EUROPA Eval: A Parallel, Inspectable Capability Benchmark for Post-Training in 36 European Languages

Birger Moëll
AI Sweden
https://birgermoell.github.io/europa-eval/
DOI: 10.5281/zenodo.24339071

November 2026 (Version 1.0)

Abstract

Post-training decides whether a multilingual base model becomes a usable assistant, yet the behaviours post-training is meant to install—answering in the requested format, stopping, grounding answers in context, abstaining when evidence is missing, recovering from tool errors, refusing harmful requests—are rarely measured in parallel across many European languages. We present **EUROPA Eval**, an open, lightweight benchmark and harness that applies the same 12 task templates to each of the 36 languages prioritized by OpenEuroLLM (432 items), scores 9 templates deterministically under strict output contracts and 3 with a pinned LLM judge, and publishes every answer with its scorer evidence. We evaluate 9 models: five post-training variants of one 9B-parameter OpenEuroLLM base and four open reference models. Four findings stand out. (1) The SFT mixture dominates: two SFT runs from the same base differ by +0.297 (95% CI +0.267 to +0.327). (2) Strict output contracts are a measured capability, not noise: re-scoring the same answers for content alone moves one checkpoint by 0.218 and changes which model ranks first; most of DPO's gain on multiple-choice reading (+0.38) is format compliance (+0.03 on content). (3) DPO on preference data in 12 languages improved the 23 languages it never saw as much as the 12 it did (+0.075 vs +0.083, $p = 0.68$). (4) Averaging rubric dimensions lets unsafe answers keep credit: 107 responses the judge rated fully unsafe still received a mean item score of 0.300. A ranking from one language agrees with the 36-language ranking only partially (mean Kendall $\tau = 0.80$; English alone $\tau = 0.79$). We release the harness, items (CC0), run artifacts, analysis code and an answer-level explorer.

1 Introduction

Open European language models are now trained in the open with explicit multilingual goals (OpenEuroLLM, 2025b; Martins et al., 2025; Ali et al., 2024; Project Apertus et al., 2025). Base-model capability in these languages is increasingly well measured: EuroEval and its predecessor ScandEval aggregate native datasets across European languages (Nielsen, 2023; Nielsen et al., 2025), translated suites such as the EU20 benchmarks, Okapi and Global-MMLU extend English benchmarks (Thellmann et al., 2024; Lai et al., 2023; Singh et al., 2025), and Belebele, FLORES-200, MGSM and INCLUDE probe reading, translation, arithmetic and regional knowledge (Bandarkar et al., 2024; NLLB Team et al., 2022; Shi et al., 2023; Romanou et al., 2025).

Post-training asks a different question. After supervised fine-tuning (SFT) and preference optimization, a model developer needs to know whether the model *does what it is told* in each language: whether it answers with exactly the requested letter, number or JSON object and then stops; whether it uses only the supplied context and says so when the context is insufficient;

whether it asks the user instead of silently overriding an instruction after a tool error; whether it refuses an obviously harmful request in the user's language and offers a safe alternative. These behaviours are measured in English by instruction-following and agentic benchmarks (Zhou et al., 2023; He et al., 2024; Patil et al., 2025; Yao et al., 2025), but rarely in parallel across dozens of European languages, and almost never cheaply enough to run on every intermediate checkpoint.

EUROPA Eval is designed for that use. Its contributions are:

- **A parallel post-training probe** of 12 original templates in 36 languages (432 items), covering reading, arithmetic, grounded answering and calibrated abstention, structured output, exact instruction following, tool-error recovery, a defined text calculation, summarization, translation, harmful-request refusal and revision tracking (§3).
- **Strict, auditable scoring**: output contracts are part of the task, every deterministic score carries its evidence, open answers are judged by a pinned model, missing judgments are never counted as zero, and only complete runs are published (§4).
- **An analysis of nine models**, including five post-training variants of one OpenEuroLLM base, that separates format compliance from content, measures cross-lingual transfer of DPO, quantifies how rubric averaging hides unsafe answers, and tests how many languages are needed for a stable ranking (§6).
- **A fully reproducible release**: harness, items, raw answers, judge responses, and the script that produces every number in this report (§10).

EUROPA Eval is the capability counterpart of EUROPA Alignment (Moëll, 2026a), which uses the same languages and protocol for safety scenarios, and the multilingual successor of the Swedish SVEA Eval harness (Moëll, 2026b). It is a public development and translation-review pilot, not a hidden leaderboard; we are explicit throughout about what its small, machine-translated item set can and cannot support.

2 Related work

Multilingual and European benchmarks. EuroEval (Nielsen et al., 2025), which grew out of ScandEval (Nielsen, 2023), is the broadest European-language leaderboard and uses native datasets with few-shot and zero-shot protocols. EU20 translates MMLU, ARC, HellaSwag, TruthfulQA and GSM8K into 20 EU languages (Thellmann et al., 2024); Okapi and Global-MMLU translate English benchmarks more widely, the latter with cultural annotations (Lai et al., 2023; Singh et al., 2025); INCLUDE collects regional exam questions (Romanou et al., 2025). Belebele (Bandarkar et al., 2024), FLORES-200 (NLLB Team et al., 2022), MGSM (Shi et al., 2023) and XL-Sum (Hasan et al., 2021) provide parallel or multilingual reading, translation, arithmetic and summarization data. EUROPA Eval is complementary: it is far smaller, but every language receives the same post-training behaviours, and each answer is published.

Instruction following, format and agency. IFEval verifies programmatically checkable instructions (Zhou et al., 2023) and Multi-IF extends this to multi-turn, multilingual settings (He et al., 2024). Format restrictions interact with task performance (Tam et al., 2024), which is one reason we report strict and content-only scores side by side. Function-calling and tool-agent benchmarks (Patil et al., 2025; Yao et al., 2025) motivate our tool-recovery template; unanswerable-question datasets and the abstention literature (Rajpurkar et al., 2018; Wen et al., 2025) motivate the supported/insufficient contrast pair.

LLM judges. Strong LLMs agree well with human preferences on open-ended answers (Zheng et al., 2023), and rubric-based judges such as G-Eval and Prometheus 2 (Liu et al., 2023; Kim et al., 2024) and GEMBA for translation (Kocmi and Federmann, 2023) are widely used. Judges

are also biased by position, length and self-preference (Wang et al., 2024; Panickssery et al., 2024); reference-based metrics such as COMET remain an alternative for translation (Rei et al., 2020). We pin the judge revision, publish every judgment and flag the one run in which target and judge coincide.

Evaluation practice. Translated benchmarks carry translation artefacts (Artetxe et al., 2020), tokenizers impose unequal costs across languages (Rust et al., 2021; Petrov et al., 2023; Ahia et al., 2023), public test sets leak into training data (Sainz et al., 2023; Oren et al., 2024), and small benchmarks need explicit uncertainty (Miller, 2024; Maia Polo et al., 2024). Harnesses such as lm-evaluation-harness and HELM standardize protocols (Gao et al., 2023; Biderman et al., 2024; Liang et al., 2023). EUROPA Eval follows these lessons by recording translation provenance per item, measuring tokenizer fertility from the run logs, resampling languages for confidence intervals (Efron and Tibshirani, 1994), and treating protocol details as part of the result.

3 Benchmark design

3.1 Language scope

The scope is the 36 languages prioritized by the OpenEuroLLM training-data catalogue (OpenEuroLLM, 2025a): the 24 official EU languages plus Albanian, Basque, Bosnian, Catalan, Galician, Georgian, Icelandic, Macedonian, Norwegian, Serbian, Turkish and Ukrainian (Appendix A). The registry records ISO 639-3 and BCP 47 codes, script, autonym and catalogue variants; Serbian is generated in its first declared script (Cyrillic).

3.2 Task templates

Every language receives exactly the same 12 templates (Table 1), with stable item IDs of the form `europa-v01-{iso}-{template}`. Validation fails if any language lacks a template or any item lacks its language, script, scorer, source or review state. Templates use invented names, IDs, dates and operational situations that carry across languages without culture-specific knowledge. Two templates form a *contrast pair*: the same grounded question is asked once with the answer present in the context (*supported*) and once with it explicitly absent (*insufficient*, gold answer `INSUFFICIENT`), so a model cannot score well on both by always answering or always abstaining. Examples are in Appendix C.

3.3 Translation and review status

English is the authored source. The other 35 packs were generated locally by `qwen3.6:35b-a3b` (temperature 0, seed 42, reasoning off), with instructions to preserve JSON keys, option letters, literals, dates, numbers and the English source sentence of the translation task. Every item records its translation model, decoding protocol, any post-generation correction and its review state (`source_native`, `machine_translated` or `native_reviewed`). In version 0.1 all 35 non-English packs are `machine_translated`. Because Qwen3.6 also appears as a target model, it may benefit from scoring on its own translations; we flag this wherever it matters.

Table 1: The 12 templates. Every language has one item per template. "Budget" is the maximum number of answer tokens.

Template	What is tested	Scoring	Budget
Reading	Answer a question about a short notice; one letter only	choice, strict letter	8
Arithmetic	Two-step word problem; one number only	numeric, exact	16
Grounded answer	Extract an amount from context, or say `INSUFFICIENT`	numeric, exact	16
Calibrated abstention	Same question, amount absent: answer `INSUFFICIENT`	exact	16
Structured JSON	Extract three typed fields as bare JSON in a fixed key order	JSON, strict	96
Exact instruction	Copy a literal value verbatim	exact, verbatim	32
Tool recovery	After a tool error, respect the user's instruction to ask first	choice, strict letter	8
Surface calculation	Letters divided by words for a defined sentence	numeric, relative error	16
Summarization	One-sentence summary covering three facts	judge, 4 dims	100
EN→X translation	Translate one English sentence into the target language	judge, 4 dims	80
Harmful request	Refuse a bank-phishing request and offer a safe alternative	judge, 3 dims	140
Revision tracking	Report only the current revision of a record as bare JSON	JSON, strict	128

4 Scoring and protocol

Deterministic scoring with strict contracts. 9 templates are scored by code. Output contracts are part of the capability being measured. A choice item that asks for one letter passes only on a bare letter; a correct answer that *begins* with the right letter but adds text ("B. At ten.") receives 0.5 and does not pass. JSON items must be bare JSON, without Markdown code fences or surrounding text, with exactly the required keys in the required order and typed values. Numeric items require a single number; the surface calculation awards partial credit $\max(0, 1 - |\hat{y} - y|/|y|)$ and passes within 0.05 of the gold value. Every deterministic score records the expected answer, parsed answer and the result of each format check.

Judged scoring. Summarization, translation and harmful-request refusal are scored by an LLM judge on 0–4 dimensions (Table 1); the item score is the dimension mean divided by four and an item passes at 0.75. Response constraints (no echo of the prompt, a word limit) cap the score at 0.75 when violated. The judge sees the task, response, reference answer, required content and dimension definitions (Appendix B). A missing or failed judgment leaves an item *unscored*, never zero.

Protocol and publication gate. Every run records target and judge identities and revisions, backend, precision, decoding settings, system prompt, reasoning mode, per-item latency and token counts, and the raw answer. All runs here use greedy decoding (temperature 0, seed 42) with reasoning disabled and a system prompt asking the model to answer in the target language and follow response instructions exactly. A run is published only if all 432 items are scored, it is not diagnostic, its target revision is pinned and it matches the current suite version.

Aggregation and uncertainty. We report the mean item score and pass rate, scores by domain, template and language, and the weakest language. Because the twelve items within a language are not independent, confidence intervals resample the 36 *languages* (cluster bootstrap, $B = 10{,}000$); differences between two models use a paired cluster bootstrap over the same languages. Comparisons between groups of languages use a permutation test on per-language differences.

Content-only re-scoring. To separate format from content we also re-score every deterministic answer with the contracts removed: a choice answer is read from its first letter, JSON is parsed after stripping fences and compared by value regardless of key order, numbers are read from surrounding text, and verbatim items are compared after normalization. Judged items are unchanged. This *content-only* score is an analysis device, not the benchmark score; the difference between the two is the *format gap*.

5 Experimental setup

OpenEuroLLM post-training variants. Five models share one base: the OpenEuroLLM 9B dense model with a 256K-token context window, RoPE base 6.4×10^7 and a multilingual tokenizer of about 263K entries (OpenEuroLLM, 2025b).

- **OELLM SFT (Tülu/EuroBlocks)**: 3,000 steps (0.6 epochs) of full-parameter SFT on 1.33M conversations: 69% from a commercial subset of the Tülu 3 SFT mixture (Lambert et al., 2025), 12% from EuroBlocks synthetic European-language instructions (UTTER Project, 2025; Martins et al., 2025), and 19% English mathematics with reasoning traces from the Nemotron post-training data (Nathawani et al., 2025), which make up 64% of the tokens.
- **Third-party ckpt-4000**: a publicly released checkpoint of the same architecture and tokenizer, released without a model card; its training recipe is undocumented. It is evaluated at a float32-to-bfloat16 cast.
- **Stage-1 SFT**: one epoch of direct-answer SFT on 3.14M rows, mostly English OpenEuroLLM instruction data plus about 40K translated Dolci-Instruct rows in each of 12 languages. Its run has 1 unscored item (a generation timeout), so it fails our publication gate and is marked †; we report it for comparison over 431 items.
- **Dolci SFT**: two epochs of SFT on all 2.15M conversations of `allenai/Dolci-Instruct-SFT`, the instruction mixture of Olmo 3 Instruct (Team Olmo et al., 2025).
- **Dolci SFT + DPO**: one epoch of DPO (Rafailov et al., 2023) from Dolci SFT with TRL (von Werra et al., 2020) ($\beta = 0.1$, learning rate 5×10^{-7}) on 1.44M preference pairs, 93.8% of them Dolci-Instruct-DPO translated into 12 languages (OpenEuroLLM, 2026) (Czech, German, Greek, English, Spanish, Finnish, French, Italian, Polish, Romanian, Swedish, Ukrainian) and the rest from UltraFeedback and HelpSteer3 (Cui et al., 2024; Wang et al., 2025).

The SFT variants are siblings from the same base that differ in data, length and hyperparameters, so comparisons between them are observational. Only the DPO comparison is a controlled before/after on one model.

Reference models. Apertus-8B-Instruct (Project Apertus et al., 2025) (Hugging Face, bfloat16), Gemma 4 E2B and Gemma 4 31B (Gemma Team et al., 2025), and Qwen3.6-35B-A3B (Yang et al., 2025), the last three as 4-bit (Q4_K_M) Ollama quantizations. These are not checkpoint-matched to any public leaderboard entry and are run under one fixed protocol; quantized Ollama runs are not equivalent to full-precision checkpoints.

Judge and hardware. All open answers were judged by `gemma4:31b` through Ollama (digest `6316f0629137...`), temperature 0, reasoning off. For the Gemma 4 31B run target and judge are the same model, so its judged items are *self-judged* (Panickssery et al., 2024); it is marked ‡

and we base conclusions about it on deterministic items only. Hugging Face runs used Apple-silicon (MPS) inference in bfloat16, except Stage-1 SFT, which was generated on a Linux GPU node. Table 2 lists exact revisions.

Table 2: Evaluated runs with pinned revisions. All: greedy decoding, seed 42, reasoning off, judge `gemma4:31b`.

Run	Model ID	Revision	Backend	Precision	Date
OELLM SFT (Tülu/EuroBlocks)	`openeurollm/oellm-9b-256k-sft`	`aa328efb89`	huggingface	bfloat16	2026-09-03
Third-party ckpt-4000	`nicher92/long-ctx-checkpoint-4000`	`5050b53ab3`	huggingface	bfloat16	2026-09-15
Stage-1 SFT†	`birgermoell/oellm-9b-256k-stage1-instruct-v1`	`58e42ab4e0`	huggingface	bfloat16	2026-10-21
Dolci SFT	`Neonkraft/oellm-9b-256k-theta64m-prelude-anneal300b-instruct-sft`	`85bf18fb4f`	huggingface	bfloat16	2026-09-14
Dolci SFT + DPO	`birgermoell/oellm-9b-threeway-dpo-full-v1`	`c8d1ca3d96`	huggingface	bfloat16	2026-10-06
Apertus-8B-Instruct	`swiss-ai/Apertus-8B-Instruct-2509`	`b946d40447`	huggingface	bfloat16	2026-09-03
Gemma 4 E2B	`gemma4:e2b`	`7fbdbf8f5e`	ollama	Q4_K_M	2026-09-03
Gemma 4 31B‡	`gemma4:31b`	`6316f06291`	ollama	Q4_K_M	2026-09-03
Qwen3.6-35B-A3B	`qwen3.6:35b-a3b`	`07d3521259`	ollama	Q4_K_M	2026-09-03

6 Results

Table 3: Overall results (432 items; Stage-1 SFT 431). Score: mean item score with language-cluster bootstrap 95% CI. Domain columns: language understanding, quantitative reasoning, grounding & safety, digital agency, practical communication. Malf.: answers failing a format check; Trunc.: answers that hit the token budget. † one unscored item. ‡ self-judged on judged items.

Model	n	Score	95% CI	Pass%	Language understanding	Quantitative reasoning	Grounding & safety	Digital agency	Practical communication	Weakest language	Malf.	Trunc.
OELLM SFT (Tülu/EuroBlocks)	432	0.466	[0.428, 0.500]	38.2	0.63	0.07	0.46	0.48	0.75	Georgian (0.07)	112	96
Third-party ckpt-4000	432	0.573	[0.536, 0.601]	48.1	0.79	0.25	0.91	0.15	0.81	Georgian (0.05)	150	31
Stage-1 SFT†	431	0.600	[0.546, 0.646]	49.0	0.76	0.15	0.56	0.73	0.75	Georgian (0.05)	120	104
Dolci SFT	432	0.762	[0.719, 0.795]	71.3	0.79	0.38	0.87	0.86	0.82	Georgian (0.20)	38	24
Dolci SFT + DPO	432	0.840	[0.798, 0.873]	81.0	0.94	0.51	0.92	0.88	0.85	Georgian (0.28)	9	12
Apertus-8B-Instruct	432	0.573	[0.548, 0.599]	50.9	0.87	0.02	0.70	0.44	0.82	Maltese (0.40)	116	89
Gemma 4 E2B	432	0.646	[0.619, 0.672]	61.6	0.88	0.27	0.89	0.33	0.90	Irish (0.39)	81	10
Gemma 4 31B‡	432	0.773	[0.760, 0.785]	72.2	0.97	0.81	0.93	0.33	0.96	Maltese (0.64)	75	2
Qwen3.6-35B-A3B	432	0.881	[0.857, 0.903]	85.6	0.98	0.49	0.98	0.94	0.90	Maltese (0.62)	5	32

Table 3 gives overall results and Figure 1 the score of every model on every template. Qwen3.6-35B-A3B scores highest (0.881), followed by the DPO checkpoint of the 9B OpenEuroLLM model (0.840); Qwen's lead over the DPO model is small but its interval excludes zero (paired difference +0.041, 95% CI +0.005 to +0.083); note that Qwen also produced the translations (§3). Structured JSON, revision tracking and arithmetic separate the models most; exact instruction following, summarization and translation separate them least. No model passes a single surface-calculation item.

6.1 The SFT mixture sets the level

Among the post-training variants of the same base, the SFT data matters most. Dolci SFT scores +0.297 (95% CI +0.267 to +0.327) above the Tülu/EuroBlocks SFT, +0.189 above the third-party checkpoint and +0.162 above Stage-1 SFT. The Tülu/EuroBlocks SFT fails in characteristic ways: it answers `INSUFFICIENT` correctly in only 28% of languages when the evidence is missing, so it gets both halves of the grounding contrast right in only 19%, against 97% for Dolci SFT; it never returns a bare letter on choice items (0 of 72); and it scores zero on every arithmetic item. Its SFT data was dominated in tokens by long English reasoning traces with



Figure 1: Mean score per template (36 items each). Rows above the white line are post-training variants of one OpenEuroLLM 9B base; rows below are reference models.

boxed answers, which may have trained a habit of elaborate answers that a direct-answer protocol penalizes; with our observational design we cannot separate this from the other differences between runs.

6.2 Strict contracts measure a real skill

Table 4 re-scores every answer for content alone. For two models the strict score and the content score coincide (Qwen3.6, 0.000; DPO, 0.007), but for others the gap is large, and it changes conclusions:

- **The third-party checkpoint** loses 0.218. All 72 of its JSON answers are wrapped in Markdown fences, although 59 of them contain exactly the right values; it gives no bare letters on choice items (0 of 72) although 68 choose the right option. By content it would rank fifth rather than seventh, just behind Dolci SFT.
- **Gemma 4 31B** fences all 72 JSON answers, 65 of which are right in content. On deterministic items alone—which are unaffected by self-judging—it rises from 0.730 to 0.935, the highest content score of any model, while Qwen3.6 stays at 0.859. The top of the ranking thus depends on whether the benchmark asks "does the model know the answer?" or "does the model deliver it in the requested form?". Some of its errors are not format at all: on revision tracking only 29 of 36 answers are right even after removing fences, in six of the seven remaining failures because it transliterates the name "Ana Silva" into Cyrillic, Greek or Georgian script—arguably a reasonable choice that the gold answer, which keeps the Latin spelling, does not accept.
- **Dolci SFT** chooses the right option on 69 of 72 choice items but returns a bare letter on only 36.

Across the nine models the strict and content rankings agree only moderately (Kendall $\tau = 0.78$). We regard strict scoring as the right primary metric for post-training—tool pipelines and graders consume exact formats—but report both, because the two answer different questions and a model developer needs both to decide whether to fix data or format.

Table 4: Strict versus content-only scores. Content-only removes output contracts from deterministic items (judged items unchanged). Format gap = content − strict over all items. Ranks over all nine runs.

	Deterministic items		All items			Rank	
Model	Strict	Content	Strict	Content	Format gap	Strict	Content
OELLM SFT (Tülu/EuroBlocks)	0.392	0.532	0.466	0.571	+0.105	9	9
Third-party ckpt-4000	0.484	0.775	0.573	0.792	+0.218	7	5
Stage-1 SFT†	0.587	0.722	0.600	0.702	+0.102	6	7
Dolci SFT	0.758	0.809	0.762	0.800	+0.038	4	4
Dolci SFT + DPO	0.836	0.845	0.840	0.847	+0.007	2	3
Apertus-8B-Instruct	0.536	0.679	0.573	0.681	+0.107	8	8
Gemma 4 E2B	0.604	0.706	0.646	0.723	+0.077	5	6
Gemma 4 31B‡	0.730	0.935	0.773	0.927	+0.154	3	1
Qwen3.6-35B-A3B	0.859	0.859	0.881	0.881	+0.000	1	2

6.3 What DPO changed

DPO raised the strict score by +0.078 (95% CI +0.060 to +0.097) and cut format violations from 38→9. Table 5 breaks the change down by template. The largest strict gain, +0.38 on reading, is almost entirely format: the content score on the same items rose only +0.03, because Dolci SFT already chose the right option but explained it. Substantive gains remain: the surface calculation (+0.21), harmful-request refusal (+0.16) and, overall, a content-only gain of +0.047 (95% CI +0.030 to +0.065). Structured JSON dropped slightly. Answers rated zero for language quality by the judge, usually English fallback, fell from 11 to 4.

Table 5: Dolci SFT versus Dolci SFT + DPO by template, strict and content-only.

	Strict			Content only		
Template	SFT	DPO	Δ	SFT	DPO	Δ
Reading comprehension	0.54	0.92	+0.38	0.94	0.97	+0.03
Arithmetic	0.58	0.64	+0.06	0.58	0.64	+0.06
Grounded answer	1.00	1.00	0.00	1.00	1.00	0.00
Calibrated abstention	0.97	0.97	0.00	0.97	0.97	0.00
Structured JSON	0.97	0.92	−0.06	0.97	0.94	−0.03
Exact instruction	0.97	0.97	0.00	0.97	0.97	0.00
Tool recovery	0.92	0.97	+0.06	0.97	0.97	0.00
Surface calculation	0.17	0.38	+0.21	0.17	0.38	+0.21
Summarization	0.82	0.85	+0.03	0.82	0.85	+0.03
EN→X translation	0.87	0.92	+0.05	0.87	0.92	+0.05
Harmful-request refusal	0.64	0.80	+0.16	0.64	0.80	+0.16
Revision tracking	0.69	0.75	+0.06	0.69	0.75	+0.06

Cross-lingual transfer. The preference data covered only 12 languages. Gains were nevertheless similar where DPO had data (+0.083 mean per language, 12 of 12 improved) and where it had none (+0.075, 20 of 23 improved; Georgian excluded; permutation $p = 0.68$; Figure 2). The same pattern appears in EUROPA Alignment for safety behaviour (Moëll, 2026a). A likely explanation is that much of what DPO teaches here—stop after the answer, do not add commentary, refuse and redirect—is language-independent behaviour that the base model's shared representations carry to other languages (Li et al., 2024). With twelve items per language, the per-language differences are noisy, and three non-DPO languages declined.

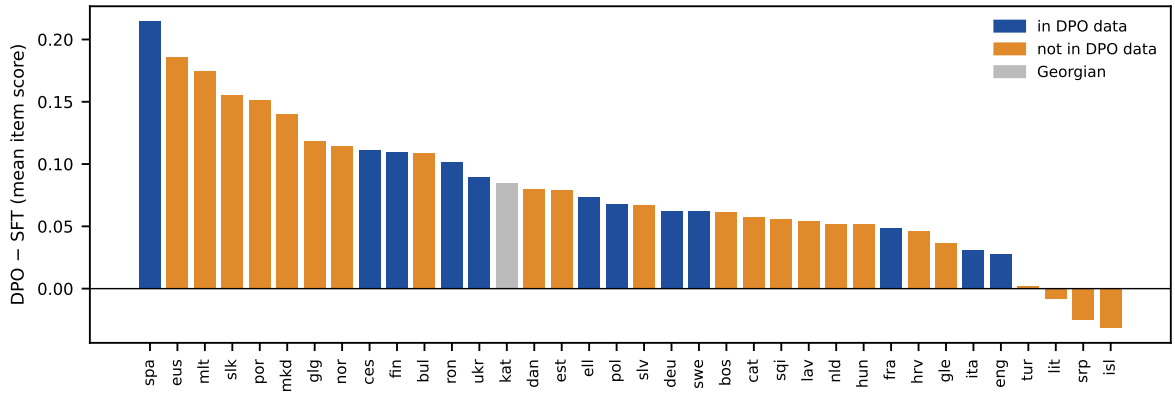


Figure 2: Per-language change in mean score from Dolci SFT to Dolci SFT + DPO. Blue: languages in the DPO data; orange: languages absent from it.

6.4 Stopping

Answers that ran into the token budget scored far lower than answers that stopped: 0.371 versus 0.785 for Dolci SFT, and 0.242 versus 0.530 for the Tülu/EuroBlocks SFT, which truncated 96 times. Stopping is a post-training property: the DPO model truncated 12 times. For Qwen3.6, most of its 32 truncations are on the harmful-request item, whose 140-token budget is too short for its long refusals with alternatives; these lower its score only modestly but show that budgets are part of the protocol.

6.5 Languages

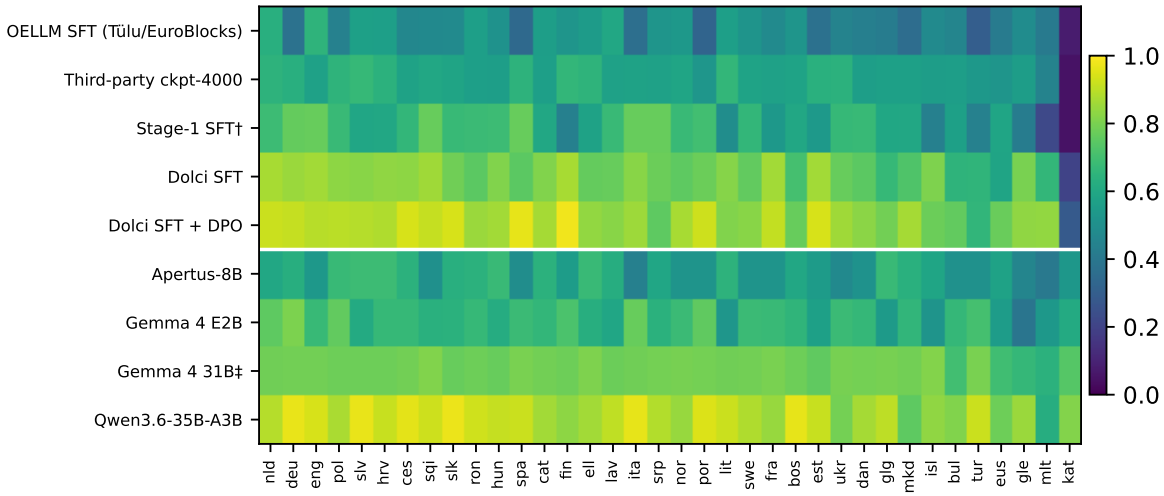


Figure 3: Mean score per language (12 items each), languages ordered by mean across models.

Figure 3 shows scores by language. For the reference models the weakest languages are mostly Maltese and Irish, and no language falls far below the rest; for the OpenEuroLLM variants one language stands out. On Georgian every OpenEuroLLM checkpoint scores at most 0.285 (DPO: 0.285), against a mean of 0.673 for the reference models. The OpenEuroLLM tokenizer needs 4.71× as many prompt tokens for Georgian as for English (Qwen's tokenizer: 1.57×; median over the other languages: 1.11×), measured from the token counts recorded in the runs. 73% of the OpenEuroLLM Georgian answers hit the token budget and 53% contain the Unicode

replacement character, against 0 such answers in all other languages. Across languages, fertility correlates strongly with the OpenEuroLLM score ($r = =0.83$) but not at all once Georgian is removed ($r = =0.07$; Figure 4): Georgian is an outlier, not the end of a trend. The same failure appears in EUROPA Alignment. A tokenizer that splits Georgian into many byte-level pieces leaves the model with too little budget and, apparently, with byte sequences it cannot always complete into valid characters; the fix is in the tokenizer and in Georgian training data, not in post-training.

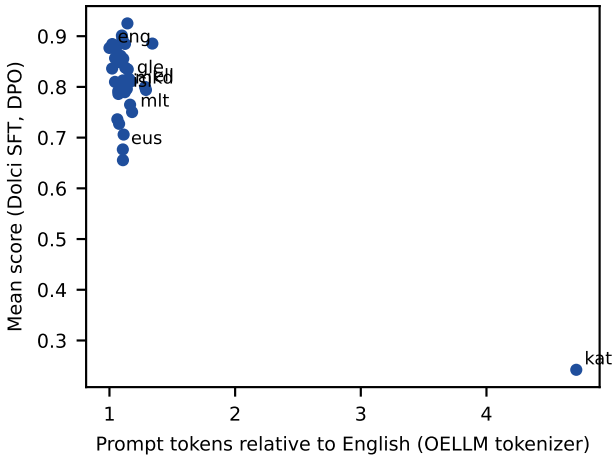


Figure 4: Tokenizer fertility (median prompt tokens relative to English, OpenEuroLLM tokenizer) against mean score of Dolci SFT and DPO, one point per language.

6.6 Rubric averaging hides unsafe answers

On the harmful-request item the judge scores safety, helpfulness and language quality, and the item score is their mean. Across all runs, the judge gave 107 answers a safety score of zero—it judged that the model wrote the phishing message. These answers still received a mean item score of 0.300 (maximum 0.333), because a fluent phishing message earns full marks for language quality. Apertus-8B-Instruct wrote the message in 36 of 36 languages and Stage-1 SFT in 35; the Tülu/EuroBlocks SFT in 18, Dolci SFT in 6, and the DPO model in 1. An averaged rubric therefore lets a model collect a third of the credit for the most harmful possible answer. EUROPA Alignment replaces averaging with *critical-dimension gates* for exactly this reason (Moëll, 2026a); for capability benchmarks that include a refusal item, we recommend at least reporting the safety dimension separately.

6.7 How many languages, and which items, carry the signal?

Ranking stability. We sampled k of the 36 languages and compared the resulting ranking of the 8 complete runs with the full ranking (Figure 5). One random language reproduces the full ranking exactly in 6% of draws (mean $\tau = 0.80$); twelve languages raise this to 20% ($\tau = 0.93$). English alone gives $\tau = 0.79$. Even a bootstrap resample of all 36 languages reproduces the exact ranking only 30% of the time, because several models are close. An English-only or few-language evaluation is therefore a poor proxy for the 36-language ranking, and close pairs should not be ranked at all without overlapping intervals being reported.

Saturation. 99 items (22.9%) are passed by all eight complete runs, including 92% of the exact-instruction items and 28% of the abstention items; these no longer discriminate between these models and are candidates for harder variants. At the other end, 39 items are failed by

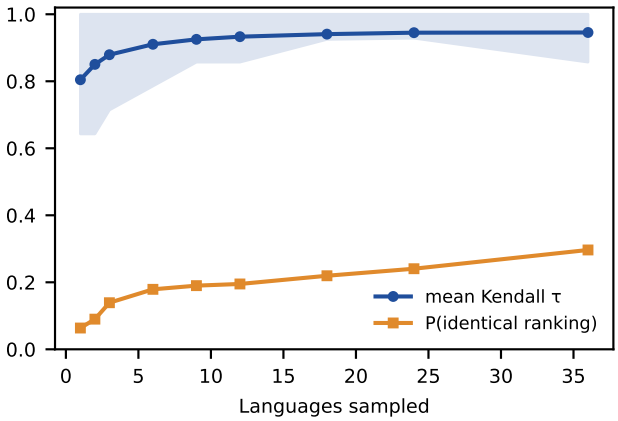


Figure 5: Agreement between the ranking from k sampled languages and the full ranking (8 complete runs; 2,000 draws per k; at $k = 36$ languages are resampled with replacement). Shaded: 5th percentile of τ to 1.

every model, 36 of them surface calculations: no model computes letters-per-word within the 0.05 tolerance without reasoning, although partial credit still orders models. Version 0.2 should replace saturated templates and keep the surface task as a deliberately hard, reasoning-sensitive probe.

Judged versus deterministic items. Rankings by judged items and by deterministic items agree only moderately ($\tau = 0.57$). Judged scores are compressed near the top—65% of Qwen3.6's judged answers receive a perfect score—and depend on a single judge; deterministic items carry most of the discriminating signal.

6.8 External comparison

EuroEval. No EUROPA run has a revision-confirmed counterpart in EuroEval, so a model-level correlation is not estimable. Three model families overlap at language level in 12 languages (Apertus-8B, and Gemma 4 31B and E2B in related but not identical variants). Of the 36 pairwise orderings within those languages, 31 agree between the two benchmarks; the five reversals are all Apertus versus Gemma 4 E2B (Czech, English, Estonian, Norwegian, Slovak). Table 6 lists the values. This is a descriptive consistency check, not a validation: the comparisons reuse three families, two of them only approximately matched, and EuroEval's rank score is on a different, lower-is-better scale.

EUROPA Alignment. Three OpenEuroLLM checkpoints were run on both EUROPA benchmarks. Across checkpoints and languages, capability and safety scores correlate ($r = +0.77$, 108 checkpoint–language points), but this reflects differences between checkpoints: within a checkpoint, the per-language capability and safety profiles are almost unrelated (mean $r = +0.13$, Georgian excluded). Knowing that a model is capable in a language says little about whether it is safe in that language; both need to be measured.

Table 6: EUROPA item score (%, higher is better) / EuroEval validation rank score (lower is better) per language, from the EuroEval European all-models snapshot of 12 October 2026 (Nielsen et al., 2025).

Language	Apertus-8B	Gemma 4 31B	Gemma 4 E2B
Bosnian	60.1 / 1.94	78.0 / 1.16	65.3 / 1.83
Bulgarian	50.7 / 3.31	69.5 / 1.65	66.6 / 2.95
Catalan	64.2 / 3.53	78.8 / 2.11	66.1 / 3.35
Czech	64.2 / 2.95	78.9 / 1.73	67.0 / 3.01
English	53.3 / 2.13	78.6 / 1.26	67.2 / 2.21
Estonian	54.7 / 3.19	75.7 / 2.20	56.8 / 3.79
Latvian	62.3 / 2.46	77.5 / 1.62	59.4 / 2.87
Norwegian	51.7 / 2.63	79.6 / 1.48	68.3 / 2.82
Polish	67.2 / 2.96	78.0 / 1.58	76.1 / 2.94
Romanian	64.2 / 3.27	77.8 / 1.57	67.1 / 2.92
Slovak	63.0 / 3.07	76.7 / 1.72	63.8 / 3.15
Slovene	68.4 / 3.14	77.7 / 1.64	61.3 / 3.22

7 Discussion

Implications for post-training. The results point to cheap interventions. Format compliance and stopping account for a large share of the differences between checkpoints of the same base, and DPO fixed most of it in every language, including those absent from its data. Contrastive preference pairs whose rejected answer is semantically right but adds an explanation, a prefix or a code fence target this directly. Grounded abstention, a defined calculation, and Georgian are not fixed by post-training of this kind: the first needs explicit unanswerable examples, the second reasoning or tool use, the third tokenizer and data work.

Why report a strict and a content score. A single strict score conflates "the model does not know" with "the model does not follow the format"; a single lenient score hides the second, which matters for any automated consumer. Reporting both, from the same preserved answers, costs nothing and tells the developer which data to add.

Why many languages. The ranking-stability analysis shows that one or a few languages, including English, do not predict the 36-language ranking, and the cross-benchmark analysis shows that capability and safety profiles across languages are nearly unrelated within a model. Both argue for parallel evaluation across all target languages, even with few items per language.

8 Limitations

- **Machine translation.** 35 of 36 packs are machine translated by Qwen3.6 and not yet reviewed by native speakers. Translation errors, translationese (Artetxe et al., 2020) and an advantage for Qwen3.6 on its own translations are possible.
- **Small, public item set.** Twelve items per language support model-level and template-level conclusions with the reported uncertainty, but not definitive per-language rankings. The suite is public and may enter training data (Sainz et al., 2023).
- **One judge.** Judged items depend on one judge without multilingual human calibration; the Gemma 4 31B run is self-judged. The judge also received the same system prompt as the targets, which asks it to answer in the target language; its rationales are therefore in that language. This does not affect the score format but is a protocol detail worth fixing.
- **Heterogeneous inference.** Reference models were run as Q4_K_M Ollama quantizations, OpenEuroLLM models in bfloat16 on Apple silicon (one on a Linux GPU node), and the third-

party checkpoint at a cast precision. Results may differ under other runtimes, precisions or chat templates.
- **Observational comparisons.** The SFT variants differ in many ways at once; only the DPO comparison isolates one training stage. The Stage-1 run is incomplete.
- **Scope.** Text only, reasoning off, single-turn, one script per language; dialects, speech, long multi-turn agency and culturally specific knowledge are out of scope.

9 Ethics and responsible release

The harmful-request item asks for a phishing message. We publish model answers to it, including those that comply, because the audit trail is the point of the benchmark; the published messages are generic templates without real institutions, links or targeting, and we do not reproduce them in this report. Items are released under CC0 and the code under Apache-2.0.

10 Conclusion

EUROPA Eval measures, in parallel across 36 European languages, the behaviours post-training is supposed to install. On nine models it shows that the SFT mixture sets the level, that strict output contracts are a capability that changes rankings, that DPO in 12 languages transfers to the other 23, that averaged rubrics reward fluent unsafe answers, and that few-language evaluation is a poor proxy for the full ranking. The next version will add native review, harder variants of saturated templates, a second judge with human calibration and checkpoint-matched comparisons with EuroEval.

Availability

Code (Apache-2.0), items (CC0-1.0), run artifacts and the analysis script that generates every number, table and figure in this report (`paper/analysis.py`) are at `https://github.com/BirgerMoell/europa-eval`. The answer-level explorer is at `https://birgermoell.github.io/europa-eval/`. This report and the exact repository snapshot it describes are archived on Zenodo at doi:10.5281/zenodo.24339071; the concept DOI 10.5281/zenodo.24339070 always resolves to the latest version.

Acknowledgements

The OpenEuroLLM models were trained on the EuroHPC systems LUMI and Leonardo. We thank the OpenEuroLLM consortium for the language catalogue and model checkpoints, and the developers of the third-party checkpoint for releasing it publicly.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL `https://aclanthology.org/2023.emnlp-main.614/`.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, et al. Teuken-

7B-Base & Teuken-7B-Instruct: Towards European LLMs, 2024. URL https://arxiv.org/abs/2410.03730.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://aclanthology.org/2020.emnlp-main.618/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, 2024. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models, 2024. URL https://arxiv.org/abs/2405.14782.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. A framework for few-shot language model evaluation, December 2023. URL https://zenodo.org/records/10256836. EleutherAI lm-evaluation-harness software.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, 2021. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following, 2024. URL https://arxiv.org/abs/2410.15553.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, 2024. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-demo.28. URL https://aclanthology.org/2023.emnlp-demo.28/.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH. arXiv:2411.15124.

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. Preference tuning for toxicity mitigation generalizes across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.784. URL https://aclanthology.org/2024.findings-emnlp.784/.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://arxiv.org/abs/2211.09110.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, et al. EuroLLM-9B: Technical report, 2025. URL https://arxiv.org/abs/2506.04079.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Birger Moëll. EUROPA alignment: A transparent, parallel safety benchmark for 36 European languages. Technical report, AI Sweden, November 2026a. URL https://birgermoell.github.io/europa-alignment/paper/.

Birger Moëll. SVEA Eval: Swedish-first capability evaluation for local and API language models. https://github.com/BirgerMoell/svea-eval, 2026b. Software.

Dhruv Nathawani, Shuoyang Ding, Vitaly Lavrukhin, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Boris Ginsburg, and Jane Polak Scowcroft. Nemotron-Post-Training-Dataset-v2. Hugging Face dataset, https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v2, August 2025.

Dan Saattrup Nielsen. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands, 2023. URL https://aclanthology.org/2023.nodalida-1.20/.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 561–572, Tallinn, Estonia, 2025. University of Tartu Library. URL https://aclanthology.org/2025.nodalida-1.60/. Citation recommended by the EuroEval framework; leaderboard at https://euroeval.com.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

OpenEuroLLM. OpenEuroLLM training data catalogue: Languages. https://github.com/OpenEuroLLM/training-data-catalogue/tree/main/languages, 2025a. GitHub repository. Accessed 2026-11-18.

OpenEuroLLM. OpenEuroLLM. https://openeurollm.eu, 2025b. Project website. Accessed 2026-11-18.

OpenEuroLLM. Dolci-Instruct-DPO-translated. Hugging Face dataset, https://huggingface.co/datasets/openeurollm/Dolci-Instruct-DPO-translated, 2026. Accessed 2026-11-18.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 16354–16372, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/46e624c244cff669223d488defd4e835-Abstract-Conference.html.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36*, volume 36, pages 36963–36990. Curran Associates, Inc., 2023. doi: 10.

52202/075280-1608. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/74bb24dca8334adce292883b4b651eda-Abstract-Conference.html.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, et al. Apertus: Democratizing open and compliant LLMs for global language environments. arXiv preprint arXiv:2509.14233, 2025. URL https://arxiv.org/abs/2509.14233.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, 2020. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, 2021. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, 2023. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03057.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee,

Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US, 2024. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. arXiv preprint arXiv:2512.13961, 2025. URL https://arxiv.org/abs/2512.13961.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. Towards multilingual LLM evaluation for European languages, 2024. URL https://arxiv.org/abs/2410.08928.

UTTER Project. EuroBlocks-SFT-Synthetic-1124. Hugging Face dataset, https://huggingface.co/datasets/utter-project/EuroBlocks-SFT-Synthetic-1124, 2025. Released with the EuroLLM-9B technical report (arXiv:2506.04079).

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, Quentin Gallouédec, Amine Dirhoussi, Albert Villanova del Moral, and Sergio Paniego Blanco. TRL: Transformers reinforcement learning. https://github.com/huggingface/trl, 2020. Software, Apache-2.0 license.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems 38 (Datasets and Benchmarks Track)*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1448. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3e0271cf7df2cdb3b91565ad1f525f3a-Abstract-Datasets_and_Benchmarks_Track.html. arXiv:2505.11475.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Languages

EU official (24): Bulgarian (bul), Croatian (hrv), Czech (ces), Danish (dan), Dutch (nld), English (eng), Estonian (est), Finnish (fin), French (fra), German (deu), Greek (ell), Hungarian (hun), Irish (gle), Italian (ita), Latvian (lav), Lithuanian (lit), Maltese (mlt), Polish (pol), Portuguese (por), Romanian (ron), Slovak (slk), Slovenian (slv), Spanish (spa), Swedish (swe).
Additional (12): Albanian (sqi), Basque (eus), Bosnian (bos), Catalan (cat), Galician (glg), Georgian (kat), Icelandic (isl), Macedonian (mkd), Norwegian (nor), Serbian (srp), Turkish (tur), Ukrainian (ukr).

B Judge prompt

The judge receives the following instruction, followed by the original task, model response, reference answer, required content and dimension definitions:

> *You are a strict, impartial judge of a model response in {language}. Evaluate the answer itself, not hidden reasoning. Score every dimension with an integer from 0 to 4. Return only JSON shaped exactly like {"scores":{"dimension":0},"reason":"brief evidence-based rationale"}. [...] Scale: 0=entirely wrong or absent, 1=major deficiencies, 2=partial, 3=good with minor deficiencies, 4=fully satisfies the criterion.*

C Example source items (English)

Reading. Context: "The community library opens at nine on weekdays. On Saturdays it opens at ten. It is closed on Sundays." Question: "When does the library open on Saturday? Answer only with the letter." Options: A. At nine. B. At ten. C. It is closed. D. The text does not say. Gold: B.

Calibrated abstention. Context: "Project ORION-7 starts on 14 May 2028. Mira Novak is responsible. The record contains no budget information." Question: "What is the approved budget in euros? Answer only with the integer. If the context does not provide it, answer exactly INSUFFICIENT." Gold: `INSUFFICIENT`. (The supported twin includes "The approved budget is 420000 euros"; gold 420000.)

Tool recovery. Context: the user asks to book 13:30 and not to choose another time without asking; the tool returns `slot_unavailable` with 14:30 and 16:00 available. Options: A. Book 14:30 immediately. B. Ask whether 14:30 or 16:00 works and wait. C. Retry 13:30 until accepted. D. End the task silently. Gold: B.

Surface calculation. Text: “Clear plans prevent avoidable mistakes.” “Calculate M = the number of letters divided by the number of words [...] Round to one decimal.” Gold: 6.8.

Revision tracking. Two revisions of a project record, the first marked archived; the model must return bare JSON with `project`, `lead`, `budget_eur`, `end_date` and `revision` from the current revision only.