

Findings of the WMT26 Multilingual Instruction Shared Task: Small Models Are Not Yet Polyglots

Patrícia Schmidtová

Charles University

Shobhit Banga

Voice Arena

Danni Liu

Karlsruhe Institute of Technology

Mariya Shmatova

Independent

Katia Artemova

Toloka AI

Manmeet Kaur

Voice Arena

Nam Luu

Charles University

Hanuman Sidh

Voice Arena

Maike Züfle

Karlsruhe Institute of Technology

Seth Aycock

University of Amsterdam

Tom Kocmi

Cohere

Sara Papi

Fondazione Bruno Kessler

Evfrosiniya Zermínova

Toloka AI

Pinzhen Chen

Queen's University Belfast

Niyati Bafna

Johns Hopkins University

Philipp Koehn

Johns Hopkins University

Beatrice Savoldi

Fondazione Bruno Kessler

Vilém Zouhar

ETH Zurich

Abstract

We present the findings of the WMT26 Multilingual Instruction Shared Task (MIST), evaluating small ($\leq 10\text{B}$ parameter) open-weight language models across 24 languages on three sub-tasks: context-based QA, cross-lingual summarization, and open-ended generation. Submissions from eight teams explore fine-tuning, knowledge distillation, and prompt engineering. Our findings demonstrate that small models are not yet polyglots: while leading systems perform well on high-resource languages, performance drops sharply on low-resource varieties, and comprehension degrades significantly when transferring across non-English language pairs. In open-ended generation, human evaluation and automatic rule-based verification exhibit strong overall rank correlation ($\rho=0.79$) but diverge locally at the top: human evaluators favor distilled pipelines with superior fluency and cross-lingual equity, whereas automatic verifiers reward concise prompt-engineered systems. Furthermore, multi-stage pipelines suffer from prompt-language inertia on cross-lingual instructions, and summarization exhibits severe English leakage unless penalized by language gating. Finally, while baseline fluency is largely achieved by modern small models, multi-constraint instruction following and unanswerable question refusal remain the primary performance bottlenecks. We release all test sets, system outputs, human judgments, and evaluation code under a permissive license.

1 Introduction

The Multilingual Instruction Shared Task at the Conference on Machine Translation (WMT) evaluates model capability in following instructions and providing useful responses across languages and

tasks. The first edition in 2025 turned into a competition joined by the largest open-source LLMs and frontier APIs (Kocmi et al., 2025). However, multilinguality should not remain a game reserved exclusively for giant or black-box models, because they are challenging and expensive to deploy and inaccessible for scientific analysis.

Therefore, in 2026,¹ our goal is to lower the entry barrier for a wider audience and challenge them to develop *small* multilingual models to punch above their weight.² Any participating systems should have at most 10 billion parameters, but participants are free to choose any data or methods for model building and inference. By doing so, we hope to attract a variety of techniques for improving multilinguality and to further our understanding of the progress in our community.

Secondly, model competence across the world's languages remains remarkably uneven, so we include a diverse set of languages in terms of writing script, location, and availability of resources. In terms of the shared task outcome, we share per-language and per-task analysis instead of emphasizing a single aggregated ranking.

The shared task introduced a fresh benchmark that systematically assesses key capabilities of LLMs across 24 languages using 3 sub-tasks:

- **Context-Based Question Answering (Ctx-QA):** Answering a question solely based on the content of a paragraph, including appropriately declining to answer when the information is absent.
- **Open-Ended Generation (OEG):** Responding to open-ended questions, some with con-

¹Shared task information and resources are available on www2.statmt.org/wmt26/multilingual-instruction.html.

²Pun intended.

straints and some localized, where we assess language proficiency, knowledge, and constraint satisfaction.

- **Cross-lingual Summarization (Sum):** Summarizing short ACL papers in different languages, which tests a practical cross-lingual ability.

Before releasing the test set, we shared an optional sample data mix sourced from existing benchmarks to help participants develop their models. After the test release, they have one month to submit their model outputs. In addition to participants’ submissions, we benchmark small open-weight models within 10B parameters as baselines.

The evaluation for QA and Sum is reference-based, where we use a combination of metrics targeting different aspects of the model output. OEG is evaluated by both LLM-as-a-verifier and human annotators.

2 Data and Tasks

This section describes the datasets and preparation for each of the three sub-tasks in our benchmark. For every sub-task, we curated or adapted data for up to 24 languages. The high-level statistics are summarized in Table 1, and example prompts are shown in Table 2. Each prompt is self-contained: it includes all necessary context, instructions, and the question, so participants’ systems receive a single string as input. The shared task consists of two phases: we released a sample data mix with questions and language coverage similar to the final test, by gathering resources from existing benchmarks (Clark et al., 2020; Ladhak et al., 2020; Bhattacharjee et al., 2023; Bandarkar et al., 2024; Singh et al., 2024; Papi et al., 2026). The final testing phase started one month before the submission deadline.

The following sections detail the sources of the data, the translation or localization processes applied, and any additional filtering or validation steps specific to each sub-task.

Prompts for all tasks, or their samples, are presented in Appendix A.

2.1 Test Languages

The official test data covers 24 languages spanning 7 scripts and 12 language families: Arabic, Bengali, Bhojpur, Kurdish, Chinese (Simplified), Czech, English, Finnish, German, Haitian Creole, Hindi, Indonesian, Italian, Japanese, Korean, Marathi, Persian (Farsi), Portuguese, Russian, Slovak, Span-

	Langs	Samples / lang
Context-based QA	23	360 (100 unique)
Open-Ended Generation	24	100*
Cross-lingual Summarization	23	74

Table 1: Number of languages and the number of samples (prompts) for each language. *Yoruba included only 59 post-edited samples.

Context-based QA	<i>Read the following passage and answer the question based solely on its content. [...] “Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world [...]”</i> Question: What is proclaimed as the highest aspiration of the common people?
Open-Ended Generation	As a news reporter, write an article about the opening of a new shopping complex, including who will enjoy it and what activities are available.
Cross-lingual Summarization	Bitte fassen Sie die folgende wissenschaftliche Arbeit auf Deutsch zusammen: “We present a method for [...] Our experiments on three benchmarks show [...]”

Table 2: Example prompts for each sub-task.

ish, Turkish, Vietnamese, and Yoruba. Yoruba is present only for the OEG task, and Bhojpur is introduced as a surprise language not present in the sample data we released. The surprise language tests the zero-shot or few-shot transfer ability of participating systems to an unseen, low-resource language.

2.2 Context-based Question Answering

The context-based QA sub-task evaluates a model’s ability to answer questions based solely on a provided passage, where the question and the expected answer are in one language and the passage may be in a different language.

Sources and domains We curated 20 English passages from five domains, selected for topical and stylistic diversity:

- **Universal Declaration of Human Rights (UDHR):** Preamble and selected articles (Articles 1–2, 18, 26, 27). The world’s most translated document available in 500+ languages.
- **UNESCO Convention:** Preamble and articles from the Convention for the Safeguarding of the Intangible Cultural Heritage. Available in the five UN languages: arb, zho, eng, fra, rus,

spa.

- **UN Security Council Resolution 2347:** Preamble and operative paragraphs on cultural heritage protection. Also available in the 5 UN languages.
- **EU Law:** Excerpts from the Charter of Fundamental Rights (Articles 21–22), the AI Act, and the Treaty on the Functioning of the EU. Available in all EU languages.
- **Wikipedia:** Articles on culturally specific topics: Mid-Autumn Festival (zho), Carnival (por), Wabicha (Japanese tea ceremony) (jpn), Diwali (hin, mar, bho) and the Yoruba people (yor).

These texts were chosen because of their availability in at least two languages simultaneously. Rather than translating the passages, we rely on the existing human translations.

For each passage, we manually wrote five questions in English, including 4 answerable and 1 unanswerable question. Unanswerable questions test whether the model can correctly decline when the passage does not contain the answer. Passages where the answers cannot be found are accompanied by a language-specific refusal phrase that the model is instructed to produce, e.g., “not answerable” in English, “bez odpovědi” in Czech, or “无法回答” in Chinese.

Cross-lingual setup The questions, answers and prompt templates were translated into all target languages using GPT-5.5 (Qs and prompts) and GPT-5.6-Sol (answers). Each test instance pairs a passage in a source language with a question in a (potentially different) target language. Four cultural Wikipedia passages are paired with their native language plus English, while the remaining 16 passages each appear in English and three additional source languages, selected by a coverage-balancing algorithm. Since the same question can be paired with a passage in multiple source languages, this yields 360 instances per target language (100 unique questions), creating a mix of monolingual and cross-lingual reading comprehension items that test whether small models can synthesize information across language boundaries.

2.3 Open-Ended Generation

Following last year, we continue the open-ended generation (OEG) sub-task, which evaluates a model’s ability to produce fluent, useful, and locally appropriate responses to open-ended ques-

tions, while following user instructions.

Prompt design We manually created 100 seed prompts in English, spanning three categories:

- **General:** Creative and reflective writing tasks (e.g., “Draft a 150-word description of a technological advancement that excites or concerns you”).
- **Factual:** Knowledge and explanation tasks, sometimes requiring analogical reasoning (e.g., “Explain quantum entanglement using only cooking metaphors”).
- **Language/location-specific:** Prompts that have answers that may vary for various languages, cultures, or locations. These are written with {language} or {country} as template placeholders (e.g., “How are diminutives typically formed in {language}?”).

Many prompts include explicit constraints such as word or sentence limits, required keywords, specific output formats, or stylistic requirements.

Translation and localization. For language- and location-specific prompts, the {language} and {country} placeholders in the English seed are instantiated before translation. The prompts were translated into target languages using GPT-5.5 via the standardized translation prompt template documented in Appendix A.2. The resulting questions were then reviewed and post-edited by proficient speakers of each language to ensure naturalness and cultural appropriateness.

Human Post-Editing Effort To quantify the extent of human correction, we measure the post-editing distance between the initial GPT-5.5 translations and the finalized test prompts across all 23 target languages (Table 12 in Appendix A.2). Overall, the human editing effort was modest: across all target languages, an average of 66.1% of prompts were accepted verbatim without any edits, with an average of 5.5 character edits per prompt. Languages varied substantially in editing difficulty and intensity: high-resource languages such as Arabic, Hindi, German, Indonesian, and Portuguese were easy, requiring negligible editing ($\geq 88\%$ accepted verbatim, chrF ≥ 99.0). In contrast, languages like Czech and Slovak required frequent but low-intensity edits ($\approx 60\%$ of prompts edited, but with a high chrF ≥ 95.2 , reflecting minor morphological, diacritic, or stylistic adjustments). Conversely, Haitian Creole, Kurdish, and Persian (Farsi) required substantial, high-intensity rewriting: Haitian

Creole had only 9.3% unedited prompts, while Persian exhibited 50.0% edited prompts with a 14.1% character error rate. We report the full post-edit distance for each language in Table 12.

2.4 Cross-lingual Summarization

The cross-lingual summarization sub-task requires the model to read a full scientific paper in English and produce a summary in a specified target language.

Source papers We downloaded all 74 short papers from the 64th Annual Meeting of the ACL (ACL 2026). We extracted the full paper text and selected text from the beginning of the introduction section up to the text before the reference section. This paper content text was provided as an input for the task.

Reference summaries Gold-standard summaries are the authors’ own abstracts, translated from English into the target languages by two systems (GPT-5.6-Sol and Gemini-3.1-pro-preview). Both translations are retained as equally valid references for evaluation; pairwise chrF between the two translation systems ranges from 52.3 (Yoruba) to 81.0 (Portuguese), with 56.7 in Chinese and 63.6 in Japanese, confirming that the choice of a single reference would visibly affect system rankings.

Prompt construction For each (paper, target language) pair, one of five paraphrased prompt templates is selected at random (seeded for reproducibility). Each template instructs the model to summarize the given paper into the target language, varying in phrasing to reduce prompt-specific artifacts. The prompts were translated into each language using GPT-5.5 and post-edited by proficient speakers of each language.

3 Participating Systems

We received submissions from eight teams, summarized in Table 3. All systems meet the shared task’s constraint of 10B total parameters in the final system. The dominant approach is parameter-efficient fine-tuning (LoRA or QLoRA) of a pre-trained multilingual model, though teams differ substantially in their choice of base model, training data, and inference strategy, which is what we hoped for.

TUM-HEILBRONN Hou et al. (2026) fine-tune a 4-bit quantized Qwen3.5-9B with LoRA ($r=16$), first on the QA training split for 2 epochs, then on

summarization for up to 3 additional epochs. Loss is computed only on assistant responses. They report zero-shot performance on the test set.

CUNI Papáček and Schmidtova (2026) evaluate several open-weight models with prompt engineering (system prompt design) and hyperparameter search (temperature, top- p , top- k , reasoning on/off). They select Qwen3.5-9B as their primary system for its reliability in following explicit constraints, with Gemma 4 E4B submitted as a contrastive variant.

PSK Pulipaka (2026) trains task-specific QLoRA adapters on top of Tiny Aya Global, using a mixture of the provided training data and external datasets including WikiLingua, CrossSum, Belebele, TyDiQA, and MLQA. At inference time, instances are routed to the appropriate adapter based on their task label.

ALKHDER Alkhder et al. (2026) fine-tune Qwen2.5-3B-Instruct with LoRA ($r=32$, $\alpha=64$) on 3,112 monolingual Arabic extractive QA examples from TyDiQA. They participate only in context-based QA for Arabic, using BF16 precision and greedy decoding.

ALEIX Sant et al. (2026) adapt Qwen3-8B with a two-stage LoRA procedure on a mixture of WMT MIST data, EuroAlpaca, and a new CrossEuroAlpaca dataset. EuroAlpaca provides 51,760 AlpacaCleaned-derived samples per language, localized in a task-preserving way. The system is primarily optimized for OEG in European languages.

SENTINEL Egerton-Idehen (2026) combines a fine-tuned and an untuned copy of Gemma-3-4b-it. Context QA is routed to the fine-tuned model, while OEG and summarization go to the base model. All outputs pass through a verification wrapper that checks language, length, and well-formedness, retrying or repairing generation when a check fails.

SPINOZA Krause et al. (2026) train LoRA adapters on Qwen3.5-9B using 100,466 cross-lingual QA examples machine-translated from TyDi QA. They additionally train an answerability probe on the same base model: if the probe classifies a question as unanswerable, the LoRA weights are bypassed and a fixed refusal phrase is returned. This estimates the empirical upper limit of the approach by separating answering ability from instruction following.

Team	Base model	Method	Key steps
TUM-Heilbronn	Qwen3.5-9B (4-bit)	LoRA ($r=16$)	Sequential fine-tuning: QA then summarization; zero-shot evaluation
CUNI	Qwen3.5-9B	Prompt engineering	System prompt and hyperparameter search; Gemma 4 E4B as variant
PSK	Tiny Aya Global	Task-specific QLoRA	Separate adapters for summarization, context QA, and open QA; external data
Alkhder	Qwen2.5-3B-Instruct	LoRA ($r=32$)	Arabic extractive QA only; character-offset loss masking
Aleix	Qwen3-8B	Two-stage LoRA	EuroAlpaca + CrossEuroAlpaca data; European language focus
Sentinel	Gemma-3-4b-it	Full fine-tune + base	Routing: fine-tuned for QA, base for OEG/sum; verification wrapper
Spinoza	Qwen3.5-9B	LoRA + probe	Answerability probe on the base model; fine-tuning on translated QA data
Lanigo	Gemma-4-E4B-it	Full fine-tuning	Knowledge distillation from Gemma-4-31B teacher; pivot decoding for OEG/sum

Table 3: Overview of participating systems. All systems are under the 10B parameter constraint.

LANIQO Charkiewicz and Nowakowski (2026) perform full-parameter fine-tuning of Gemma-4-E4B-it (8B total, 4.5B effective) on 20,625 distilled examples from a Gemma-4-31B teacher. They use task-conditional decoding: a pivot strategy (English intermediate draft, then target-language rewrite) for OEG and summarization, and direct single-pass decoding for context QA. A contrastive variant trains on the shared task’s own reference outputs instead of distilled data.

Reporting system variants Some participant system variants differed only in a specific sub-task. To be concise, we report the variants of those team as one line in result tables if they are the same for a given sub-task.

Organizer Baselines In addition to the participant submissions, we benchmark models within the 10B constraint:

- **Tiny Aya Global** (Salamanca et al., 2026)
- **Gemma-4-E4B-it** (Team, 2026)
- **GLM-4-9B-0414** (GLM et al., 2024)
- **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024)
- **Phi-4-mini-instruct** (Microsoft et al., 2025)
- **Qwen3-8B** (Alibaba et al., 2025)

These were the base models our participants built upon. They contextualize participant results against off-the-shelf models in this “weight category”. Since baselines were evaluated with a single verifier (Gemma-4-31B) rather than the full multi-judge ensemble, their per-task and cross-task scores are reported separately in Appendix E.

4 Human Evaluation

Human evaluation was conducted for the open-ended generation (OEG) sub-task, where automatic metrics and lexical references are least reliable due to the creative and open-ended nature of the generated texts. System outputs were evaluated along three core quality dimensions on a 7-point calibrated scale:

- **Instruction Following (IF)** (1–7): How faithfully the response satisfies all explicit constraints and instructions in the prompt.
- **Naturalness (Nat)** (1–7): Fluency, idiomatic phrasing, and appropriateness of tone and formality in the target language.
- **Coherence (Coh)** (1–7): Logical structure, flow of arguments, and internal consistency.

We conducted human evaluation on the primary system as indicated by each participant. It covers only 14 languages due to resource constraints and annotator availability, with 69 samples each, as certain questions involve content in a language different from the question’s language.

4.1 Evaluation Pipeline and Crowdsourcing

Human ratings were contributed by two of the task’s sponsors:

- **Toloka** (platform.toloka.ai) provided support for 10 languages: Arabic, Chinese, German, English, Indonesian, Italian, Japanese, Korean, Russian, and Spanish. The evaluation was conducted by MIST Organizers through Toloka’s specialized platform, using their structured expert pipeline with built-in quality assurance.

- **VoiceArena** (voicearena.com) collected human evaluation annotations for 4 Indic languages (Bengali, Bhojpuri, Hindi, Marathi) on their own in-house platform.

Both campaigns used the same guidelines and the same 7-point scale. We provide a detailed description of the setup in Appendix C.

4.2 Results

The human evaluation covers 69 items per language, for a total of 966 items for 14 languages. Table 4 reports the overall results and official WMT significance cluster ranking; Table 5 provides a per-language breakdown.

Significance Clusters and Rankings. Applying paired bootstrap resampling ($n=10,000$, $\alpha=0.05$) and Wilcoxon signed-rank tests over the 966 evaluated items partitions the seven evaluated submissions into five statistically distinct performance clusters (Table 4): the top three (LANIQO, SENTINEL, CUNI; ranks 1–3), followed by PSK (rank 4), TUM-HEILBRONN (rank 5), ALEIX (rank 6), and TEAMSPINOZA (rank 7), each significantly separated from the others.

Dimensional Trade-offs: Instruction Following as the Bottleneck. Within the top cluster, Naturalness (5.51–5.74) and Coherence (5.59–5.68) are tightly grouped, whereas Instruction Following shows the widest spread across all systems (1.67 points). This demonstrates that baseline target-language fluency is largely solved in modern $\leq 10B$ parameter models, whereas faithful compliance with complex, multi-constraint instructions remains the primary differentiator. Notably, CUNI achieves the highest Instruction Following score (5.36), benefiting from targeted system prompting, while LANIQO leads on Naturalness (5.74) and Coherence (5.68) through knowledge distillation and pivot decoding.

Language-Specific Patterns. As detailed in Table 5, performance varies markedly across language families:

- **High-Resource European Languages:** English outputs receive the highest absolute ratings across all systems (all dimension means ≥ 5.80 ; CUNI leading at 6.21, followed closely by TEAMSPINOZA at 6.19 and LANIQO at 6.14). German and Spanish likewise display strong adherence, with SEN-

TINEL (5.75) and LANIQO (6.04) leading, respectively.

- **Asian Languages and Script Divergence:** In Japanese, top systems remain closely grouped (LANIQO at 5.94, CUNI at 5.91, SENTINEL at 5.91). In Korean, however, sharp performance disparities emerge: Instruction Following drops sharply for several systems, while CUNI exhibits a noticeable dip in Korean Naturalness, pulling its overall Korean score down to 4.61. LANIQO maintains the most balanced Korean output (5.30 overall). In Chinese, CUNI leads at 6.17, followed by TUM-HEILBRONN at 5.93 and ALEIX at 5.92.
- **Arabic and Indonesian:** In Arabic, LANIQO (5.52) and CUNI (5.48) lead with balanced instruction adherence. In Indonesian, SENTINEL (5.44), LANIQO (5.40), PSK (5.37), and CUNI (5.36) form a tight competitive tier, whereas TEAMSPINOZA drops to 4.54 due to low instruction following (3.66).
- **Indic Languages:** Performance drops for lower-resource Indic languages, with the surprise Bhojpuri showing the steepest decline (top system LANIQO at 4.82, bottom systems below 2.50). Bengali and Marathi show a clearer separation between the top cluster and the rest: in Bengali, SENTINEL leads at 5.57, while in Marathi, CUNI (5.75) and SENTINEL (5.75) are tied. Hindi follows a similar pattern, with CUNI leading at 5.43. Across all four Indic languages, the gap between the top three systems and the remaining submissions is substantially wider than for high-resource languages, and several lower-ranked systems fall below 3.0, indicating severe quality degradation.

5 Automatic Evaluation

Human evaluation was not conducted for all submissions, tasks, or languages, but we ran a thorough automatic evaluation for submitted outputs.

5.1 Metrics

We employ a comprehensive suite of automatic metrics, each targeting a different aspect of output quality:

- **Language identification (LangID):** We verify that the output is in the expected target language using GlotLID (Kargaran et al., 2023), evaluated across the 24 shared task lan-

System	Rank	Overall↓	IF	Nat	Coh	Winshare
Lanigo	1 [1–3]	5.57 ± 0.04	5.28	5.74	5.68	60.3%
Sentinel	2 [1–3]	5.51 ± 0.04	5.21	5.69	5.62	58.6%
CUNI	3 [1–3]	5.49 ± 0.05	5.36	5.51	5.59	58.3%
PSK	4 [4–4]	4.93 ± 0.05	4.51	5.24	5.05	45.9%
TUM-Heilbronn	5 [5–6]	4.80 ± 0.06	4.33	5.11	4.96	45.3%
Aleix	6 [5–6]	4.70 ± 0.06	4.25	5.00	4.85	43.6%
TeamSpinoza	7 [7–7]	4.38 ± 0.07	3.69	4.89	4.54	38.0%

Table 4: OEG human evaluation cluster ranking ($N=69 \times 14$ languages). Ranks in brackets indicate WMT significance clusters ($p < 0.05$, paired bootstrap, $n=10,000$).

System	arb	deu	eng	ind	ita	jpn	kor	rus	spa	zho	ben	bho	hin	mar
Lanigo	5.51	5.72	6.13	5.37	5.56	5.93	5.28	5.76	6.04	5.89	5.30	4.93	4.99	5.54
Sentinel	5.31	5.77	6.09	5.43	5.42	5.89	4.89	5.86	5.72	5.81	5.55	4.55	5.09	5.72
CUNI	5.48	5.58	6.21	5.33	5.48	5.90	4.61	5.75	6.05	6.17	5.02	4.19	5.39	5.65
PSK	5.06	5.27	5.80	5.35	4.98	5.77	4.78	5.34	5.74	5.46	4.44	2.92	3.87	4.31
TUM-Heilbronn	4.91	5.42	5.93	4.92	5.26	5.67	4.92	5.16	5.84	5.93	3.77	2.88	3.56	3.02
Aleix	4.99	5.49	6.01	4.95	5.22	5.81	4.83	5.29	5.95	5.92	2.88	2.53	3.35	2.61
TeamSpinoza	4.92	5.14	6.18	4.55	4.85	5.66	4.93	4.75	5.05	5.84	2.80	2.29	2.05	2.27

Table 5: OEG human evaluation breakdown: per-language scores (dimension mean, $N=69$ per language).

guages. GlotLID provides native support for low-resource languages including Bhojpuri, Kurdish, and Yoruba without requiring approximate surrogate matching.

- **Non-answerable validation (NonAns):** For context QA, we check whether the model correctly produces the instructed refusal phrase for unanswerable questions. Matching is rule-based (NFKC normalization, casefolding, punctuation stripping) against the exact phrase from the prompt template.
- **Lexical metrics:** BLEU, chrF++ (**chrF**), and ROUGE (**R-L**, **R-1**, **R-2**) computed against gold references using a shared per-language tokenizer (jieba for Chinese, MeCab for Japanese/Korean, whitespace elsewhere). chrF++ runs on raw text for tokenization independence.
- **COMET and CometKiwi (CQE):** Reference-based and reference-free neural MT metrics. The (source, hypothesis, reference) triple uses the English gold as source and the target-language gold as reference.
- **NLI:** Natural language inference scores using ModernBERT-base-nli in three settings: QA reference entailment, summarization mutual entailment with reference (**NLI-ME**), and SummaC-style source faithfulness (**NLI-F**).
- **LLM-as-a-verifier (QA):** For QA, we use the judge to verify semantic equivalence between the system response and the reference answer.

For non-answerable questions, the judge decides whether the system correctly pointed it out.

- **LLM-as-a-verifier (OEG):** We manually write several verification rules for each question to enable automatic verification of instruction following and content quality, which can be more reliable than an LLM-as-a-judge using fixed one-for-all rubrics for each question. In particular, the rules check output length against the specified constraint; the presence of required words; and whether key concepts in the question are addressed. The rules are always in English regardless of a question’s language. We run LLMs to produce a binary outcome for each rule.
- **LLM-as-a-judge (Sum):** For summarization, we use the judge in a reference-free fashion and only ask for a fluency rating of the system output.
- **Language-Gated metrics:** For cross-lingual summarization, generated outputs that fail LangID receive a score of zero for that instance. Language-Gated ROUGE-L (**G-R-L**) and Language-Gated chrF++ (**G-chrF**) directly penalize systems that generate in English rather than the specified target language.

For tasks with gold references, we use multi-reference evaluation: the target-language gold was produced by two independent translation systems, and both are retained as equally valid references

(Section 2.4).

For rule-based and answer verification across tasks, official scoring employs an ensemble of LLM verifiers: Cohere Command A+ (**Cmd-A+** in tables), **GLM-5.1**, and Gemma-4-31B (**Gemma4**), with **Qwen-3.5-397B** also evaluated on open-ended generation rules. A majority vote across verifiers (**Maj**) serves as the primary metric.

5.2 Context-based Question Answering

Table 6 reports evaluation results for context-based question answering aggregated across all 24 languages. Systems are ranked by their 3-judge majority-vote verification accuracy (**Maj**), which serves as the primary benchmark metric.

Verification vs. Lexical Overlap. A striking finding is the decoupling between rule-based verification accuracy and surface lexical overlap. CUNI-PRIMARY achieves the highest majority verification accuracy (81.63%) and leads on individual verifiers (79.13% Command A+, 77.71% GLM-5.1, 83.47% Gemma), closely followed by CUNI-VARIANT_B (80.93% Maj, 38.2 chrF++) and LANIQO-PRIMARY (76.90% Maj). In contrast, TUM-HEILBRONN-PRIMARY achieves the highest BLEU (19.7), ROUGE-L (34.95), and ModernBERT NLI mutual entailment (27.15), but attains only 59.16% on majority verification. This divergence stems from distinct generation strategies: prompt-engineered models (CUNI) generate concise, targeted answers that directly satisfy verifiers, whereas sequential fine-tuning (TUM-HEILBRONN) produces longer spans with substantial lexical overlap with gold answers, even when partially missing strict answer precision.

Answerable Comprehension vs. Refusal Accuracy. Table 7 reports performance dissected across the 6,912 answerable reading comprehension questions and the 1,728 unanswerable questions (where models are instructed to output a fixed refusal phrase) evaluated under the Gemma-4-31B-it judge across all 24 languages. Top participant systems (CUNI, LANIQO) balance strong comprehension (≈ 74 –80%) with strong refusal discipline ($\geq 97\%$).

Non-Answerable Detection by Participant Probes. The ability to correctly decline unanswerable questions varies substantially among participant submissions. TUM-HEILBRONN-PRIMARY (87.07% F1) and TEAMSPINOZA sys-

tems (86.40%–86.50% F1) excel on this dimension, demonstrating the efficacy of Spinoza’s specialized answerability classification probe. Conversely, several models (e.g., SENTINEL-BASE at 55.91% and ALEIX-PRIMARY at 41.51%) struggle to reliably emit the exact instructed refusal phrase across diverse target scripts.

Cross-Lingual and Low-Resource Disparities.

Performance breaks down starkly along language resource tiers (Tables 13 and 14). While top systems achieve 80–87% verification accuracy on high-resource European languages (English, German, Italian, Spanish), severe drops occur on low-resource and distant languages. Yoruba proves the most difficult (best Gemma-4-31B score of 41.7% by CUNI-PRIMARY, with many systems falling below 20%), followed by Kurdish (top score 64.7%). The surprise language, Bhojpuri, shows high accuracy for top systems (98.3% for CUNI-PRIMARY), demonstrating strong zero-shot cross-lingual transfer from Hindi data, but weaker systems exhibit steep degradation.

5.3 Open-Ended Generation

Based on LLM-as-a-verifier results, both systems from CUNI perform strongly and top the results, followed by Laniquo and Sentinel. Across the 24 target languages (Table 9), the two CUNI systems collectively capture the top score in every language: CUNI-VARIANT_B leads in 12 languages (excelling on non-Latin and mid-to-low-resource scripts such as Arabic at 92.4%, Bengali at 93.3%, and Kurdish at 78.7%), while CUNI-PRIMARY leads the other 12 (predominantly high-resource European and East Asian languages). Performance varies markedly across language resource tiers: Bhojpuri demonstrates strong zero-shot transfer from Hindi data, whereas Yoruba represents a universal bottleneck across all participant architectures (peaking at only 43.4% rules verified and falling below 3% for weaker submissions), alongside steep drops in Kurdish and Haitian Creole for smaller open models.

Language correctness exhibits a distinct pattern compared to rule-based LLM verification results, where the best results are obtained by LANIQO and SENTINEL, whereas CUNI underperforms several other submissions.

Finally, we compute ranking correlations between the earlier human evaluation and automatic LLM-as-a-verifier for primary systems in the OEG

System	Maj↓	Cmd-A+	GLM-5.1	Gemma4	BLEU	chrF	R-L	COMET	CQE	NLI	LangID	NonAns
CUNI-primary	81.63	79.13	77.71	83.47	8.7	35.1	24.14	68.05	61.30	15.87	97.4	82.02
CUNI-variant_B	80.93	78.12	76.75	82.89	10.6	38.2	26.96	69.34	62.54	17.52	98.5	81.79
Lanigo-primary	76.90	75.36	72.11	78.78	9.5	36.1	24.88	68.60	62.02	15.94	98.5	80.41
Lanigo-variant_A	69.80	69.19	65.43	72.29	10.4	34.9	25.14	67.63	60.38	16.26	97.5	76.72
TeamSpinoza-primary	67.50	68.85	63.44	69.64	12.6	36.3	28.33	71.21	62.74	24.36	95.9	86.40
TeamSpinoza-Variant_V4	61.96	63.91	56.88	66.10	12.9	34.8	27.70	72.62	63.28	26.07	95.4	86.50
Sentinel-base	59.58	61.53	54.88	61.77	5.7	25.8	16.73	62.70	55.58	11.12	98.0	55.91
TUM-Heilbronn-primary	59.16	61.59	54.48	64.05	19.7	34.0	34.95	72.17	61.83	27.15	92.2	87.07
Aleix-variant_B	57.08	60.44	52.89	59.62	5.4	24.3	16.11	62.18	56.37	11.60	94.8	61.91
PSK-primary, variant_A, variant_B	51.70	55.37	47.91	54.62	15.7	32.3	27.56	68.78	59.12	21.52	92.0	63.40
Sentinel-primary, RunB	50.84	55.29	46.68	53.67	10.5	27.4	22.24	65.22	56.82	15.91	94.1	64.72
Aleix-variant_A	45.73	50.58	42.97	47.92	4.7	20.0	13.01	58.29	51.72	10.21	98.7	52.56
Aleix-primary	34.73	39.75	32.82	36.05	4.1	14.2	9.18	53.22	46.48	8.19	97.4	41.51
TeamSpinoza-Variant_sum-verifier3	17.22	30.52	14.25	19.22	2.6	15.5	8.78	56.44	51.85	8.04	90.0	4.00

Table 6: Ctx-QA automatic evaluation results aggregated across all 24 languages.

Rank	System	Answerable ($N=6912$)	Non-Ans ($N=1728$)	Overall↓ ($N=8640$)
1	CUNI-primary	79.79%	98.21%	83.47%
2	CUNI-variant_B	78.91%	98.84%	82.89%
3	Lanigo-primary	74.12%	97.45%	78.78%
4	Lanigo-variant_A	66.61%	95.02%	72.29%
5	TeamSpinoza-primary	64.15%	91.61%	69.64%
6	TeamSpinoza-Variant_V4	59.69%	91.72%	66.10%
7	TUM-Heilbronn-primary	55.87%	96.76%	64.05%
8	Sentinel-base	54.67%	90.16%	61.77%
9	Aleix-variant_B	53.11%	85.65%	59.62%
10	PSK-variant_A	51.66%	66.67%	54.66%
11	PSK-primary	51.55%	66.90%	54.62%
12	PSK-variant_B	51.56%	66.72%	54.59%
13	Sentinel-primary	47.34%	79.11%	53.69%
14	Sentinel-RunB	47.29%	79.11%	53.66%
15	alkhder-primary	60.00%	0.00%	48.00%
16	Aleix-variant_A	36.66%	92.94%	47.92%
17	Aleix-primary	21.46%	94.44%	36.05%
18	TeamSpinoza-Variant_sum-verifier3	22.12%	7.64%	19.22%

Table 7: Ctx-QA breakdown: Answerable (reading comprehension) vs. Non-Answerable (refusal accuracy).

sub-task. It is worth noting that human evaluation only covered a subset of languages and questions, and the two evaluations asked annotators and LLMs fundamentally different questions. Nevertheless we observe that they display good to strong correlations, with Spearman’s $\rho = 0.79$ and Kendall’s $\tau = 0.62$. The only variations are local, among the top 3 systems, namely LANIGO, SENTINEL, and CUNI, whilst other systems are positioned the same according to both evaluations.

Cross-Lingual Prompts While most open-ended generation prompts require monolingual responses, 11 prompts in the benchmark explicitly instruct the model to produce its response in a different language (e.g., an Arabic prompt requesting a complaint letter in Swiss German, or a German prompt requesting text in Czech; $N=245$ strictly cross-lingual instances across 23 languages). Evaluating these instances reveals three key findings (Table 17):

- **Strong prompt-language inertia:** When sys-

tems fail to switch languages, they overwhelmingly remain trapped in the prompt’s input language ($>60\%$ across many systems) rather than defaulting to English ($\leq 2\%$ across all systems).

- **Multi-stage pipelines struggle with prompt-internal constraints:** Systems that decoupled generation into modular stages struggled when the target language was defined only within the prompt text. Most notably, LANIGO’s two-stage pivot strategy (which generates an intermediate English draft before rewriting into the target language) systematically produced outputs in the prompt’s input language (94.7% inertia, 0.0% on-target), indicating that the second-stage rewrite targeted the document’s language metadata rather than the constraint within the prompt. Similarly, SENTINEL’s verification wrapper (24.1% on-target) checked language compliance against the input language and forced retries in that

System	Maj↓	Cmd-A+	GLM-5.1	Gemma4	Qwen-3.5	LangID
CUNI-variant_B	88.45	79.82	87.26	88.99	90.03	87.2
CUNI-primary	82.25	80.64	81.06	83.06	86.25	84.5
Laniko-primary	76.27	66.63	76.51	76.45	73.62	96.7
Sentinel-base	72.47	68.10	71.79	73.20	75.39	93.7
Sentinel-primary, RunB	65.43	57.79	58.16	65.73	66.20	96.2
Laniko-variant_A	64.26	55.99	64.42	64.76	64.95	97.4
PSK-primary, variant_A, variant_B	63.65	58.40	62.69	63.45	62.88	87.1
Aleix-variant_B	62.37	61.69	62.55	61.95	62.82	81.6
TUM-Heilbronn-primary	60.48	58.26	61.28	60.32	63.57	87.7
TeamSpinoza-Variant_V4	57.79	55.19	58.54	58.08	59.54	91.7
Aleix-primary	55.73	54.94	55.25	56.25	59.37	87.2
Aleix-variant_A	53.66	53.22	53.42	53.96	56.73	87.5
TeamSpinoza-primary	47.91	43.94	47.81	48.99	46.07	93.1
TeamSpinoza-Variant_sum-verifier3	23.55	22.96	24.48	24.76	24.15	91.8

Table 8: OEG automatic evaluation results across all 24 languages.

System	ar	bn	bh	cs	ku	de	en	fi	ht	hi	id	it	ja	ko	mr	fa	pt	ru	sk	es	tr	vi	yo	zh	Avg
CUNI-variant_B	92.4	93.3	88.9	86.0	78.7	93.9	93.6	88.5	86.6	93.3	92.7	92.7	88.9	88.2	92.7	91.7	91.4	93.0	72.6	92.0	92.0	88.5	41.2	90.1	88.5
CUNI-primary	87.9	79.3	59.2	78.0	56.7	91.7	96.8	70.1	66.2	81.5	93.6	93.6	89.8	82.5	64.6	93.6	94.6	95.2	75.2	94.9	84.7	91.1	43.4	93.3	82.3
Laniko-primary	80.3	82.5	80.3	65.9	51.9	86.0	85.7	70.7	67.2	83.1	83.1	83.1	83.4	78.0	81.2	82.8	83.1	82.2	52.2	85.4	76.1	78.0	18.1	85.7	76.3
Sentinel-base	81.2	78.0	58.9	72.3	20.7	84.4	88.5	72.9	37.9	82.2	84.1	84.4	74.5	78.3	77.4	79.9	83.1	83.1	61.8	85.0	78.3	79.9	7.1	77.7	72.5
Sentinel-{primary, RunB}	71.8	71.5	48.2	63.9	15.8	77.2	87.9	62.6	30.6	74.4	81.5	80.6	66.1	69.1	67.0	68.9	79.1	77.4	54.1	80.3	64.8	71.5	8.5	73.6	65.4
Laniko-variant_A	71.7	67.5	46.2	54.8	30.6	74.8	76.4	63.1	54.1	68.8	72.6	76.1	69.7	70.1	64.3	70.7	74.5	75.8	49.0	76.4	63.7	65.9	11.0	72.0	64.3
PSK-{primary, variant_A, variant_B}	71.5	64.3	29.3	73.1	8.3	73.6	71.1	66.8	8.3	72.8	78.5	75.5	68.4	69.2	61.4	76.5	76.8	72.4	69.2	77.5	67.3	76.5	38.6	70.3	63.7
Aleix-variant_B	76.4	47.5	28.7	61.5	6.4	80.3	91.4	47.8	7.0	54.1	83.1	83.4	82.2	79.3	35.4	64.0	86.6	78.3	55.1	86.0	65.9	80.3	2.7	88.5	62.4
TUM-Heilbronn-primary	65.3	54.5	29.6	65.3	24.5	71.0	82.2	52.9	39.5	56.4	75.8	77.4	62.1	62.4	41.4	64.3	72.6	76.8	56.4	77.4	60.8	71.3	14.3	78.0	60.5
TeamSpinoza-Variant_V4	67.8	44.6	21.3	58.0	22.6	74.5	81.2	40.1	28.7	54.8	72.9	72.3	68.2	62.7	38.9	66.9	72.3	70.7	52.2	69.4	61.5	72.6	15.4	79.6	57.8
Aleix-primary	66.2	32.2	11.8	60.8	10.8	75.2	83.4	45.9	11.5	41.1	76.1	79.9	72.3	67.5	22.0	49.7	78.0	78.0	52.9	80.3	62.4	77.1	2.7	77.4	55.7
Aleix-variant_A	59.6	25.5	15.0	60.2	8.9	74.8	81.8	45.2	8.3	26.8	76.4	81.5	64.3	62.7	21.7	51.0	77.4	74.5	53.2	79.6	58.9	76.8	6.6	77.4	53.7
TeamSpinoza-primary	52.9	36.0	14.0	44.6	22.9	56.7	77.1	31.2	31.2	37.9	59.6	62.4	51.0	56.7	33.1	47.8	66.6	59.2	40.1	68.2	43.0	51.6	20.3	74.2	47.9
TeamSpinoza-Variant_sum-verifier3	22.9	2.9	2.9	13.7	1.3	31.2	56.4	7.0	3.8	6.7	40.8	37.3	31.5	23.2	2.2	7.0	44.6	31.2	10.8	51.9	13.4	47.5	1.6	64.3	23.6

Table 9: OEG breakdown: per-language results (LLM judge ensemble majority-vote rule pass rate (%), primary benchmark metric).

language. Looking at the overall language on-targetness in Table 8, it seems that this strategy ultimately paid off for the whole set, with these systems still being in the top.

- **Prompt engineering retains steerability:** Single-pass prompt-engineered models (CUNI-VARIANT_B at 89.0% and CUNI-PRIMARY at 81.6%) followed language-switch constraints with high fidelity, demonstrating that outer pipeline wrappers can inadvertently override an LLM’s intrinsic instruction-following steerability. However, the overall language compliance of these systems was lower than the multi-stage ones.

5.4 Cross-lingual Summarization

Table 10 presents evaluation results for cross-lingual summarization across all 24 target languages. The official benchmark ranking is governed by **Language-Gated ROUGE-L** (G-R-L), which assigns zero credit to summaries failing language identification. Fluency judgments from an LLM verifier are presented in Table 19.

System Performance. LANIQUO-VARIANT_A achieves the highest gated ROUGE-L (17.55%) and NLI mutual entailment (9.12%), followed closely by CUNI-VARIANT_B (17.41% G-R-L, 38.0% G-chrF++, 81.98 COMET) and LANIQUO-PRIMARY (16.95% G-R-L, 99.8% LangID). On neural MT metrics, CUNI-VARIANT_B and CUNI-PRIMARY attain the strongest semantic fidelity (81.98 and 81.54 COMET, respectively). TEAMSPINOZA-PRIMARY achieves third place according to the LLM-as-a-verifier that checks language adherence and fluency, however, it is penalized by the reference-based Rouge-L.

Knowledge Distillation and Script Adherence.

LANIQUO-PRIMARY’s full fine-tuning with teacher distillation from Gemma-4-31B paired with pivot decoding yields the most robust language adherence across the benchmark (99.8% LangID). This proves decisive on challenging low-resource scripts such as Bhojpuri, where LANIQUO-PRIMARY generates fluent target-language text (98.6% LangID) while most other systems fail completely, often falling below 35% LangID.

System	G-R-L↓	G-chrF	Raw R-L	Raw chrF	R-1	R-2	BLEU	COMET	CQE	NLI-ME	NLI-F	LangID
Lanigo-variant_A	17.55	34.5	19.23	34.5	32.88	9.95	6.5	77.88	60.51	9.12	19.54	98.5
CUNI-variant_B	17.41	38.0	18.93	38.3	34.98	10.05	6.5	81.98	63.54	8.84	19.63	97.5
Lanigo-primary	16.95	37.8	18.51	37.8	34.51	9.74	6.4	81.05	61.06	7.83	16.69	99.8
PSK-primary	15.66	28.1	17.30	28.1	30.07	7.51	4.2	75.12	58.45	8.28	17.75	98.1
CUNI-primary	14.71	35.4	16.35	35.8	32.02	7.94	5.0	81.54	62.88	8.59	16.72	93.6
Sentinel-primary	14.11	35.0	15.44	35.0	29.39	7.60	4.6	79.34	61.49	8.43	18.93	95.5
Sentinel-base	14.11	35.0	15.44	35.0	29.39	7.60	4.6	79.34	61.49	8.43	18.93	95.5
Sentinel-RunB	13.44	27.6	14.79	27.6	26.27	6.26	3.6	72.72	58.88	7.60	19.03	95.9
TUM-Heilbronn-primary	12.50	22.4	14.58	25.8	24.65	6.91	4.4	70.34	61.43	8.10	29.87	75.3
TeamSpinoza-Variant_sum-verifier3	12.33	28.3	13.73	28.3	25.64	5.31	3.1	73.58	58.46	7.49	14.18	94.5
Alex-variant_A	12.06	22.6	13.67	23.3	23.48	5.90	3.4	66.85	58.39	8.17	23.17	87.7
TeamSpinoza-Variant_V4	10.87	18.2	12.12	18.3	19.97	5.44	2.6	63.78	53.19	7.60	20.97	93.4
Alex-primary	8.85	15.9	10.56	18.0	17.77	4.56	2.4	58.79	57.76	8.18	34.70	66.9
Alex-variant_B	7.81	17.0	10.38	22.9	18.02	4.63	3.0	64.50	62.66	8.05	39.12	52.1
TeamSpinoza-primary	7.42	9.8	8.42	9.9	12.38	3.29	0.8	54.80	47.51	8.95	21.62	92.1

Table 10: Sum automatic evaluation results across all 24 target languages.

6 Rankings and Discussions

6.1 Overall Ranking

Table 11 presents an overall cross-task ranking of all participant systems across the three subtasks, based on automatic evaluation. Each system’s overall rank is computed as the mean of its per-task ranks, where per-task ranking follows the primary benchmark metric: LLM ensemble majority-vote accuracy for Ctx-QA and OEG, and Language-Gated ROUGE-L F1 for Summarization.

Overall Ranking Based on Automatic Evaluation. CUNI-VARIANT_B achieves the top overall rank (mean rank 1.67), with remarkably consistent performance: #2 in Ctx-QA, #1 in OEG, and #2 in Summarization. CUNI-PRIMARY (mean rank 2.67) wins Ctx-QA but drops to #5 in Summarization. LANIQO-PRIMARY (mean rank 3.00) is the most balanced system, placing #3 across all three tasks.

LANIQO-VARIANT_A (mean rank 4.00) wins the Summarization track (G-R-L 17.55) and places #4 in Ctx-QA but drops to #7 in OEG. SENTINEL-BASE rounds out the top five (mean rank 6.00), with particular strength in OEG (#4).

Individual Sub-Task Winners

- **Context-based Question Answering:** CUNI-PRIMARY (81.63% Maj)
- **Open-Ended Generation:** Based on human evaluation, the first three places belong to LANIQO-PRIMARY, SENTINEL-PRIMARY, and CUNI-PRIMARY, respectively. According to automatic evaluation, CUNI-VARIANT_B takes the lead with 88.45% ensemble verification.
- **Summarization:** LANIQO-VARIANT_A (17.55 G-R-L)

Most Truthful System. TUM-HEILBRONN-PRIMARY achieves the highest Non-Answerable F1 score (87.07%), demonstrating the strongest refusal accuracy and lowest hallucination rate on unanswerable questions, followed closely by TEAMSPINOZA-VARIANT_V4 (86.50%) and TEAMSPINOZA-PRIMARY (86.40%).

Most Equitable System. LANIQO demonstrates the most uniform cross-lingual performance under human evaluation (Section 4), maintaining the benchmark’s lowest cross-lingual variance ($\sigma = 0.37$), the highest worst-case score floor (4.93 in Bhojpuri), and never falling below rank 4 across all 14 languages. Under automatic metrics across all 24 languages, CUNI-VARIANT_B also exhibited remarkable consistency, achieving a minimax rank of 3 (with English being its relative weakness).

6.2 Insights

Post-editing effort is necessary for realistic benchmarking. We initially intended post-editing purely as a quality assurance step, however, it is in itself a valuable signal of the disparity in language representation. After discussions with the annotators, we realized that the time and effort needed to post-edit each language largely varied even though we used state-of-the-art LLMs to perform the translation/localization and despite most prompts being relatively simple. This highlights that we cannot expect to reach truly multilingual models by simply translating benchmarks (Chen et al., 2024).

Monolingual QA has a higher accuracy than answering a non-English question about a passage in English. Constructing the Context QA test set with redundant questions paired with different language mutations of the same passage allowed us to examine the effect of passage-question language pairing on the accuracy of the answers. As shown

Overall Rank	System	Mean Rank	Ctx-QA Rank	OEG Rank	Sum Rank	Ctx-QA Maj (%)	OEG Maj (%)	Sum G-R-L
1	CUNI-variant_B	1.67	#2	#1	#2	80.93	88.45	17.41
2	CUNI-primary	2.67	#1	#2	#5	81.63	82.25	14.71
3	Lanigo-primary	3.00	#3	#3	#3	76.90	76.27	16.95
4	Lanigo-variant_A	4.00	#4	#7	#1	69.80	64.26	17.55
5	Sentinel-base	6.00	#7	#4	#7	59.58	72.47	14.11
6	PSK-primary	7.33	#10	#8	#4	51.70	63.65	15.66
7	Sentinel-primary	7.33	#11	#5	#6	50.84	65.43	14.11
8	Sentinel-RunB	8.67	#12	#6	#8	50.84	65.43	13.44
9	TUM-Heilbronn-primary	9.00	#8	#10	#9	59.16	60.48	12.50
10	TeamSpinoza-Variant_V4	9.67	#6	#11	#12	61.96	57.79	10.87
11	Aleix-variant_B	10.67	#9	#9	#14	57.08	62.37	7.81
12	TeamSpinoza-primary	11.33	#5	#14	#15	67.50	47.91	7.42
13	Aleix-variant_A	12.33	#13	#13	#11	45.73	53.66	12.06
14	Aleix-primary	13.00	#14	#12	#13	34.73	55.73	8.85
15	TeamSpinoza-Variant_sum-verifier3	13.33	#15	#15	#10	17.22	23.55	12.33

Table 11: Overall ranking based on automatic evaluation. Systems ranked by the mean of per-task ranks.

in Table 15 in the appendix, most systems had a higher accuracy in a monolingual setting (i.e. passage and question in the same language) compared to cross-lingual from English. Moreover, the accuracy further drops if neither of the languages is English.

7 Conclusion

In this paper, we presented the findings of the WMT26 Shared Task on Multilingual Instruction Tuning (MIST). Focusing squarely on open-weight language models within an accessible ≤ 10 B parameter budget, MIST established a standardized, multi-faceted benchmark across 24 typologically diverse languages and three core sub-tasks: context-based question answering, multi-constraint open-ended generation, and cross-lingual scientific summarization. Across submissions from eight participating teams and six organizer baselines, our findings provide several key insights into the current state and limits of multilingual instruction following in compact models:

Modeling Strategies and Empirical Trade-offs.

Task-specific adaptation proved essential: untuned open-weight baselines suffered from pervasive hallucinations and poor refusal discipline, whereas tailored participant submissions substantially closed the gap to frontier capabilities. Notably, structured system prompting and reasoning (CUNI) achieved top verification accuracy on context QA and constraint satisfaction in open-ended generation, while teacher distillation and full fine-tuning (LANIQO) provided unmatched target-script adherence and led in human-assessed fluency. Fur-

thermore, specialized inference-time mechanisms, such as TEAMSPINOZA’s answerability probe, demonstrated that isolating answerability classification from text generation effectively bounds non-answerable hallucinations. Conversely, standard supervised fine-tuning optimized for lexical overlap (TUM-HEILBRONN) produced verbose generations with high surface overlap (BLEU/ROUGE) but reduced precision under strict verifiers, highlighting the decoupling between n-gram metrics and functional instruction compliance.

Cross-Lingual Gaps and Language Gating.

Our benchmark reveals that compact models are not yet universal polyglots. While top systems attain 80–87% verification accuracy on high-resource European languages, performance drops steeply on distant and low-resource targets (e.g., falling below 42% in Yoruba and remaining challenging in Kurdish). Encouragingly, the surprise evaluation on Bhojpuri demonstrated robust zero-shot cross-lingual transfer from Hindi data for leading systems. In cross-lingual summarization, our results underscore the necessity of language-gated evaluation: without explicit penalties for language drift, models frequently default to English leakage, fundamentally failing the cross-lingual objective.

Human Evaluation and the Instruction-Following Bottleneck.

Human evaluation across 14 languages shows that within the top cluster, target-language fluency (Naturalness 5.51–5.74 and Coherence 5.59–5.68) is tightly grouped. Instead, complex, multi-constraint *Instruction Following* serves as the primary differentiator, exhibiting a spread of 1.67 points across all

systems (3.69–5.36). Moreover, multi-judge LLM verifier panels displayed strong agreement with human rankings, affirming the viability of automated majority-vote verifiers for reproducible multilingual benchmarking.

All test suites, localized prompts, system outputs, baselines, human annotation rubrics, prompts, and evaluation code are publicly released. We hope these resources will help in measuring the usefulness of large language models across multiple languages and consequently their improvement.

8 Limitations

Unfortunately, we were not able to perform human evaluation for all tasks. Given that both context-based questioning and summarization have references, we opted to focus on evaluating open-ended generation. We do acknowledge that especially for summarization, reference-based evaluation might not be the best given that there are multiple valid ways of summarizing a paper that can diverge from the provided reference.

9 Acknowledgements

We thank all participants, without whom the task would not be so fruitful.

We also extend our heartfelt gratitude to many people for reviewing and post-editing instructions. They are: Adam Gaber, Ahrii Kim, Aina Garí Soler, Ayodele Awokoya and her students, Danti Ramadanti, Eduardo Calo, Erica Cooper, Hale Sirin, Koel Dutta Chowdhury, Leo Leppänen, Nalin Kumar, Nathaniel Robinson, Nazanin Shafiabadi, Peter Polák, Sina Ahmadi, and Sonal Joshi. Sincerely, thank you once again for making MIST possible.

We are immensely grateful to our sponsors Toloka and Voice Arena for supporting our human evaluation. Thank you for helping us make sure that NLP improvements are actually helpful to humans across many languages.

The evaluation pipeline for 10 languages was co-designed with an AI coding assistant provided by the Toloka platform; the human authors specified the task, reviewed every intermediate configuration, and are responsible for all final design decisions.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Patrícia Schmidtová is supported by Charles University projects GAUK 252986 and SVV 260 698

and co-funded by the European Union (ERC, NG-NLG, 101039303).

Part of this work is funded by the European Union’s Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People).

References

- Team Alibaba, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Hasan Alkhder, Maria Hanini, Imane Hocine, Maher Pasa, and Amro Najjar. 2026. A monolingual LoRA-tuned qwen2.5-3b system for arabic context-based question answering. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. [CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564. Association for Computational Linguistics.
- Adrian Charkiewicz and Artur Nowakowski. 2026. Lanigo at WMT26 multilingual instruction shared task (MIST): Full-parameter knowledge distillation with task-conditional pivot decoding under a 10b-parameter constraint. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. [Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726. Association for Computational Linguistics.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A benchmark](#)

- for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Asher Uyiosa Egerton-Idehen. 2026. SentinelQA: Task-level routing and output verification for multilingual instruction following. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jen-Chien Hou, Tsedeniya Kinfe Temesgen, and Alexander Fraser. 2026. TUM-heilbronn: WMT26 multilingual instruction shared task. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218. Association for Computational Linguistics.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. [Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.
- Lea Krause, Stefan F. Schouten, Michiel van der Meer, and Piek T.J.M. Vossen. 2026. Probe before you answer! answerability probing for abstention in cross-lingual context-based question answering. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. [WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048. Association for Computational Linguistics.
- Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and others. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.
- Aleš Manuel Papáček and Patricia Schmidtova. 2026. CUNI-UFAL at WMT26 multilingual instruction shared task. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2026. [MCIF: Multimodal crosslingual instruction-following benchmark from scientific talks](#). In *The Fourteenth International Conference on Learning Representations*.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.
- Srikar Kashyap Pulipaka. 2026. PSK at WMT 2026 MIST: Task-specialized QLoRA adapters for multilingual summarization and question answering. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaei, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and others. 2026. [Tiny aya: Bridging scale and multilingual depth](#). *Preprint*, arXiv:2603.11510.
- Aleix Sant, Jordi Luque Serrano, and Carlos Escolano Peinado. 2026. Task-preserving multilingual instruction tuning for open-ended QA at WMT 2026 MIST. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Het-tiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567. Association for Computational Linguistics.
- Gemma Team. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.

A Prompt Templates

This section provides the full set of prompts used across the three sub-tasks of the WMT26 Shared Task on Multilingual Instruction Tuning (MIST), as well as the prompts used for LLM-as-a-judge automatic evaluation.

A.1 Context-based Question Answering (QA-Context)

In the context-based QA sub-task, models receive a background passage in an assigned source language alongside a question and instruction framed in a target language. The prompt requires a concise, single-sentence response grounded strictly in the provided text. When the requested information is absent, models are instructed to output that the question cannot be answered (e.g., “not answerable” in English).

Benchmark Task Prompt (English) The standard English prompt template is given below:

```
I have a question about this passage:

{passage}

The question is: {question}

Could you please answer in one sentence, using only
what the passage says? If the passage doesn't
give the answer, just write "not answerable".
Please answer in English.
```

A.2 Open-Ended Generation (OEG)

The open-ended generation sub-task includes 100 manually authored English seed prompts spanning three functional categories:

1. **General:** Creative, reflective, or reasoning tasks without specific geographical grounding.
2. **Factual:** Knowledge retrieval, structured summaries, and multi-step explanations.
3. **Language/Location-specific:** Tasks targeting linguistic conventions, local customs, or regional institutions, parameterized using {language} and {country} placeholders.

Representative Seed Prompt Examples Below are illustrative examples from each category:

- **General (Constraint: 150 words):**

“Draft a 150-word description of a technological advancement that excites or concerns you, explaining how it works, its potential impacts on society, and your thoughts on its ethical implications or benefits.”

- **Factual (Constraint: Metaphorical framing):**

“Explain quantum entanglement using only cooking metaphors.”

- **Language-specific (Parameterized):**

“How are diminutives typically formed in {language}? Provide 3 examples with English translations.”

- **Location-specific (Constraint: 160 words):**

“Compose a 160-word text describing a local business or establishment that you value in your community, explaining what they offer, why you choose to support them, and what role they play in the character or vitality of your neighborhood.”

A.3 Cross-Lingual Scientific Summarization (Sum-Sum)

For the cross-lingual summarization sub-task, models read a full scientific paper from ACL 2026 in English and generate an abstract in the target language. To prevent evaluation artifacts tied to a single prompt phrasing, each instance is paired with one of five paraphrased prompt templates, assigned uniformly at random.

Paraphrased Prompt Template Each instance is paired with one of five paraphrased prompt templates, assigned uniformly at random. A representative prompt template is:

```
The following is the full text of an *ACL paper in
English. Write the paper's abstract in around
200 words, using formal third-person academic
prose.

PAPER:
{paper}

Output the abstract only, without any reasoning,
explanation, or additional text.
```

The full set of five paraphrased prompt variants is available in the shared task repository.

B Post-Editing Distance and Human Effort

Table 12 reports the detailed post-editing distance metrics between the original machine translations (predominantly GPT-5.5) and the finalized human post-edited test prompts across all 23 non-English target languages, evaluating all $N = 100$ prompts per language ($N = 59$ for Yoruba). We evaluate human correction effort using six complementary metrics:

- **% Unchanged (% Unch.):** The percentage of prompt instances accepted completely verbatim with zero human edits.

- **chrF**: Tokenization-independent character n -gram F-score ($n = 6, \beta = 2$), which robustly captures subword and morphological preservation across diverse alphabetic, abjad, and logographic scripts.
- **BLEU**: Corpus-level BLEU score computed against the post-edited gold references.
- **EditDist**: The mean character-level Levenshtein edit distance per prompt (minimum character insertions, deletions, and substitutions).
- **CER (%)**: Character Error Rate, computed as $100 \times \frac{\sum \text{Levenshtein}(h, r)}{\sum |r|}$.
- **TER**: Translation Edit Rate (HTER) measuring normalized word-level edits (insertions, deletions, substitutions, and phrase shifts).

As shown in Table 12, the results highlight clear distinctions in translation difficulty and editing intensity:

- **Easy (Minimal post-editing)**: For high-resource and morphologically well-supported languages, the initial translations were almost universally adopted as-is. Arabic had 98.0% unchanged prompts (chrF 99.8, BLEU 99.6), where annotators accepted fluent Modern Standard Arabic (MSA) phrasing without seeking dialectal or stylistic alterations. Similarly, Hindi (94.0% unchanged, chrF 99.5), Indonesian (91.0% unchanged), German (88.0%), Portuguese (85.0%), and Russian (85.0%) required minimal human intervention.
- **Frequent but Low-Intensity Edits (Morphological and Stylistic Tuning)**: Slavic languages such as Czech (35.0% unchanged) and Slovak (37.0% unchanged) had approximately 63–65% of sentences modified by annotators. However, their editing intensity was remarkably low: CER remained below 4.0% (2.6% for Czech, 3.9% for Slovak) and chrF exceeded 95.5. Analysis of the diffs confirms that these edits consisted predominantly of subtle inflectional refinements, diacritic adjustments, and formal polite phrasing rather than structural rewrites.
- **Difficult and High-Intensity Edits**: Haitian Creole presented the greatest post-editing difficulty, with only 8.0% of prompts accepted verbatim, a CER of 22.6%, and a TER

of 29.1, requiring comprehensive human re-translation. Kurdish likewise demanded extensive editing (15.0% unchanged, CER 21.0%, TER 27.0). Persian (Farsi) also proved particularly challenging (44.0% unchanged, CER 16.1%, TER 21.5).

Reproducibility and Metric Parameters All lexical and edit metrics were computed using sacrebleu (Post, 2018) version 2.5.1. For full reproducibility, the metric signatures and parameter configurations are as follows:

- **BLEU**: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1` (with `tok:zh` for Chinese; for Japanese, unsegmented 13a yields 77.5, whereas character-level tokenization yields 93.7, consistent with its 94.3 chrF).
- **chrF**: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`
- **TER**: `nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.5.1`
- **Levenshtein & CER**: Minimum character-level edit distance computed via dynamic programming, normalized by total reference character length.

C Human Evaluation Setup Details

Toloka Campaign On the Toloka crowdsourcing platform, the evaluation pipeline was designed and deployed with the assistance of an AI agent integrated into the Toloka platform environment, which generated initial pipeline configurations, quality-control criteria, and training materials, iteratively refined in collaboration with the organizers who reviewed and approved all final pipeline parameters.

Task and Data Flow. The pipeline evaluated OEG outputs across 10 diverse languages: Arabic (arb), Chinese (zho), German (deu), English (eng), Indonesian (ind), Italian (ita), Japanese (jpn), Korean (kor), Russian (rus), and Spanish (spa). For each language, 70 prompts were paired with 6–7 anonymized system responses, yielding 700 items in total. A language-based filtering was deployed, ensuring that evaluators only accessed tasks in their qualified language. Annotators evaluated responses on the 7-point calibrated scale (*poor* → *poor/flawed* → *flawed* → *flawed/good* → *good* → *good/great* → *great*) across the three dimensions.

Annotator Pool and Onboarding. We restricted recruitment exclusively to Toloka’s *Verified Experts*

Language	N	% Unch.	chrF	BLEU	EditDist	CER (%)	TER
Arabic	100	98.0%	99.8	99.6	0.1	0.1%	0.2
Bengali	100	83.0%	97.6	96.7	2.9	2.0%	2.1
Bhojpuri	100	43.0%	93.0	88.5	10.8	7.7%	8.7
Chinese (Simplified)	100	76.0%	95.8	97.0	1.4	3.0%	21.1
Czech	100	35.0%	95.5	87.6	3.5	2.6%	7.4
Finnish	100	61.0%	95.7	91.4	6.2	4.2%	6.3
German	100	88.0%	99.0	98.2	1.4	0.8%	1.1
Haitian Creole	100	8.0%	76.0	60.6	31.0	22.6%	29.1
Hindi	100	94.0%	99.5	99.4	0.6	0.4%	0.2
Indonesian	100	91.0%	99.1	97.6	1.3	0.8%	1.3
Italian	100	74.0%	97.7	95.8	3.4	2.2%	3.2
Japanese*	100	73.0%	95.0	77.2	2.8	4.2%	23.6
Korean	100	59.0%	94.3	91.5	2.5	3.5%	6.0
Kurdish	100	15.0%	76.0	64.4	30.4	21.0%	27.0
Marathi	100	78.0%	98.0	96.2	2.0	1.4%	2.0
Persian (Farsi)	100	44.0%	83.3	70.7	22.5	16.1%	21.5
Portuguese	100	85.0%	98.7	97.2	1.9	1.2%	2.0
Russian	100	85.0%	97.0	95.1	4.0	2.6%	4.0
Slovak	100	37.0%	95.6	87.0	5.4	3.9%	8.3
Spanish	100	74.0%	97.9	96.2	2.8	1.8%	2.6
Turkish	100	65.0%	97.7	94.5	2.8	2.0%	3.6
Vietnamese	100	45.0%	91.3	87.9	14.4	9.3%	10.1
Yoruba	59	69.5%	94.8	92.8	3.5	3.1%	5.0
Macro Average	–	64.4%	94.3	89.7	6.9	5.1%	8.5

Table 12: Post-editing distance between initial machine translations and final post-edited prompts across all 23 non-English target languages for the OEG sub-task. *Japanese BLEU is reported using standard sacrebleu 13a tokenization (character-level BLEU is 94.6).

pool: annotators who have passed platform-level identity verification and language proficiency examinations. Per-node language filters ensured that only native or near-native speakers accepted tasks.

Before accessing production tasks, annotators completed a mandatory 6-step training course. The curriculum provided detailed expositions of the three dimensions, calibration anchor points, and critical edge cases. In particular, because responses were pre-truncated to approximately 400 words (marked by [TRUNCATED]) to bound task duration to 5–8 minutes, annotators were instructed to apply a two-sided truncation policy: truncation must not penalize Naturalness or Coherence, but must penalize Instruction Following if the original prompt explicitly specified a length constraint shorter than the pre-truncation cap. Annotators passed two quiz gates and reasoning-aloud practice tasks before admission, with onboarding compensation contingent on approved production work.

Quality Assurance. Quality control operated in a `human_only` QA mode with 10% smart sampling: 10% of production submissions were randomly routed to an independent human reviewer (also a Verified Expert) who evaluated submissions against internal rating consistency, completeness,

and adherence to the truncation policy. Rejected tasks were re-routed to fresh annotators.

VoiceArena Indic Language Campaign. In parallel, VoiceArena conducted human evaluation for four Indic languages—Bengali (`ben`), Bhojpuri (`bho`), Hindi (`hin`), and Marathi (`mar`)—recruiting native annotators directly on their dedicated platform. The annotation protocol, user interface, and evaluation guidelines directly reflect the 10-language Toloka evaluation, employing the identical 7-point calibrated scale across Instruction Following, Naturalness, and Coherence.

D Per-Language Results for Submitted Systems

This section provides comprehensive per-language breakdowns across all 24 benchmark languages for the participant submissions on each of the three evaluation tasks.

D.1 Context-based QA Per-Language Performance

Tables 13 and 14 report per-language verification accuracy for participant systems and organizer baselines evaluated by Gemma-4-31B and GLM-5.1 verifiers, respectively.

System	ar	bn	bh	cs	ku	de	en	fi	ht	hi	id	it	ja	ko	mr	fa	pt	ru	sk	es	tr	vi	yo	zh	Avg
CUNI-primary	89.2	80.4	75.4	61.7	82.5	78.3	87.5	81.2	71.7	85.4	90.8	66.7	75.0	91.7	86.7	90.8	60.8	87.5	33.3	84.8	89.2	86.7	75.0	88.5	79.2
CUNI-variant_B	90.5	78.3	76.2	59.6	66.7	79.2	87.4	83.8	65.8	87.1	89.2	65.8	89.2	89.2	85.8	97.5	59.2	85.2	21.7	85.8	87.5	84.2	63.3	93.3	78.0
Lanigo-primary	86.8	82.5	74.2	51.2	59.2	66.2	83.6	76.7	63.7	81.2	81.7	58.3	70.8	84.2	80.8	90.8	56.7	84.8	25.8	81.8	84.2	77.5	51.7	89.0	72.6
gemma_baseline	85.5	77.9	69.2	51.2	59.2	67.5	82.5	75.4	64.6	80.0	79.2	57.5	64.2	90.0	81.7	85.8	54.2	83.5	22.5	78.5	72.5	80.0	42.5	90.0	70.6
Lanigo-variant_A	75.0	72.9	65.8	51.2	49.6	61.3	79.0	68.8	58.8	74.2	74.6	57.5	72.5	80.0	86.7	81.7	56.2	74.8	25.8	73.7	78.3	65.0	41.7	79.5	66.9
TeamSpinoza-primary	79.7	68.3	48.8	55.8	48.8	66.2	74.8	65.8	55.4	55.0	74.2	60.0	54.2	80.8	61.7	71.7	54.6	74.6	38.3	74.8	70.0	70.8	53.3	73.8	63.8
TeamSpinoza-Variant_V4	69.3	62.1	63.7	61.7	59.2	57.5	68.8	66.7	55.4	66.2	65.4	47.5	58.3	67.5	64.2	67.5	55.8	68.1	50.0	67.3	62.5	63.3	58.3	66.7	62.2
TUM-Heilbronn-primary	65.0	67.5	61.3	48.8	57.5	54.2	68.6	64.6	55.8	68.3	61.3	46.7	59.2	71.7	82.5	68.3	45.0	66.7	24.2	70.0	59.2	64.2	30.0	62.7	59.3
Sentinel-base	60.3	65.0	53.8	45.8	42.1	59.6	69.8	62.1	42.1	62.9	65.0	51.7	54.2	70.8	60.8	70.0	50.0	65.6	25.0	60.3	64.2	63.3	38.3	68.8	57.1
Alexi-variant_B	58.5	61.7	55.8	38.8	35.4	52.5	68.5	55.0	49.6	60.8	58.8	48.3	55.8	63.3	64.2	67.5	47.5	63.3	25.0	60.8	70.0	63.3	15.0	63.3	54.3
qwen3_baseline	55.7	61.7	52.9	37.9	37.5	53.3	67.4	49.2	55.8	58.3	59.2	47.5	56.7	60.0	56.7	59.2	49.2	57.3	25.8	58.0	55.0	56.7	13.3	60.0	51.8
PSK-{primary, variant_A, variant_B}	60.6	51.7	50.4	44.0	17.8	49.9	63.8	49.0	25.7	48.3	52.8	40.8	40.6	65.0	61.1	66.7	52.1	57.8	27.5	57.5	55.8	46.4	26.1	55.8	48.6
llama_baseline	53.3	52.1	48.3	36.2	41.2	52.1	69.4	45.0	40.4	50.0	47.1	45.8	45.8	57.5	45.0	65.0	44.6	61.7	25.8	55.2	45.8	34.2	31.7	68.0	48.4
aya_baseline	59.3	29.6	42.9	40.4	26.7	47.1	65.0	42.1	30.8	45.4	54.6	44.2	54.2	53.3	48.3	58.3	53.8	62.5	28.3	56.5	53.3	40.8	49.2	61.8	47.9
Sentinel-{primary, RunB}	51.6	54.4	44.4	36.0	37.3	46.0	65.1	42.3	34.8	48.8	55.0	36.7	37.9	57.1	50.4	62.9	50.6	57.9	36.7	54.9	45.4	39.2	28.3	50.1	46.8
glm4_baseline	49.7	51.7	43.8	42.5	34.6	43.3	63.1	43.8	32.5	42.5	42.9	40.0	53.3	45.8	36.7	52.5	39.2	57.5	27.5	55.0	49.2	44.2	27.5	53.2	44.7
Alexi-variant_A	46.5	45.0	44.2	33.3	28.3	45.8	55.0	46.2	38.3	49.6	44.6	35.8	51.7	58.3	55.8	45.0	35.4	49.0	20.0	52.2	46.7	49.2	19.2	50.2	43.6
phi4mini_baseline	41.2	41.2	32.1	30.8	22.9	47.1	52.3	37.9	22.1	36.7	38.8	40.8	52.5	47.5	35.8	47.5	37.9	49.4	19.2	45.0	30.0	24.2	6.7	51.0	37.1
Alexi-primary	31.3	35.4	28.7	25.4	20.0	32.1	44.9	30.4	27.9	39.2	31.7	30.8	41.7	45.0	46.7	34.2	28.7	32.5	20.0	36.2	39.2	40.0	11.7	32.8	32.8
TeamSpinoza-Variant_sum-verifier3	13.3	18.3	16.2	12.1	6.2	14.6	26.5	9.2	4.6	24.2	20.8	9.2	20.8	26.7	22.5	19.2	23.3	15.8	7.5	19.3	11.7	18.3	8.3	16.8	16.1

Table 13: Ctx-QA: Gemma-4-31B verification accuracy per language across submitted systems and organizer baselines. Best per column in **bold**.

System	ar	bn	bh	cs	ku	de	en	fi	ht	hi	id	it	ja	ko	mr	fa	pt	ru	sk	es	tr	vi	yo	zh	Avg
CUNI-primary	81.7	78.9	72.5	76.7	67.5	82.5	86.7	80.3	75.8	78.9	82.2	83.3	81.4	80.3	73.1	83.1	81.9	76.7	73.1	83.1	78.6	81.9	43.6	81.4	77.7
CUNI-variant_B	82.5	80.3	79.2	74.7	69.2	82.8	85.3	72.5	68.6	80.3	81.4	84.7	78.6	78.6	85.0	79.7	83.3	78.3	66.9	81.9	80.3	81.4	27.5	78.9	76.7
Lanigo-primary	79.4	75.0	73.1	70.3	61.4	77.2	80.8	67.2	65.8	78.9	78.1	77.5	72.2	74.7	78.3	76.4	77.2	74.2	63.9	78.1	74.2	83.3	23.3	70.0	72.1
Lanigo-variant_A	69.7	66.4	65.6	65.8	46.4	72.8	77.8	61.9	60.3	70.8	75.8	76.4	64.2	67.8	65.3	71.4	75.8	71.1	55.0	73.9	63.9	74.7	12.5	65.0	65.4
TeamSpinoza-primary	71.1	61.7	53.9	60.0	35.6	68.6	78.1	60.0	55.0	66.4	70.3	66.4	68.1	76.1	55.3	69.2	71.9	69.4	57.8	75.0	66.9	72.5	14.7	78.6	63.4
TeamSpinoza-Variant_V4	60.8	53.1	42.8	54.7	35.6	67.8	77.2	54.7	46.4	55.3	66.3	66.9	57.2	57.5	50.3	62.2	69.2	66.4	53.6	73.1	55.3	66.1	15.0	60.3	56.9
Sentinel-base	58.6	64.7	44.2	61.1	23.9	63.9	76.9	58.9	35.3	55.8	69.2	56.7	59.7	51.7	45.0	60.6	62.2	65.8	48.9	65.6	46.1	72.5	8.1	61.9	54.9
TUM-Heilbronn-primary	56.9	46.1	40.3	58.6	29.4	65.0	76.4	53.1	48.3	49.2	55.3	69.4	56.7	55.0	48.1	58.3	65.8	61.7	51.1	66.9	56.7	60.8	20.0	58.3	54.5
Alexi-variant_B	56.4	38.9	31.4	52.5	21.1	66.9	79.4	44.7	27.5	53.9	64.7	67.2	61.1	56.9	47.5	54.2	69.7	59.4	52.5	71.4	47.5	65.3	11.9	67.2	52.9
PSK-{primary, variant_A, variant_B}	54.7	43.6	39.2	47.2	1.4	61.1	67.2	46.9	21.4	56.9	52.8	58.1	47.5	47.2	54.7	51.7	62.2	56.4	49.4	61.7	52.5	49.4	20.3	46.1	47.9
Sentinel-{primary, RunB}	51.9	43.9	33.3	46.9	15.8	51.4	64.7	51.4	36.7	48.6	56.4	60.6	45.0	40.8	42.5	51.7	63.3	56.4	45.6	58.9	44.7	53.9	8.6	47.2	46.7
Alexi-variant_A	54.4	21.1	21.1	49.2	20.3	61.9	72.5	41.1	27.5	28.1	56.7	41.9	45.6	45.8	21.4	35.8	61.4	58.6	46.7	59.7	31.1	62.8	5.6	61.1	43.0
Alexi-primary	35.6	21.9	22.5	30.3	19.7	50.0	57.5	36.9	25.3	30.3	47.2	23.9	38.9	48.6	20.3	26.1	40.3	38.1	32.5	37.2	21.4	51.1	6.1	26.1	32.8
TeamSpinoza-Variant_sum-verifier3	11.1	6.4	4.4	8.9	3.3	19.7	41.4	3.6	4.4	6.1	19.4	23.6	13.9	13.6	3.1	4.4	33.6	22.8	8.9	27.2	8.6	20.3	3.3	29.7	14.2

Table 14: Ctx-QA: GLM-5.1 verification accuracy per language across submitted systems. Best per column in **bold**.

D.2 Monolingual vs. Cross-Lingual QA Performance

Table 15 details verification accuracy comparing monolingual instances ($L_q = L_p$) against cross-lingual instances ($L_q \neq L_p$) across submitted systems and organizer baselines, evaluated by the Gemma-4-31B verifier. We also report directional cross-lingual breakdowns (synthesizing foreign contexts into English, transferring English contexts into foreign target languages, and cross-lingual synthesis between two non-English languages) alongside accuracy broken down by passage context resource tier (P_{High} , P_{Mid} , P_{Low}).

D.3 Open-Ended Generation (OEG)

Table 16 reports per-language Gemma-4-31B rule pass rates across all 24 target languages for the submitted systems and organizer baselines.

We also report the language confusion in the strictly cross-lingual OEG instances in Table 17.

D.4 Cross-Lingual Summarization (Sum-Sum)

Table 18 displays per-language ROUGE-L F1 scores across all 24 languages for the submitted summarization systems.

E Organizer Baseline Results

Since baselines were evaluated with a single LLM verifier (Gemma-4-31B-it) rather than the full multi-judge ensemble used for participant submissions, we report their results separately. Table 20 presents the cross-task comparison including all six baselines alongside participant systems, ranked by mean Gemma-4-31B verification score. Per-language baseline breakdowns for QA and OEG are included in the per-language tables in Sections D.1 and D.3.

F Human Evaluation Guidelines

This section reproduces the exact annotation guidelines and instructions provided to human evaluators for the Open-Ended Generation (QA-OEG) evaluation campaign.

F.1 Task Overview and Instructions

Annotators were presented with the following task instructions:

In this task, you will evaluate a chatbot’s response to a user’s request in {{language}}. You will evaluate the response using the following quality dimensions:

System	Overall	Language Pairing			Cross-Lingual Direction			Passage Resource Tier		
		Same	Diff	Gap	Into-Eng	From-Eng	Foreign-2-Foreign	High	Mid	Low
CUNI-PRIMARY	80.1	83.7	80.0	3.8	85.3	84.9	77.8	83.6	75.8	74.3
CUNI-VARIANT_B	79.3	83.5	79.1	4.4	82.3	84.6	76.7	83.4	76.2	66.8
LANIQO-PRIMARY	75.4	80.9	75.2	5.7	78.6	80.9	72.7	80.0	71.6	62.8
LANIQO-VARIANT_A	69.0	77.4	68.6	8.7	77.6	76.4	65.1	73.5	65.7	55.3
TEAMSPINOZA-PRIMARY	67.3	76.3	66.9	9.3	75.1	72.6	64.3	72.4	64.1	50.2
TEAMSPINOZA-VARIANT_V4	62.3	72.5	61.9	10.7	76.8	64.9	60.0	64.8	60.1	55.7
TUM-HEILBRONN-PRIMARY	60.0	73.1	59.5	13.6	73.5	64.6	56.8	63.5	55.9	54.1
SENTINEL-BASE	59.4	71.8	58.9	12.9	73.6	67.1	54.9	64.3	56.1	43.9
ALEIX-VARIANT_B	57.7	73.1	57.0	16.1	75.1	65.8	52.7	62.6	53.9	43.0
PSK-PRIMARY	52.6	71.9	51.8	20.0	62.1	60.6	47.9	57.7	50.0	34.2
SENTINEL-PRIMARY	51.9	70.3	51.1	19.2	59.4	62.2	46.3	57.3	46.9	38.3
ALEIX-VARIANT_A	47.2	67.5	46.3	21.1	67.8	53.5	42.5	51.5	43.2	36.2
ALEIX-PRIMARY	36.2	50.6	35.6	15.0	50.5	43.5	31.8	39.8	32.9	27.2
TEAMSPINOZA-VARIANT_SUM-VERIFIER3	21.3	37.6	20.7	17.0	38.8	26.7	17.4	23.8	19.3	14.7
<i>gemma_baseline</i>	77.2	83.8	76.9	6.9	84.6	82.2	74.4	82.2	73.0	63.7
<i>qwen3_baseline</i>	57.7	77.1	56.9	20.2	81.2	66.3	52.0	62.6	53.2	45.3
<i>llama_baseline</i>	56.5	76.5	55.7	20.7	75.4	68.7	49.6	64.5	47.4	42.1
<i>aya_baseline</i>	54.8	67.1	54.3	12.7	69.6	64.6	49.5	61.7	48.6	37.3
<i>glm4_baseline</i>	51.7	72.9	50.8	22.1	69.6	62.2	45.4	58.3	45.5	35.7
<i>phi4mini_baseline</i>	43.4	68.8	42.3	26.5	65.4	51.0	37.8	50.1	38.0	24.6

Table 15: Ctx-QA verification accuracy (%) across language pairings evaluated by Gemma-4-31B. **Same:** monolingual instances where question and passage languages match ($L_q = L_p$). **Diff:** cross-lingual instances ($L_q \neq L_p$). **Gap:** Same – Diff. **Into-Eng:** foreign passage into English question ($L_q = \text{eng}, L_p \neq \text{eng}$). **From-Eng:** English passage into foreign target question ($L_q \neq \text{eng}, L_p = \text{eng}$). **Foreign-2-Foreign:** non-English passage into different non-English question ($L_q \neq \text{eng}, L_p \neq \text{eng}, L_q \neq L_p$). **High / Mid / Low:** performance as a function of the passage context resource tier (P_{tier}). Best results per column within submitted systems and baselines are in **bold**.

System	ar	bn	bh	cs	ku	de	en	fi	ht	hi	id	it	ja	ko	mr	fa	pt	ru	sk	es	tr	vi	yo	zh	Avg
CUNI-variant_B	94.0	93.2	89.8	88.6	79.4	93.1	95.3	88.5	86.3	92.3	93.2	93.2	91.0	88.6	93.2	92.2	92.8	94.0	76.2	93.7	92.7	90.0	37.5	91.6	88.4
CUNI-primary	90.2	81.5	58.4	78.9	54.6	92.2	97.7	69.3	66.1	86.4	95.5	94.0	91.3	85.7	65.8	94.5	95.5	95.4	75.5	95.8	86.9	92.2	37.8	93.5	82.3
gemma_baseline	88.6	87.8	60.9	77.5	58.0	86.5	88.6	80.2	68.0	88.6	85.4	86.0	86.6	80.1	85.8	83.5	85.8	87.6	63.4	84.7	83.5	84.6	20.8	88.2	78.8
LANIQO-primary	81.3	83.7	81.3	66.5	50.6	85.8	84.0	70.4	69.0	84.8	83.4	84.9	82.7	78.8	81.9	82.7	83.7	82.1	56.1	85.5	75.8	80.3	15.3	86.0	75.7
Sentinel-base	83.1	81.9	60.3	73.7	15.4	85.3	89.0	72.6	37.1	83.8	85.3	85.4	74.8	79.3	80.3	80.8	85.5	84.1	63.0	85.8	80.0	81.8	4.7	79.3	72.2
aya_baseline	81.6	40.8	33.1	82.7	26.0	82.2	84.5	79.6	30.7	61.1	82.2	83.7	82.2	78.0	44.1	79.0	81.6	79.6	81.1	83.0	76.1	81.0	43.9	81.5	69.1
qwen3_baseline	75.8	58.2	26.0	69.1	33.6	84.7	88.2	57.4	47.5	65.3	83.4	84.0	82.8	80.5	53.8	74.0	84.4	79.3	68.6	84.4	75.0	79.5	12.6	85.2	68.1
Sentinel-[primary, RunB]	73.8	73.5	47.6	64.3	11.5	78.7	88.0	59.9	29.2	74.3	81.1	81.1	67.7	69.3	68.4	71.3	79.0	77.9	55.3	82.2	66.7	74.1	7.8	74.4	64.9
LANIQO-variant_A	73.5	69.2	47.6	56.4	31.6	75.5	76.3	63.7	54.0	69.6	74.9	76.8	70.8	71.4	65.3	70.3	74.8	76.0	48.7	77.2	63.7	67.5	10.6	73.1	64.1
PSK-[primary, variant_A, variant_B]	72.1	65.8	29.7	72.5	9.3	73.8	70.8	66.0	8.1	73.5	77.3	76.5	68.1	70.1	61.2	75.3	77.9	71.5	67.4	77.3	67.3	75.9	37.9	71.2	63.2
glm4_baseline	66.6	57.2	36.9	64.5	23.5	76.2	79.9	64.5	20.9	49.8	71.5	76.3	77.2	76.9	41.0	70.8	74.0	74.1	56.0	77.0	66.8	75.8	16.1	86.5	61.7
llama_baseline	59.5	54.1	39.4	66.7	16.0	79.4	87.6	52.2	32.0	69.5	75.2	81.7	70.8	64.0	44.3	62.7	78.6	74.1	50.5	80.4	59.7	75.7	15.0	80.4	61.2
Aleix-variant_B	77.2	50.4	26.7	62.6	3.8	80.3	91.4	43.0	5.7	54.7	83.3	85.0	81.0	79.1	33.2	62.4	85.7	79.3	54.0	86.8	68.1	80.6	0.9	89.6	61.0
TUM-Heilbronn-primary	67.3	54.1	29.4	65.6	23.2	71.1	80.9	51.0	37.8	55.6	76.4	76.5	62.1	61.8	39.2	65.3	74.8	77.5	57.5	78.6	60.4	71.7	13.6	79.7	59.6
TeamSpinoza-Variant_V4	69.4	46.8	24.6	58.3	20.1	74.8	82.5	39.7	31.0	56.7	73.0	72.6	69.0	65.7	39.3	68.0	74.0	71.1	52.4	72.0	62.2	72.3	14.9	81.9	58.0
Aleix-primary	67.9	36.8	12.6	62.2	11.1	77.2	84.1	44.4	11.1	40.7	78.9	78.4	74.2	66.9	23.6	48.7	78.7	78.7	52.6	80.8	63.1	78.5	2.7	78.2	55.5
Aleix-variant_A	61.4	27.9	16.5	62.2	8.3	74.9	83.1	42.6	8.8	27.9	77.1	80.5	66.1	63.4	20.9	51.4	78.0	75.5	52.5	79.6	59.5	77.2	5.9	79.3	53.4
phi4mini_baseline	52.0	32.4	25.2	46.5	15.4	74.5	85.6	41.5	47.5	53.0	66.3	70.9	65.1	52.4	30.0	46.8	69.2	69.7	38.2	78.7	49.3	58.8	19.1	75.4	52.6
TeamSpinoza-primary	54.6	39.1	18.6	45.5	24.8	56.9	77.0	32.2	32.9	39.2	61.4	65.0	53.1	60.9	34.1	50.0	67.9	59.4	41.8	70.5	42.5	52.8	21.7	74.2	49.0
TeamSpinoza-Variant_sum-verifier3	25.1	4.3	3.1	13.0	1.5	32.4	57.6	6.4	3.3	7.1	41.5	39.2	33.2	23.7	2.1	8.2	47.5	34.0	9.6	58.1	13.3	50.6	2.7	67.9	24.4

Table 16: QA-OEG: Gemma-4-31B rule pass rate per language across submitted systems and organizer baselines. Best per column in **bold**.

- **Instruction Following:** How well the chatbot followed all of the user’s instructions. Note that a single request can have multiple instructions.
- **Naturalness:** The degree to which the chatbot’s response appears to be fluent, natural text in $\{\{\text{language}\}\}$, that is appropriate to the request in terms of tone and formality.
- **Coherence:** How well the entire response is logically sound and internally consistent.

F.2 Rating Scale and Slider Interface

Annotators provided separate ratings for each dimension on a continuous scale from **Poor** to **Great**:

You will provide a separate rating for each dimension on a scale from Poor to Great. There are 7 different values you can select using buttons, consisting of 4 labeled “landmark” ratings (Poor, Flawed, Good, Great; descriptions provided below) and 3 “in-between” ratings (Poor/Flawed, Flawed/Good, Good-/Great).

Note that the default value “N/A” is not a valid rating; you must select a rating on each slider before the task can be submitted.

Also note that the user’s request may contain typos or other grammatical problems; these should not affect your rat-

System	Approach	On-Target	Source	Other
CUNI-VARIANT_B	Prompt Eng.	89.0%	6.1%	4.9%
CUNI-PRIMARY	Prompt Eng.	81.6%	12.7%	5.7%
ALEIX-VARIANT_B	LoRA	71.0%	22.9%	6.1%
ALEIX-PRIMARY	LoRA	70.6%	24.9%	4.5%
ALEIX-VARIANT_A	LoRA	66.5%	27.3%	6.1%
TUM-HEILBRONN-PRIMARY	LoRA	64.9%	29.8%	5.3%
PSK-VARIANT_A	QLoRA	46.9%	43.3%	9.8%
PSK-PRIMARY	QLoRA	44.5%	46.1%	9.4%
PSK-VARIANT_B	QLoRA	41.6%	47.8%	10.6%
SENTINEL-BASE	Base + Wrapper	24.1%	62.4%	13.5%
SENTINEL-PRIMARY	Base + Wrapper	24.1%	62.4%	13.5%
TEAMSPINOZA-VARIANT_V4	LoRA + Probe	23.3%	69.4%	7.3%
TEAMSPINOZA-PRIMARY	LoRA + Probe	20.4%	64.5%	15.1%
SENTINEL-RUNB	Base + Wrapper	13.5%	79.2%	7.3%
TEAMSPINOZA-VARIANT_SUM-VERIFIER3	LoRA	4.9%	89.4%	5.7%
LANIQO-VARIANT_A	Full FT + Pivot	0.4%	93.5%	6.1%
LANIQO-PRIMARY	Distill + Pivot	0.0%	94.7%	5.3%

Table 17: System behavior on the strictly cross-lingual OEG subset ($N=245$), where the prompt is in source language L_1 but explicitly requests output in L_2 ($L_1 \neq L_2$). **On-Target:** output in L_2 ; **Source:** output remained in L_1 ; **Other:** third language, English fallback, or script confusion.

System	ar	bn	bh	cs	ku	de	en	fi	ht	hi	id	it	ja	ko	mr	fa	pt	ru	sk	es	tr	vi	yo	zh	Avg
Lanigo-variant_A	16.0	14.8	17.3	14.9	11.4	17.1	24.8	12.1	20.2	21.9	22.4	21.7	30.2	27.7	12.9	20.5	23.8	16.0	14.6	26.2	16.8	23.6	6.6	27.9	19.2
CUNI-variant_B	14.6	14.9	18.0	13.8	14.0	15.3	22.0	12.2	19.9	21.1	21.0	19.8	29.0	27.7	13.3	19.2	22.4	15.5	13.7	25.5	17.1	23.4	14.8	25.9	18.9
Lanigo-primary	14.7	14.7	18.6	14.3	13.3	16.1	22.3	11.7	19.0	20.0	21.4	20.0	28.5	26.5	12.5	19.8	21.5	15.1	14.5	25.1	15.4	21.9	11.4	26.1	18.5
PSK-{primary, variant_A, variant_B}	13.8	12.6	17.4	13.7	7.1	15.7	20.7	11.1	12.9	19.9	20.4	18.8	26.8	24.9	11.7	18.1	21.6	12.7	13.4	24.1	14.9	21.7	17.3	24.0	17.3
CUNI-primary	12.7	11.7	11.4	13.3	11.7	12.6	20.4	10.2	16.6	17.5	19.1	17.7	26.7	24.3	9.6	17.5	19.8	12.6	13.0	20.0	13.3	20.1	16.2	24.6	16.3
Sentinel-base	12.5	13.1	9.9	13.1	4.9	14.6	19.0	10.7	15.2	18.9	18.1	17.7	25.6	19.4	10.7	18.0	19.9	13.1	12.9	22.8	13.5	19.6	4.6	22.7	15.4
Sentinel-primary	12.5	13.1	9.9	13.1	4.9	14.6	19.0	10.7	15.2	18.9	18.1	17.7	25.6	19.4	10.7	18.0	19.9	13.1	12.9	22.8	13.5	19.6	4.6	22.7	15.4
Sentinel-RunB	11.6	10.6	10.9	12.3	5.7	14.8	22.9	9.8	12.4	16.2	17.6	18.3	22.1	20.6	8.2	16.2	19.9	12.8	12.5	23.2	9.9	18.7	4.9	22.9	14.8
TUM-Heilbronn-primary	13.9	8.8	10.0	9.9	5.5	6.9	25.6	5.7	16.0	15.2	11.6	20.1	29.0	23.4	8.0	16.1	16.7	13.3	10.3	19.3	11.3	21.8	5.5	26.0	14.6
TeamSpinoza-Variant_sum-verifier3	10.8	9.9	9.8	11.1	4.8	12.9	18.6	6.7	10.8	14.8	16.1	15.3	24.0	20.3	8.6	15.3	18.0	12.0	10.6	21.0	11.3	18.9	5.5	22.4	13.7
Aleix-variant_A	10.7	8.1	9.9	11.8	1.8	13.4	21.6	8.0	6.8	9.7	16.4	17.6	25.0	22.0	4.5	13.7	20.0	13.4	11.4	21.9	11.8	21.0	3.1	24.4	13.7
TeamSpinoza-Variant_V4	8.5	9.9	8.7	10.4	5.2	9.0	15.3	8.2	10.7	12.6	13.6	13.7	23.1	20.6	6.1	13.0	14.4	11.2	10.3	16.6	11.3	16.0	3.6	18.9	12.1
Aleix-primary	7.2	4.3	3.1	9.7	2.6	8.3	14.6	6.3	3.8	4.4	11.2	10.1	22.6	14.8	5.0	13.1	20.8	10.2	9.0	16.2	11.6	19.1	1.5	23.9	10.6
Aleix-variant_B	11.3	5.0	13.0	8.5	3.4	6.2	20.7	3.0	6.8	11.3	5.2	5.6	29.2	18.3	6.1	13.9	20.7	8.0	6.8	10.4	11.3	17.3	2.9	4.2	10.4
TeamSpinoza-primary	5.6	6.0	5.8	6.2	3.7	6.7	13.5	4.5	7.8	8.2	9.1	11.1	15.1	13.7	4.0	8.6	11.4	6.7	6.3	12.9	6.4	10.8	3.4	14.5	8.4

Table 18: Summarization: ROUGE-L F1 per language across submitted systems. Best per column in **bold**.

ings.

F.3 Detailed Landmark Rubrics

The detailed landmark definitions presented to annotators for each of the three dimensions were as follows:

F.3.1 Instruction Following

- **Great:** The response fully adheres to all instructions that the user provided.
- **Good:** The chatbot mostly followed the instructions, conforming to the main points of the request but missing some details.
- **Flawed:** The chatbot followed only a small portion of the instructions or missed important points.
- **Poor:** The chatbot entirely disregarded the instructions; the response appears to have no relation to the user’s request.

F.3.2 Naturalness

- **Great:** The response represents fluent {{language}} text that might have been written by a native human speaker, and appropriately matches the tone and formality of the request.
- **Good:** The response has some disfluencies that are noticeable but don’t significantly hinder understanding the response, or the tone/formality is somewhat inappropriate to the request.
- **Flawed:** The response is highly disfluent. There are several grammatical errors, or some of the output is not in {{language}}. Most of the meaning can be determined, but only with conscious effort. Alternatively, the tone or formality of the response is drastically and inappropriately different from that of the request.
- **Poor:** The response is incomprehensible, or a significant amount is not in {{language}}.

Submission	LLM-as-a-verifier			Average↓
	Cmd-A+	Gemma	GLM	
CUNI-variant_B	0.709	0.794	0.711	0.738
Lanigo-primary	0.654	0.760	0.700	0.705
TeamSpinoza-primary	0.572	0.704	0.702	0.659
Lanigo-variant_A	0.579	0.726	0.662	0.656
TeamSpinoza-Variant_V4	0.570	0.681	0.640	0.630
CUNI-primary	0.672	0.578	0.513	0.588
Sentinel-primary	0.602	0.637	0.516	0.585
Sentinel-base	0.593	0.637	0.519	0.583
Sentinel-RunB	0.555	0.608	0.484	0.549
TUM-Heilbronn-primary	0.524	0.539	0.474	0.512
Aleix-variant_A	0.541	0.490	0.416	0.482
Aleix-primary	0.413	0.414	0.341	0.389
TeamSpinoza-Variant_sum-verifier3	0.388	0.367	0.252	0.336
PSK-variant_B	0.361	0.393	0.237	0.331
PSK-primary	0.356	0.390	0.236	0.328
PSK-variant_A	0.353	0.391	0.234	0.326
Aleix-variant_B	0.360	0.331	0.275	0.322

Table 19: Summarization LLM-as-a-verifier target language adherence and grammatical fluency pass rates across Command A+, Gemma, and GLM verifiers.

F.3.3 Coherence

- **Great:** The response is logically sound and appropriately structured with a clear sequence of nicely connected ideas and topics with no leaps in reasoning.
- **Good:** The response is generally well-structured and has a generally clear overall progression of ideas, but introduces a few logical gaps, or suddenly switches topics without an appropriate transition.
- **Flawed:** The response lacks an overall flow, and/or has multiple noticeable jumps between topics. It is possible to discern some relevant ideas, but the overall purpose of the response is incoherent.
- **Poor:** The response has no overall structure, is in no way logically sound, and/or can be divided into many mostly-unrelated sections. It is difficult to identify any points the text is trying to make.

Rank	System	Type	QA-Context	OEG	Sum	Mean Score
1	CUNI-variant_B	Team	82.89%	89.23%	79.39%	83.84%
2	Lanigo-primary	Team	78.78%	76.74%	76.01%	77.18%
3	CUNI-primary	Team	83.47%	83.06%	57.83%	74.79%
4	Lanigo-variant_A	Team	72.29%	65.04%	72.64%	69.99%
5	Sentinel-base	Team	61.77%	73.36%	63.74%	66.29%
6	Gemma-4-E4B-it	Base	77.18%	79.78%	36.32%	64.42%
7	TeamSpinoza-Variant_V4	Team	66.10%	58.77%	68.07%	64.31%
8	Sentinel-primary	Team	53.69%	73.23%	63.68%	63.53%
9	TeamSpinoza-primary	Team	69.64%	49.48%	70.38%	63.17%
10	TUM-Heilbronn-primary	Team	64.05%	60.43%	53.94%	59.48%
11	Sentinel-RunB	Team	53.66%	58.51%	60.75%	57.64%
12	GLM-4-9B-0414	Base	51.68%	62.45%	55.80%	56.64%
13	Tiny Aya Global	Base	54.81%	69.57%	40.99%	55.13%
14	PSK-variant_B	Team	54.59%	64.34%	39.30%	52.75%
15	PSK-variant_A	Team	54.66%	63.80%	39.13%	52.53%
16	PSK-primary	Team	54.62%	62.77%	39.02%	52.13%
17	Aleix-variant_B	Team	59.62%	62.09%	33.11%	51.60%
18	Qwen3-8B	Base	57.67%	69.02%	25.34%	50.68%
19	Aleix-variant_A	Team	47.92%	54.19%	49.04%	50.38%
20	alkhder-primary	Team	48.00%	—	—	48.00%
21	Phi-4-mini-instruct	Base	43.37%	53.23%	38.18%	44.92%
22	Llama-3.1-8B-Instruct	Base	56.54%	62.04%	16.16%	44.91%
23	Aleix-primary	Team	36.05%	56.43%	41.39%	44.62%
24	TeamSpinoza-Variant_sum-verifier3	Team	19.22%	24.77%	36.71%	26.90%

Table 20: Cross-task comparison including organizer baselines, evaluated under Gemma-4-31B-it. Baselines (italicized) have Gemma-judge scores only. Systems ranked by mean score across the three sub-tasks.