

Janani Ranasinghe and Ruvan Weerasinghe. 2025. Question answering in a low-resource language: Dataset and deep learning adaptations for sinhala. In *International Conference on Deep Learning Theory and Applications*, pages 336–352. Springer.

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, and Ruslan Mitkov. 2025. Sinhala encoder-only language models and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8623–8636, Vienna, Austria. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Surangika Ranathunga, Nisansa de Silva, Aloka Fernando, Kavindu Warnakulasuriya, Isuru Wijesiri, Menan Velayuthan, Charitha Rathnayaka, Thivaharan Varatharajan, Sajeevi Silva, Piumi Kandanaarachchi, and Uthayasanker Thayasivam. 2026. EnSiTa - A Trilingual Multi-Domain Parallel Dataset and Benchmark for Domain-Specific Machine Translation. *arXiv preprint arXiv:2609.29511*.

Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian's, Malta. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Hrishikesh Terdalkar, Kirtan Bhojani, Aryan Dongare, and Omm Aditya Behera. 2025. BHRAM-IL: A benchmark for hallucination recognition and assessment in multiple Indian languages. In *Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)*, pages 102–116, Mumbai, India. Association for Computational Linguistics.

Freja Thoresen and Dan Saattrup Smart. 2026. A multilingual hallucination benchmark. In *Proceedings of the Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026)*, pages 187–192, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).

Saad Obaid Ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How much do LLMs hallucinate across languages? on realistic multilingual estimation of LLM hallucination. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29077–29098, Suzhou, China. Association for Computational Linguistics.

Xinwei Wu, Heng Liu, Jiang Zhou, Xiaohu Zhao, Linlong Xu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Challenging multilingual LLMs: A new taxonomy and benchmark for unraveling hallucination in translation. *arXiv preprint arXiv:2510.24073*.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

A Background on Sinhala

Sinhala is the mother tongue of approximately 17 million people in Sri Lanka, where it holds official-language status. It belongs to the Indo-Aryan branch of the Indo-European language family and uses a Brahmi-derived abugida script in which consonant–vowel sequences are represented as single orthographic units. The modern Sinhala lexicon also contains loanwords from Tamil, English, Portuguese, and Dutch (de Silva, 2026).

Sinhala exhibits diglossia, with distinct literary and spoken varieties. The literary variety is head-final with subject–object–verb (SOV) word order, while spoken Sinhala allows more flexible constituent order. Sinhala is highly inflected and has a rich case-marking system that distinguishes animate and inanimate nouns, contributing to substantial token-level variation and alignment challenges in machine translation and source-grounded question answering (de Silva, 2026; Ranathunga et al.,

2026). Despite its written tradition of more than two millennia, Sinhala remains under-resourced in natural language processing, with limited tools, datasets, and evaluation benchmarks compared to high-resource languages (Ranathunga and de Silva, 2022).

B Example Datapoints

Table 3 shows representative data points from the MT and QA tracks. The MT example is a FLORES plus English–Sinhala sentence pair about dinosaur feathers, used to check whether models preserve factual content despite fluent output. The QA examples include answerable and unanswerable cases; the latter describe a problem (overscan-induced subtitle cropping) but provide no technical fix, so the gold label is UNANSWERABLE. These items test whether systems correctly refuse to answer when evidence is missing and otherwise ground answers strictly in the Sinhala text.

C Qusetion-Answering Annotation

Each answer receives one of four labels: *Supported*, *Unsupported*, *Correct refusal*, *False refusal*, or *Degenerated Repetitive Output*. These labels jointly capture whether the source contains the required information and whether the model chooses to respond or abstain. This taxonomy follows the intrinsic-hallucination framing of HaluLens (Bang et al., 2025), adopts the answer/abstain distinction of DelucionQA (Sadat et al., 2023), and uses a source-verification test adapted from AIS (Rashkin et al., 2023).

C.1 Question-Answering Hallucination Taxonomy

- **Supported** – The answer is fully grounded and supported by the provided source context.
- **Unsupported** – The answer cannot be verified using the provided source context.
- **Correct refusal** – The model appropriately abstains from answering when the question is unanswerable from the given context.
- **False refusal** – The model refuses to answer when the question is actually answerable from the provided context.
- **Degenerated repetitive output** – Text that exhibits excessive repetition, incoherence, or generic filler content, indicating generation collapse rather than semantic error (Holtzman et al., 2019).

These five labels jointly cover two independent axes of model behavior: whether the question is answerable from the source, and whether the model chooses to answer or abstain. Representative examples of each label are provided in Appendix B.

C.2 Question-Answering Annotation Guidelines

The question-answering annotation was carried out by a team of three native Sinhala speakers (A1, A2, A3). Before annotation, all three completed a calibration round, jointly discussing sample question–answer pairs to align their interpretation of the four labels.

To identify the Supported/Unsupported distinction, annotators apply the source-verification test from the Attributable to Identified Sources (AIS) framework (Rashkin et al., 2023), which stipulates that generated content about the given source is attributable if and only if the source supports the claim being made. Concretely, annotators ask, for each response:

1. Can every factual claim in the answer be directly located or reasonably inferred from the source passage?
2. If the question is unanswerable from the passage, does the response explicitly abstain rather than asserting a specific answer?

If the answer to (1) is "no" for any claim, the response is labeled *Unsupported*, regardless of whether the claim happens to be true. Annotators are instructed to default to *Unsupported* when grounding is ambiguous. Figure 3 summarizes this decision procedure.

C.3 Inter-annotator Agreement of Question-Answering Task

To assess inter-annotator agreement, a subset(20%) of the QA dataset was independently annotated by all three annotators, blinded to each other's labels. Table 4 reports pairwise and three-way agreement and Fleiss' κ.

C.4 Question-Answering Labeling Examples

Table 5 shows representative examples of our QA labeling scheme. For an unanswerable question, a model response that names a specific software

Task	Example
MT (News)	**Source (EN):** "This not only confirms that at least some dinosaurs had feathers, a theory already widespread, but provides details fossils generally cannot, such as color and three-dimensional arrangement." **Reference (SI):** මෙයින්, දැනටමත් ප්‍රසිද්ධ වූ න්‍යායයක් වන සමහර ඩයිනෝසරවලට පිහාටු තිබුණු බව සනාථ කිරීම පමණක් නොව, ෆොසිලවල සාමාන්‍යයෙන් සැපයිය නොහැකි, ත්‍රිමාණ බව සහ වර්ණය වැනි විස්තර ද සැපයේ.
QA (Unanswerable)	**Passage (excerpt, SI):** චිත්‍රපටියක් ඩීවීඩී ආකෘතියට පරිවර්තන කිරීමට උත්සාහ කරන විට ඇති බව ඉතාම පොදු ගැටලුවලින් එකක් අධිපරිලෝකනය වෙයි... (overscan cropping subtitles near the frame edge) **Question (SI):** ඡේදයට අනුව, ඩීවීඩී එකක් සෑදීමේදී ඕවර්ස්කෑන් නිසා උපසිරැසි කප්පාදු වීම වැළැක්වීම සඳහා භාවිත කළ හැකි මෘදුකාංගය හෝ තාක්ෂණික විසඳුම කුමක්ද? *("What software or technical solution can be used to prevent subtitle cropping caused by overscan when creating a DVD?")* **Gold label:** UNANSWERABLE – the passage describes the overscan problem but names no specific software or technical fix

Table 3: Representative examples from the MT and QA subsets.

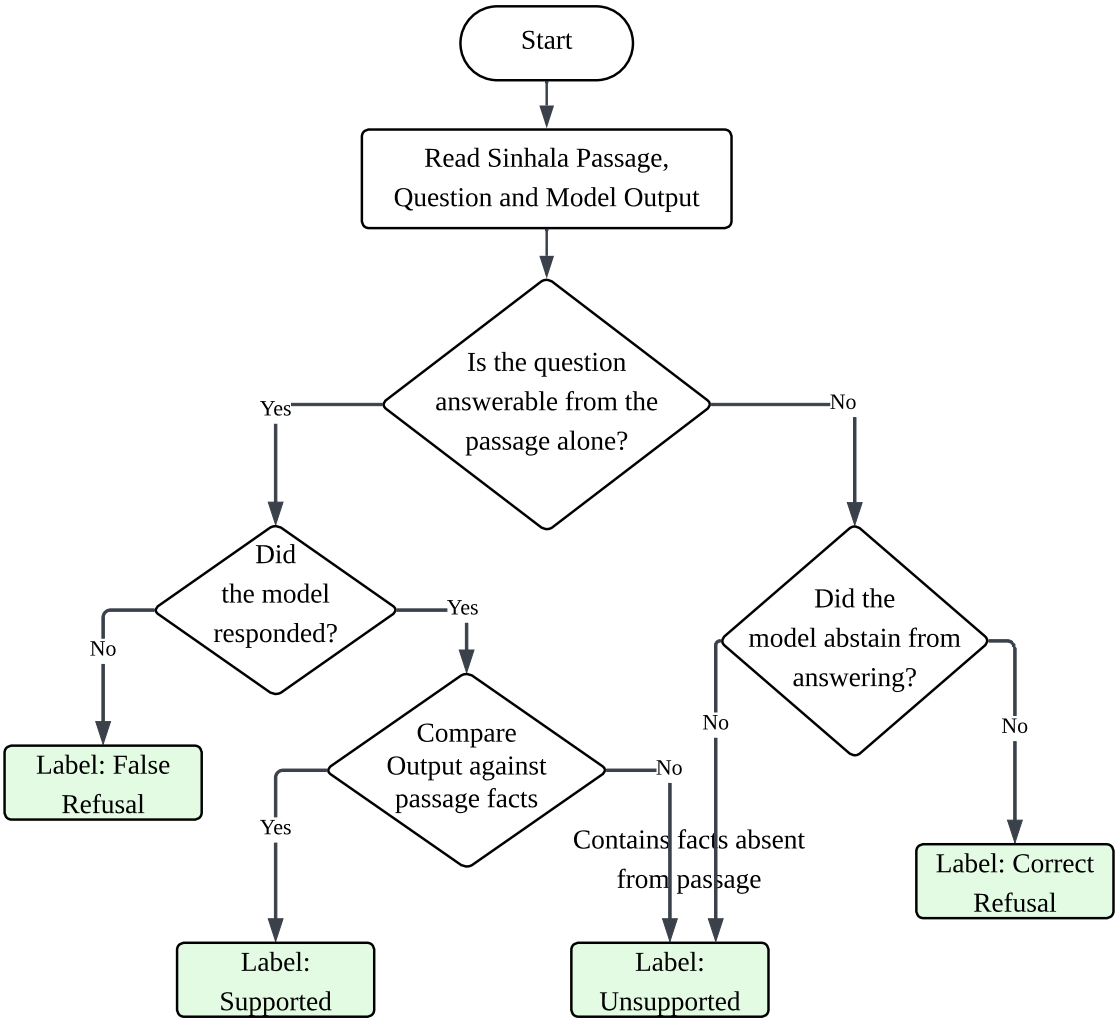


Figure 3: Decision procedure for assigning QA labels.

Annotators	Agreement (%)	κ
A1 – A2	94.52%	0.9010
A1 – A3	93.81%	0.8903
A2 – A3	95.60%	0.9195
Three-way	**92.02%**	**0.9034**

Table 4: Inter-annotator agreement for the QA annotations.

tool is labeled *Unsupported*, since the passage provides no such information. In contrast, a response that explicitly states the passage does not specify a solution is labeled *Correct refusal*, reflecting appropriate abstention. For an answerable question, a response that directly matches information in the passage is labeled *Supported*.

D Machine Translation Annotation

The annotation taxonomy used in this benchmark is adapted and synthesized from established taxonomies and annotation frameworks for machine translation errors and hallucinations. In particular, the definitions and distinctions draw on the hallucination taxonomy of Wu et al. (2025), the detachment-based framework of Guerreiro et al. (2023b), HalOmi (Dale et al., 2023b), HAT (Chatterjee et al., 2026), and the hallucination taxonomy of Dale et al. (2023a). We adopt the primary-label structure and hallucination subcategories to the requirements of our English-to-Sinhala MT annotation task, while refining the operational criteria to address language-specific and context-dependent cases. Following HAT, the distinction between *Mistranslation* and *Hallucination* is based on the degree of semantic connection between the erroneous target content and the source. The detailed definitions and annotation rules used by the annotators are provided below.

D.1 Machine Translation Hallucination Taxonomy

For machine translation, the primary labels are defined as follows:

- **Faithful** – The translation preserves the meaning of the source without introducing unsupported content.
- **Omission** – Important source content is missing from the translation.
- **Mistranslation** – Content is present and meaningfully grounded in the source, but is rendered incorrectly, such as through an incorrect word form, spelling, inflection, terminology, etc.
- **Degenerated repetitive output** – Text that exhibits excessive repetition, incoherence, or generic filler content, indicating generation collapse rather than semantic error (Holtzman et al., 2019).
- **Hallucination** – The translation introduces content that is unsupported by the source and lacks a meaningful semantic connection to it.

Exactly one primary label is assigned to each translation. If the primary label is *Hallucination*, it may be further categorized into one of the following subcategories:

- **Extraneous hallucination** – The translation fabricates semantically significant content that is unsupported by or not inferable from the source.
- **Repetition** – The translation contains repetitive phrases or tokens that form a looping or oscillatory pattern.
- **Detached hallucination** – The translation is fluent but largely unrelated to the source's meaning.
- **Partial hallucination** – A portion of the translation is unrelated to the source, or is only very indirectly related to it, while the remaining portion is meaningfully grounded in the source.
- **Incorrect target language/untranslated content** – The output is partly or entirely in an incorrect target language, or the source text is copied as-is rather than translated.
- **Other** – The hallucination does not fit any of the above categories. Annotators must provide a comment explaining the reason for assigning this label.

D.2 Machine Translation Annotation Guidelines

D.2.1 Distinguishing Mistranslation from Hallucination

The distinction between *mistranslation* and *hallucination* is based on whether the erroneous content

Type	Passage / Question / Answer	Label	Explanation
Unanswerable	**Passage (excerpt):** චිත්‍රපටියක් ඩීවීඩී ආකෘතියට පරිවර්තන කිරීමට... **Question:** ...ඕවර්ස්කෑන් නිසා උපසිරැසි කප්පාදු වීම වැළැක්වීම සඳහා භාවිත කළ හැකි මෘදුකාංගය...කුමක්ද? **Model answer:** [states a specific software name]	Unsupported	The passage never names a specific software or technical fix; any named tool is fabricated
Unanswerable	(same passage and question as above) **Model answer:** "The passage does not specify a particular software solution."	Correct refusal	Model correctly abstains given insufficient evidence
Answerable (short-answer)	[source passage] / [question] **Model answer:** [answer directly stated in passage]	Supported	Answer content matches passage directly

Table 5: Guideline examples illustrating the QA labeling framework.

remains meaningfully grounded in the source. Annotators were instructed first to identify the source content corresponding to the erroneous target content. If the erroneous content can be meaningfully associated with a source word, phrase, or proposition, but its meaning is incorrectly conveyed, it is categorized as *mistranslation*. In contrast, if the erroneous content introduces information or a proposition that cannot be meaningfully associated with any content in the source, it is categorized as *hallucination*. Thus, a change in meaning alone does not constitute hallucination. The key criterion is whether the erroneous content remains semantically grounded in the source.

Consider the following example:

> **Source:** He was injured.
>
> **Translation:** ඔහු මිය ගියේය.
>
> ("He died.")
>
> The translation conveys a different meaning from the source, but the erroneous content remains meaningfully connected to the source proposition, namely *he* + an adverse physical event. Therefore, this is categorized as *mistranslation* rather than hallucination.

In contrast:

> **Source:** He was injured.
>
> **Translation:** ඔහු ලබන සතියේ නව චිත්‍රපටයක රඟපානු ඇත.
>
> ("He will act in a new movie next week.")
>
> This is categorized as *hallucination* because the proposition introduced in the translation cannot be meaningfully associated with any content in the source and is therefore unsupported by it.

When a translation contains both source-grounded and unsupported content, annotators apply the same criterion at the content level: the source-grounded erroneous portion is treated as *mistranslation*, whereas the unsupported portion is treated as *hallucination*. Such cases may therefore be assigned the appropriate hallucination subtype when only part of the translation is unsupported.

D.2.2 Treatment of Transliterated English Words

English words that have an established Sinhala equivalent should not be considered *Mistranslation* solely because the English word is transliterated into Sinhala script. These are, in fact, legitimate *loanwords* (Jayatilleke et al., 2026). The appropriateness of a transliterated form depends on its usage and acceptability in Sinhala, including commonly used colloquial and code-mixed expressions. For example, although 'television" has the formal Sinhala equivalent රූපවාහිනිය, everyday Sinhala commonly uses the expression ටීවී එක. Therefore, such commonly accepted transliterations or code-mixed forms are not categorised as *Mistranslation* when they appropriately convey the intended meaning. A transliterated form is categorised as *Mistranslation* only when it is inappropriate or does not adequately convey the intended meaning in the given context, rather than merely because a formal Sinhala equivalent exists.

D.2.3 Inter-annotator Agreement

IAA was calculated at the output level rather than the span level. Three annotators (A4, A5, and A6) independently labelled the same 60 source sentences across 17 MT systems, producing 1,020 triple-annotated translations. The subset contained 15 source sentences from each of the four domains: News, Legal, Cultural, and General. The annotators were blinded to each other's labels.

For primary labels, we calculated raw pairwise agreement as the proportion of outputs receiving identical labels from two annotators. We averaged this agreement over the three annotator pairs. We also calculated pairwise Cohen's κ and Fleiss' κ to account for chance agreement. The latter was calculated across all three annotators using the five mutually exclusive primary labels.

For hallucination sub-labels, agreement was calculated over outputs identified as hallucinated by at least one annotator. We used the mean pairwise Jaccard index to compare assigned sub-label sets, together with exact set-match agreement and Krippendorff's α using MASI distance. These metrics measure agreement between categorical annotations rather than overlap between hallucinated text spans.

Primary-label agreement metric	**Value**
Mean pairwise raw agreement	72.39%
Fleiss' κ	0.5888
Pairwise Cohen's κ	
A4 – A5	0.6768
A4 – A6	0.5181
A5 – A6	0.5790

Table 6: Agreement for the primary MT annotation labels.

For the hallucination sub-labels, the mean pairwise Jaccard index was 0.3699, while exact set-match agreement was 10.49%. Krippendorff's α using MASI distance was 0.1019. Table 8 presents the agreement statistics for each hallucination sub-label individually.

E Model Families

Proprietary models (Gemini, GPT, Claude) were accessed via their official APIs, and open-weight models (Llama, DeepSeek, Qwen, Mistral) were run from official releases with greedy or low-temperature decoding ($T \leq 0.2$). Table 9 lists the LLMs evaluated grouped by family.

Domain	**Outputs**	**Agree. (%)**	**Fleiss' κ**
News	255	73.46	0.5790
Legal	255	76.73	0.6594
Cultural	255	75.56	0.6106
General	255	63.79	0.4881
Overall	1020	72.39	0.5888

Table 7: Output-level inter-annotator agreement for MT primary labels by domain. Each domain contains 15 source sentences evaluated across 17 systems. Raw agreement is averaged over the three annotator pairs.

Sub-label	**Agree. (%)**	κ
Extraneous hallucination	80.42%	-0.0427
Repetition	88.34%	0.0754
Detached hallucination	95.34%	0.3508
Incorrect target language / untranslated content	93.01%	0.7209
Partial hallucination	67.37%	-0.0007
Other	97.67%	0.1548

Table 8: Per-sub-label agreement for hallucination annotations, treating each sub-label as present or absent.

Table 9: Evaluated LLMs by family.

Family	Models
Gemini	Gemini-3.6-Flash, Gemini-3.1-Pro
GPT	GPT-5.5
Claude	Claude Sonnet 4.6, Claude Opus 4.6
Llama	Llama-3.1-70B-Instruct, Llama-4-Maverick, Llama-3.1-8B-Instruct
DeepSeek	DeepSeek-V4-Pro-0813, DeepSeek-V4-Flash-0731
Qwen	Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B
Mistral	Mistral-Large-2512

F Sentence-level Hallucinations

We evaluate sentence-level hallucination detection using automatic scores and human labels, treating Hallucination as positive and all other categories as negative.

As shown in Table 10, GEMBA-MQM performs best overall, followed by COMET. Higher scores when distinguishing Hallucination from Faithful alone indicate that mistranslations, omissions, and degeneration complicate hallucination detection.

G Qualitative MT Error Analysis

Table 11 compares the human-assigned error categories of 17 MT systems on selected source sentences.

All systems receive non-Faithful labels for items 2 and 241. For item 93, Google Translate succeeds despite failures by the three strongest systems. GPT-5.5 alone produces a Faithful transla-

Metric	ROC-AUC	AP	H vs. F
COMET	0.687	0.186	0.821
GEMBA-MQM	**0.770**	**0.233**	**0.941**
chrF	0.642	0.164	0.771
chrF3	0.634	0.160	0.763
Token F1	0.654	0.177	0.769

Table 10: Sentence-level hallucination detection. H vs. F denotes ROC-AUC for Hallucination versus Faithful only; other columns compare Hallucination against all non-hallucination labels. AP is average precision.

Model	2 News	241 Cultural	9 News	178 Disaster	120 Legal	93 Legal	61 News
Gemini-3.1-Pro	M	M	F	M	M	M	M
Gemini-3.6-Flash	M	M	F	M	M	M	M
DeepSeek-V4-Pro	M	M	F	H	M	M	M
GPT-5.5	M	M	M	M	**F**	M	**F**
Claude-Opus-4.6	M	M	M	**F**	M	**F**	**F**
Claude-Sonnet-4.6	H	M	M	H	M	**F**	M
DeepSeek-V4-Flash	M	M	F	H	M	M	M
Llama-4-Maverick	M	M	M	H	M	M	H
Llama-3.1-70B	O	M	**F**	H	D	D	M
Qwen3.5-9B	H	M	M	H	M	H	M
Qwen3.5-27B	M	M	M	H	M	M	M
Llama-3.1-8B	H	M	D	D	D	O	H
Qwen3.5-4B	H	M	M	H	D	M	D
Mistral-Large	M	M	H	O	O	M	M
Google Translate	M	M	F	H	O	**F**	M
NLLB-600M	M	H	M	H	O	H	M
MADLAD-400-3B	M	H	F	H	D	H	**F**

Table 11: Human-assigned primary labels for selected MT source sentences across 17 systems. F = Faithful, M = Mistranslation, H = Hallucination, O = Omission, and D = Degenerated Repetitive Output. Bold values highlight selected faithful translations. Source IDs and domains are shown in the headers.

tion for item 120, while the weaker systems exhibit different failure types on items 9 and 178.

H QA Results

Figure 4 compares token F1 with human-supported rates on answerable questions. The separation between the two measures shows that low reference overlap can coexist with source-grounded answers. Model rankings also differ: Gemini-3.1-Pro leads on token F1, whereas Claude-Opus-4.6 leads on Supported rate.

Figure 5 complements this comparison by separating successful responses from different failure modes. Supported answers and correct refusals represent appropriate behavior on answerable and unanswerable questions, respectively. Unsupported responses indicate hallucination, while false refusals and degenerated repetitive outputs capture separate failures. The larger Unsupported portions for the Llama-3.1 models show that their lower accuracy involves unsupported generation, rather than refusal to answer alone.

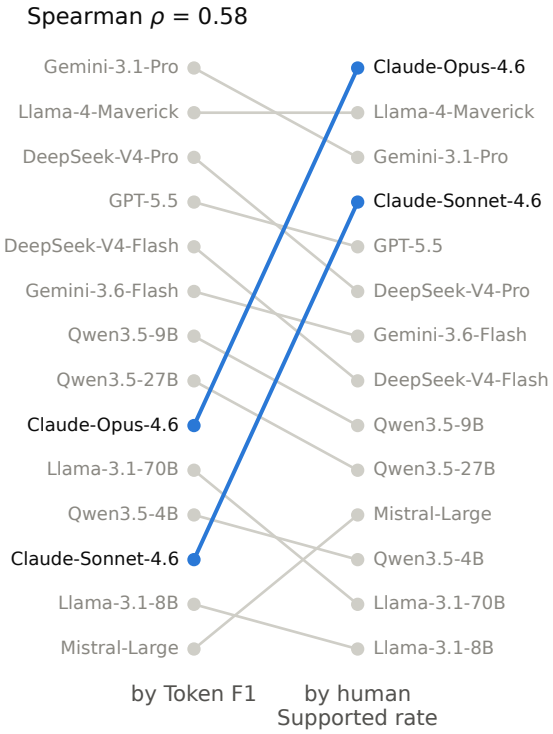


Figure 4: Comparison of token F1 and human-supported response rates on answerable Sinhala QA items across 14 models. Token F1 measures lexical overlap with the reference answer, while the Supported rate measures grounding in the source passage. Gemini-3.1-Pro achieves the highest token F1, whereas Claude-Opus-4.6 achieves the highest Supported rate, illustrating that the two evaluation methods can rank models differently.

I Machine Translation Results

This section provides analysis of human-annotated translation errors, hallucination subtypes, model size effects, topic-level variation, and automatic evaluation metrics.

Figure 6 reveals distinct failure patterns across systems. Gemini-3.1-Pro achieves the highest Faithful rate (92.5%), while Llama-3.1-8B exhibits extensive degeneration (56.9%). NLLB-600M shows substantial mistranslation and omission despite relatively low hallucination, demonstrating that low hallucination does not necessarily imply high translation quality.

Figure 7 shows that Partial Hallucination is the dominant subtype across all domains, particularly General (106) and News (81). Extraneous Hallucinations occur predominantly in News (53), while

Detached Hallucinations concentrate in General and Cultural (83). These patterns suggest that hallucination characteristics vary across domains.

Figure 8 shows improved Faithful rates with increasing model size in both families. Qwen3.5 improves from 1.0% (4B) to 25.5% (27B), while Llama-3.1 reduces degeneration from 56.9% (8B) to 15.0% (70B). However, hallucination remains substantial in larger models, indicating that increased capacity alone does not ensure source faithfulness.

Figure 9 reveals considerable topic-level variation within the General domain. Stronger models maintain high Faithful rates across most topics, whereas weaker models show inconsistent performance, including near-zero faithfulness on several topics. However, nine of twelve topics contain at most five sentences per system, so these differences should be interpreted as exploratory.

Figure 10 shows positive associations between automatic scores and human Faithful rates across 68 model–domain combinations. GEMBA-MQM exhibits stronger rank correlation ($\rho = 0.954$) than COMET ($\rho = 0.830$). However, similarly high automatic scores correspond to varying human Faithful rates, particularly for COMET, highlighting limitations in domain-level agreement.

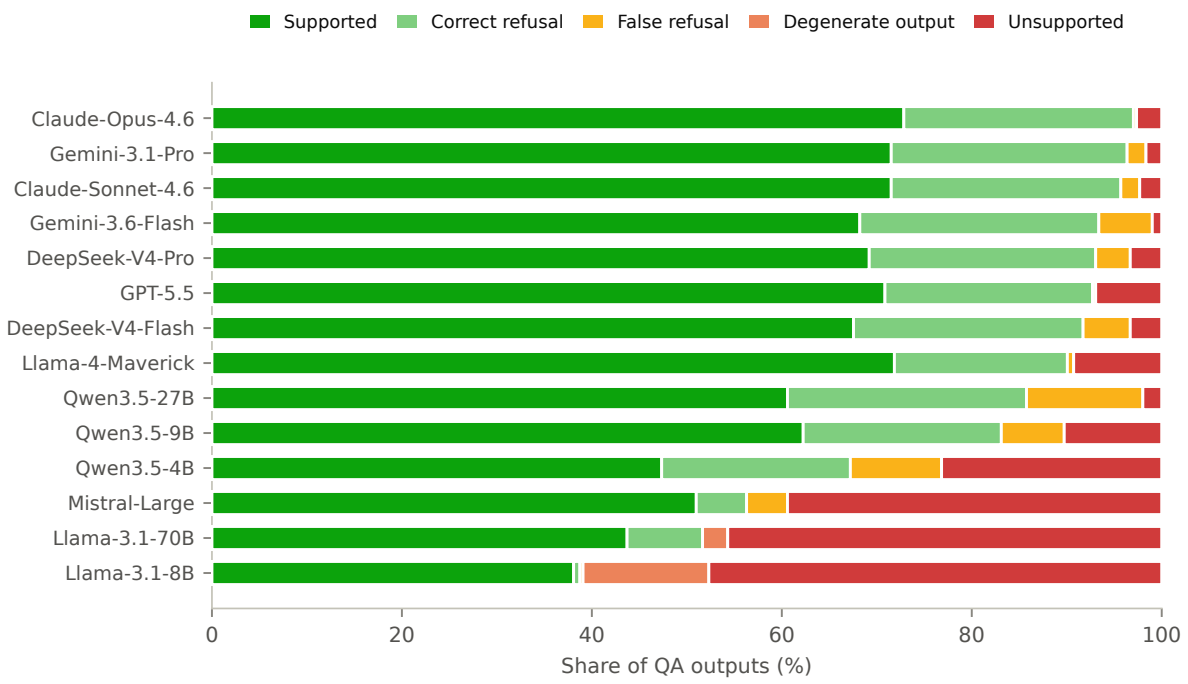


Figure 5: Distribution shows that high-performing closed-source systems generally produce more Supported answers and fewer Unsupported responses. However, the distribution also shows important differences among open-weight models. Llama-3.1-8B and Llama-3.1-70B have substantially larger Unsupported portions, showing weaker grounding in the provided Sinhala context.

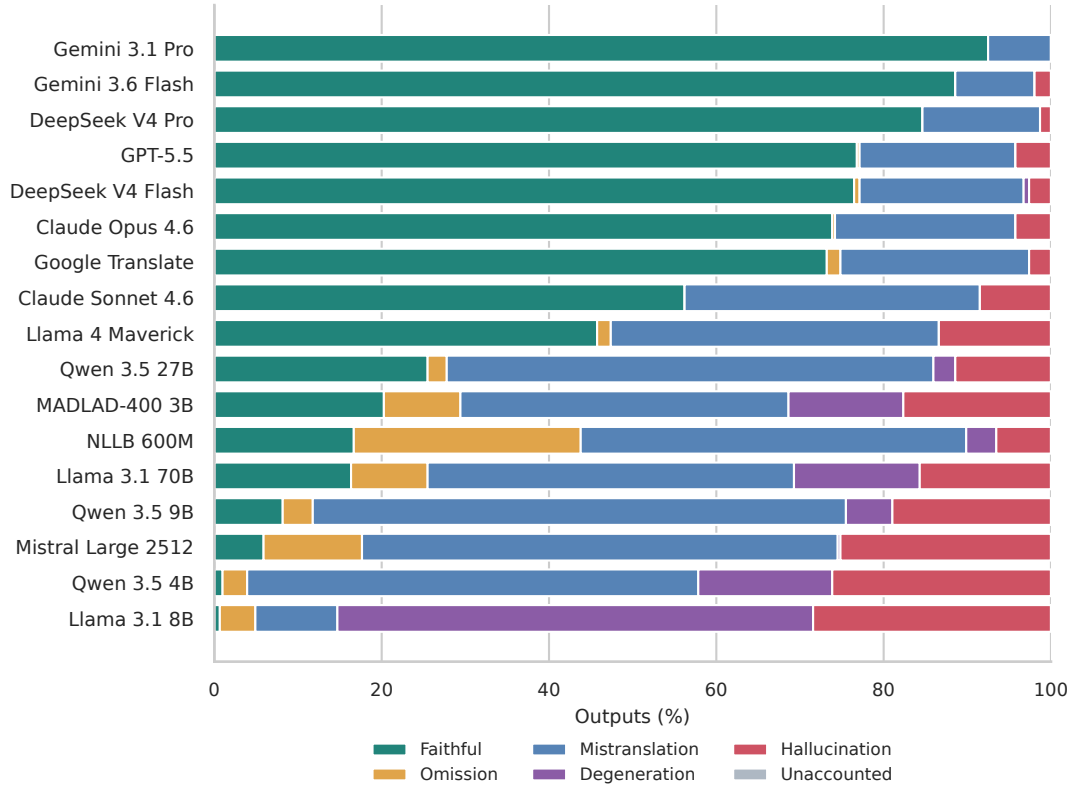


Figure 6: Distribution of the mutually exclusive primary human labels across the evaluated translation systems. Each bar represents 306 translations, and systems are ordered by their human-judged Faithful rate. The Unaccounted category is retained as an audit field and is zero in the revised primary-label data.

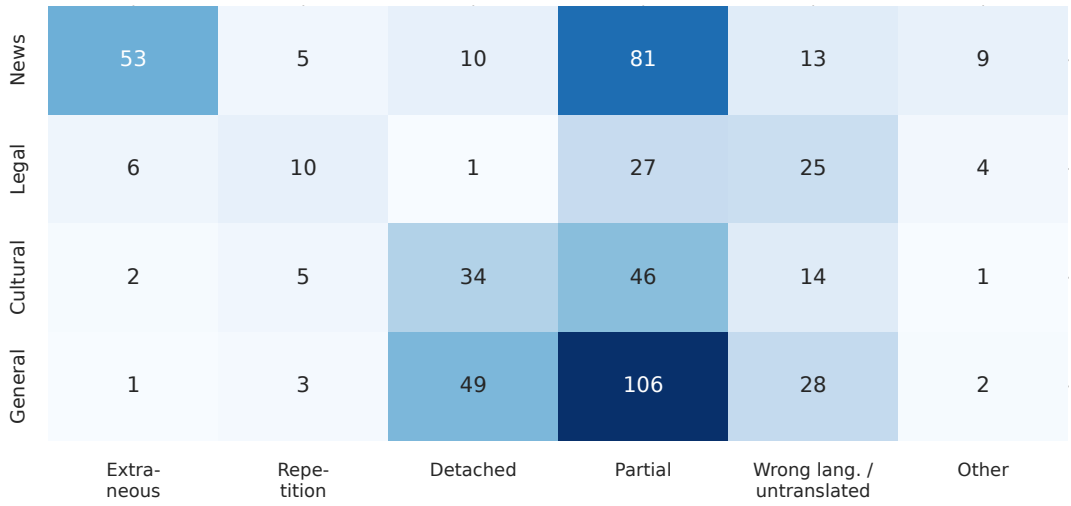


Figure 7: Recorded hallucination subtype assignments by broad domain. Partial Hallucination is the most frequently recorded subtype across all domains, whereas Extraneous Hallucinations are concentrated in News and Detached Hallucinations occur primarily in the Cultural and General domains. Incorrect target language or untranslated content is also relatively frequent in Legal. Cells report raw counts rather than normalized rates and therefore should not be interpreted as directly comparable domain-level prevalence.

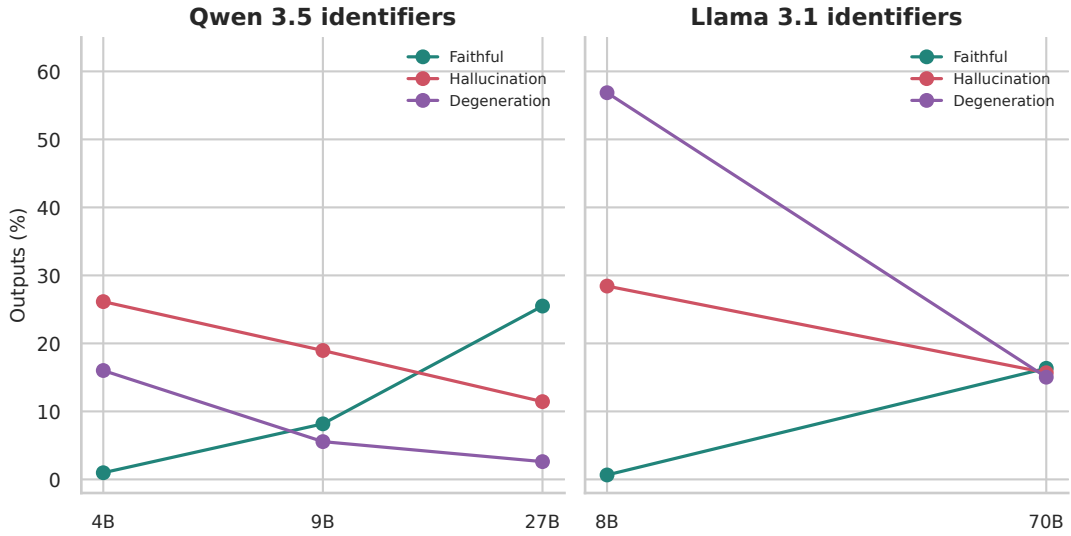


Figure 8: Faithful, Hallucination, and Degenerated Repetitive Output rates for the Qwen3.5 and Llama-3.1 parameter-labelled models. Larger parameter models within both families are associated with higher Faithful rates and lower degeneration rates.

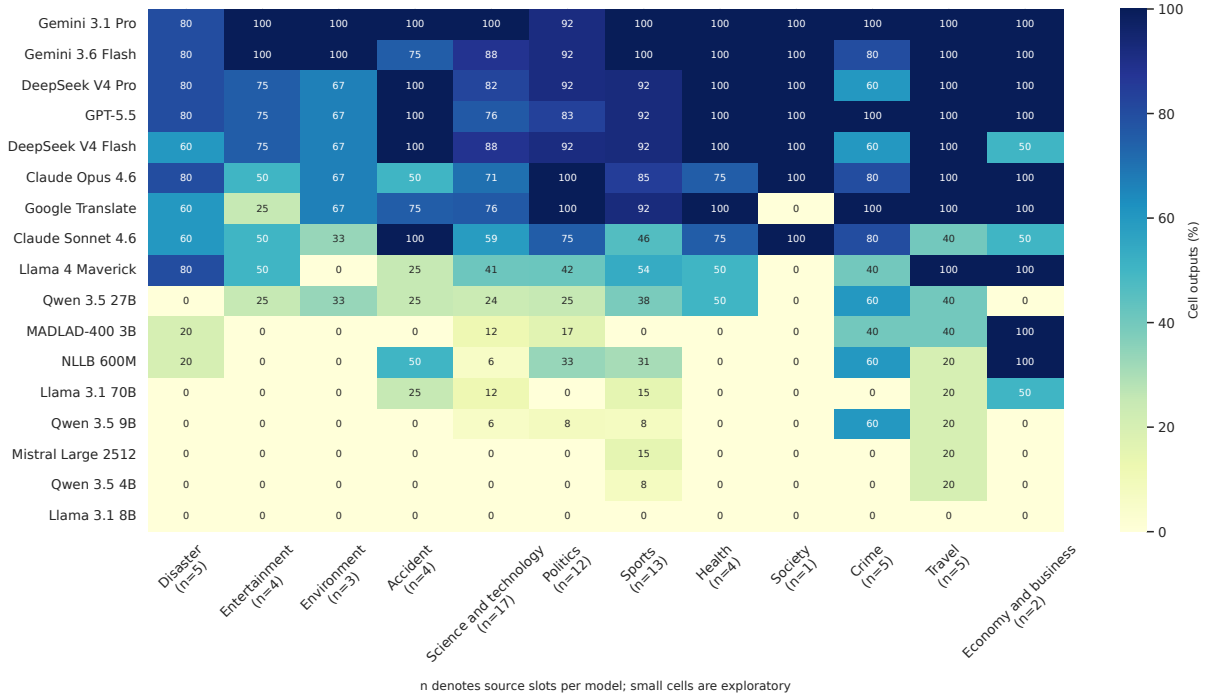


Figure 9: Human-judged faithfulness rates across the detailed topics within the General domain. The value of n under each topic gives the number of source sentences evaluated per system. Because nine of the twelve topics contain at most five source sentences, these results are exploratory and should not be interpreted as stable topic-level rankings.

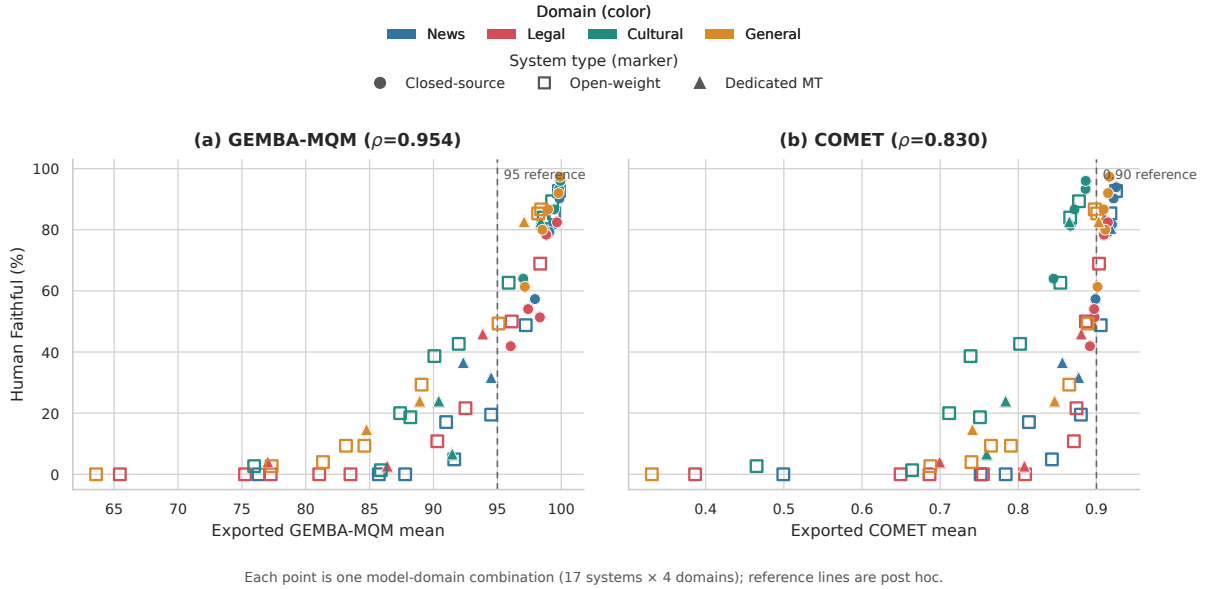


Figure 10: Relationship between human-judged Faithful rates and GEMBA-MQM and COMET scores across 68 model–domain combinations. Each point represents one system in one domain. Colors distinguish News, Legal, Cultural, and General domains, while circles, hollow squares, and triangles represent closed-source models, open-weight models, and dedicated MT systems, respectively. The reported ρ values are Spearman correlations. The dashed lines at GEMBA-MQM 95 and COMET 0.90 are descriptive reference points rather than calibrated quality thresholds.