Overview of the CLEF 2026 JOKER Task 4: Humour Generation

Generate Humorous Text from a Given Context in English, French, and Spanish

Igor Kuzmin[1,2,*], Anne-Gwenn Bosser[3], Jaap Kamps[4] and Liana Ermakova[5]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Barcelona Supercomputing Center, Barcelona, Spain
[3]Bretagne INP – ENIB, Lab-STICC CNRS UMR 6285, Brest, France
[4]University of Amsterdam, Amsterdam, The Netherlands
[5]Université de Bretagne Occidentale, HCTI, Brest, France

Abstract

This paper presents an extended overview of Task 4 (Humour Generation) of the CLEF 2026 JOKER lab, a new shared task that assesses the generative capabilities of LLM-based systems for short, constrained humorous texts. Given a structured *brief* – a target pun word and two distinct senses that the wordplay must activate – systems must generate humorous text from a given context in English, French, and Spanish. The task briefs are composed from previously curated wordplay corpora (PunEval / ExPUNations, the JOKER bilingual annotation campaign, and the JOKER 2026 French generation corpus), yielding 5,335 training and 457 test briefs across the three language tracks. Rankings are produced on the Joker Lab Arena through an Arena-style pairwise protocol: a task-specific constraint-aware judge prompt, a 3-judge cross-provider LLM ensemble with majority aggregation, and a Bradley-Terry leaderboard fitted on a sparse sample of $\approx 40{,}000$ head-to-head battles over 365 distinct submitted runs (98 English, 134 Spanish, 133 French) from four teams. We provide the task, data, submission format, and evaluation methodology descriptions. We report the official per-track results together with constraint-compliance and judge-agreement analyses.

Keywords

Computational humour, humour generation, pun generation, wordplay, LLM-as-a-judge, Bradley-Terry, evaluation arena, CLEF, JOKER

1. Introduction

Humour is a fundamental aspect of human communication that relies on creativity, cultural knowledge, and linguistic ambiguity. Its computational treatment has therefore become an important research area spanning information retrieval, natural language processing, and computational creativity. Despite recent advances in large language models, these tasks remain difficult because humour often depends on implicit knowledge, double meanings, and language-specific phenomena. The JOKER lab aims to benchmark linguistic, cultural, and creative capabilities of modern AI systems [1].

To assess the generative capabilities of LLM-based systems for constrained humorous texts we introduced Task 4 (Humour Generation). With the field advancing quickly, a gap remains in the creative production required for humour generation. While our analyses of previous years' results show that the strongest systems can excel at literal translation, humorous text often demands substantial creative adaptation: word-for-word generations rarely preserve wordplay, double meanings, or cultural allusions. We therefore aim to evaluate LLM-based generative systems not only on fidelity but also on their capacity to adapt or re-create content so that the target text conveys comparable intent and effect. This motivates a new evaluation task in which models must produce humour across languages under

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
igorkuz.tech@gmail.com (I. Kuzmin); liana.ermakova@univ-brest.fr (L. Ermakova)
https://www.joker-project.com (L. Ermakova)
0009-0001-1513-1834 (I. Kuzmin); 0000-0002-0442-2660 (A. Bosser); 0000-0002-6614-0087 (J. Kamps); 0000-0002-7598-7474 (L. Ermakova)

clear constraints, with assessment centred on semantic consistency, humorous effect, and justified adaptation rather than surface overlap.

This paper is the extended overview of Task 4. For details on other JOKER-2026 tasks, we refer the reader to their respective overview papers [2, 3, 4]. This paper complements the lab-level overview [1] with the state-of-the-art review of humour generation (Section 2), the task data with worked input examples (Section 3.1, Section B), the evaluation methodology (Section 4), and extended results including constraint-compliance and judge-agreement analyses along with the complete per-run leaderboards (Section 6, Section A).

2. State of the Art in Humour Generation

Humour generation is an important task for helping with humour translation (in context generation of jokes sharing the intent and effect of the original language) as well as conversational agents [5]. The current surge in LLM-powered conversational agents has long been anticipated and studied in the field of human-computer interaction. For instance, Humour has received substantial attention in HRI research as a tool for boosting social presence, likeability, and user engagement. Oliveira et al.'s systematic review [6] finds that humour generally has a favorable impact on how users perceive robots and rate their interaction quality. If context and style of machine-uttered humour shape different aspects of the user experience as tend to indicate some preliminary results [7], humour generation under constraints may be key to the safe development of humour equipped conversational agents.

Computational humour generation predates neural language models by decades however. Early systems were template- and schema-based: Petrović and Matthews [8] presented the first fully unsupervised joke generator, filling the "I like my X like I like my Y, Z" template from large-scale co-occurrence statistics. Valitutti and al. [9] used lexical constraints to generate adult humour by substituting one word in a text. Hong and Ong [10] automatically extracted humorous templates which were then used for pun generation. The neural era brought pun-specific architectures: Yu et al. [11] generated homographic puns from a conditional language model with a constrained-decoding algorithm that jointly attends to two senses of the target word, without training on pun data but were hindered by the lack of a sizable pun corpus; He et al. [12] operationalised the incongruity theory as a *local-global surprisal* principle, retrieving and editing sentences so that a homophone substitution is locally unexpected but globally coherent. Subsequent work incorporated humour theories more explicitly: AmbiPun [13] showed that ambiguity introduced through *context* words (rather than the pun word itself) yields funnier puns, and context-situated pun generation [14] conditioned generation on user-provided context words with sense selection, releasing 4.5k generations with human annotations.

When source information is withheld from evaluators, humour produced by LLMs tends to receive quality ratings similar to those given to human-created jokes [15]. Systems like GPT-5o, for example, have proven capable of crafting effective jokes and using humour constructively to navigate difficult social exchanges, occasionally exceeding human ability in the latter respect [16].

The arrival of instruction-tuned LLMs have thus shifted the question from *whether* systems can produce well-formed jokes to whether they can produce *novel* ones. In 2023, Jentzsch and Kersting [17] famously found that over 90% of 1,008 ChatGPT-generated jokes were the same 25 memorised jokes, highlighting a severe diversity problem in unconstrained humour generation. Proposed remedies include preference-based and theory-guided steering: the Creative Leap-of-Thought paradigm [18] tunes LLMs for associative "outside-the-box" humour on the Oogiri game, and HumorGen [19] distils an ensemble of comedic personas – a *Cognitive Synergy* approach grounded in psychological theories of humour – into open-weight models. The latter framework was adapted by the SaLT team for their winning Task 4 submissions (Section 5).

Shared tasks have recently begun to benchmark generation directly. SemEval-2026 Task 1 (MWAHAHA: *Models Write Automatic Humor And Humans Annotate*) [20] asks systems to generate jokes in English, Spanish, and Chinese under lexical or topical constraints – including two specified rare words, or relating the joke to a news headline – and evaluates them through human pairwise battles on an

Table 1
Task 4 dataset composition per track. Test rows are held out of training, and span source groups disjoint from train (see text).

Track		Train	Test
EN	Humour Generation (generate humorous text in English)	3,344	150
ES	Humour Generation (generate humorous text in Spanish)	392	151
FR	Humour Generation (generate humorous text in French)	1,599	156

Elo-style leaderboard inspired by Chatbot Arena. Its strongest systems relied on generate-many / select-best strategies with learned preference models over human pairwise judgments [21] or humour-policy planning [22].

Despite the shared arena-style ranking machinery, JOKER Task 4 poses a different generation problem. MWAHAHA's constraints are *surface-level* – they restrict which words may appear or which topic the joke must address, but any comedic mechanism satisfies them. JOKER Task 4 instead asks systems to *re-create wordplay from a decoupled double meaning*: each brief is the sense decomposition of an existing, human-validated pun, and a valid generation must reassemble a text in which *both* senses are simultaneously live, such that removing either breaks the joke. The constraint thus prescribes the ambiguity mechanism itself, not merely the vocabulary; the evaluation differs accordingly, relying on a constraint-aware LLM-judge ensemble that checks the both-senses condition explicitly (Section 4), where MWAHAHA ranks purely by human preference.

3. Task Description

Concretely, each instance presents the system with a structured *brief*: a target `pun_word` and two distinct senses (`sense_a`, `sense_b`) that the wordplay must activate, sometimes accompanied by supporting keywords or an annotator hint. A valid generation should (i) use the given pun word or a clear morphological variant, (ii) activate both senses so that removing either breaks the joke, (iii) read naturally in the target language, and (iv) be a short, self-contained text. The task is offered as three independent language tracks: English, Spanish, and French. Participants may enter any one, two, or all three; each track runs as a separate Codabench competition.[1]

Participants are free to use any model, system, or prompting strategy to produce the generations – closed-source LLMs, open-source models, fine-tuned models, retrieval-augmented pipelines, rule-based generators, or human-in-the-loop systems. Manual generations are accepted but must be flagged (`manual=1`) and are reported as a separate cohort.

3.1. Data

Rather than synthesising new prompts from scratch, the 2026 release composes its briefs from previously curated wordplay corpora distributed across the three tracks, preserving annotator-validated sense decompositions wherever possible. The composition by track is summarised in Table 1.

The **English** track merges two complementary subsources. PunEval [23] – itself a normalised consolidation of the SemEval-2017 Task 7 dataset [24] cross-referenced with ExPUNations annotations [25] – supplies 1,457 heterographic and homographic puns with WordNet sense keys, crowd-sourced keywords, and free-text explanations. We augment this with 2,037 wordplay instances drawn from the JOKER corpus and annotated with wordplay location and interpretation [26, 27, 28, 29, 30]. The **Spanish** track is derived from the JOKER-2023 bilingual annotation campaign [31, 28, 29, 30], which produced multiple Spanish renderings per English source enriched by pun location and interpretation annotation. We retain only translations the annotators flagged as both *valid* and *best in their group*, giving 543 canonical

[1]EN: https://www.codabench.org/competitions/13879/, ES: https://www.codabench.org/competitions/15742/, FR: https://www.codabench.org/competitions/15743/.

entries across 294 source jokes. The **French** track uses the JOKER French corpus (1,755 raw rows) built on the previous editions of the JOKER track [26, 27, 28, 29, 30]: Tom Swifty homophones and general puns with French pun-location and wordplay-interpretation annotations. We apply an annotation filtering, yielding 1,755 retained instances across 881 unique English sources.

Each track is split into train and test (Table 1). Test set sizes are deliberately capped near 150 per track to keep the head-to-head Arena pool within the LLM-judge budget over the competition window.

3.2. Formats

All files are UTF-8 JSON arrays; the three tracks share one schema, and fields that do not apply to a given source are simply absent. The input fields every record carries are `id`, `lang`, `pun_word`, `sense_a`, `sense_b` (and `keywords` where available).

The train files additionally provide `reference_joke` (the gold/human joke for the brief – for inspection or few-shot use only, never released at test time), `reference_en` for the translated tracks, and optional auxiliary fields: `hint` (annotator's explanation of how the pun works), `funniness` (0–5 annotator rating), and `style_tags` (free-text style descriptors).

Test files keep only the input metadata – `hint`, `reference_joke`, and annotator-quality fields are stripped. The provenance label (`source`) and the inconsistent annotator `pun_type` field are likewise dropped from the participant release to avoid per-source gaming and label noise. Worked train and test examples are reproduced in Section B; submissions are accepted on Codabench as a ZIP containing a single `prediction.json` with one {`run_id`, `manual`, `id`, `generation`} object per test id.

4. Evaluation

We adopt an Arena-style [32, 33] protocol. Codabench itself runs only a lightweight gate over each submission – format validation followed by a corpus-BLEU sanity check against organiser references – which is sufficient to filter malformed or off-language submissions but does not stand in for a humour score. The actual ranking is produced on the Joker Lab Arena: for each test brief, pairs of competing generations are judged through an LLM-ensemble pairwise protocol, and pairwise outcomes feed a Bradley-Terry leaderboard [32] with bootstrap confidence intervals, computed independently per language track.

4.1. A constraint-aware judge prompt

Because the task is constrained pun generation rather than open-ended question answering, we replace the generic MT-Bench `pair-v2` judge prompt [33] with a task-specific variant `pair-pun-v1`. The judge is shown the structured brief that the systems received (pun word and the two senses the wordplay must activate) and is instructed to weigh four criteria: (i) the pun word is used or appears in a clear morphological variant; (ii) both senses are simultaneously activated, such that removing either sense breaks the joke; (iii) the generation reads as fluent text in the brief's target language; and (iv) it lands as funny under a "short punchline beats long explanation" heuristic. The verdict format is the three-class scheme `[[A]]`/`[[B]]`/`[[C]]` (better-A, better-B, comparable) inherited from MT-Bench, deliberately preferred over Arena Hard's five-tier scheme [34] because the "slightly better" vs. "much better" distinction is not consistently distinguished by LLM judges on humour. The judge also explicitly receives the canonical anti-bias triplet (do not prefer the longer, the more formatted, or the self-explanatory response; do not let presentation order influence the verdict) used in [33]. The constraint-aware framing – particularly the explicit "both senses must be activated" check – follows the pairwise pun-explanation judge of [23], lifted from explanation evaluation to generation evaluation. The complete prompt text, as delivered to the judges, is reproduced in Section C.

4.2. A cross-provider judge ensemble

To reduce single-judge bias we ensemble three LLM judges drawn from three distinct providers: GPT-OSS-20B [35], DEEPSEEK-CHAT (DeepSeek V4 Flash, building on the V3 architecture [36]), and GEMINI-2.5-FLASH-LITE [37]. Each judge votes independently on every sampled battle, and verdicts are aggregated by majority. Per-judge votes are retained for downstream agreement analysis (Section 6.3). A judge ID exhibiting a one-side preference rate above 92% across at least 20 judgements is automatically excluded from the Bradley-Terry pool, mirroring the defensive heuristics proposed alongside the Chatbot Arena vote-manipulation attacks of Huang et al. [38] and the model-fingerprinting / rigging attack of Min et al. [39]. A fourth provider (MISTRAL-SMALL-LATEST [40]) was piloted but dropped from the production ensemble after its rate ceiling injected retry-induced tail latencies that capped end-to-end throughput.

4.3. Sparse pairwise sampling

We collected $K = 365$ distinct submitted runs across the three tracks (English: 98, Spanish: 134, French: 133); a full Cartesian product would yield ≈ 3.4 M head-to-head battles and exceed the LLM-judge budget by two orders of magnitude. Instead, we adopt the Chatbot Arena protocol [32]: a sparse pairwise sample over a connected vote graph. The Bradley-Terry MLE remains well-defined on this sparse graph because the logistic loss is convex and identifiability only requires graph connectivity rather than complete pairwise coverage; the confidence interval scales with $1/\sqrt{n_{\text{system}}}$ rather than with the number of explicit pair comparisons. We sampled $\approx 40{,}000$ battles (≈ 110 battles per system on average), yielding per-system confidence intervals on the order of ± 150 Elo, sufficient to discriminate the top-k systems within each language track at the run-level granularity that interests the participants.

5. Participants' Approaches

Four teams submitted 98 distinct runs to the English track, four teams submitted 134 distinct runs to Spanish, and four teams submitted 133 distinct runs to French. The University of Amsterdam contributions are split across two Codabench accounts (`ezrasmink` for the main run set and `zero0` for an additional EN-only fine-tuning batch); we treat them as a single *UAms* entry below and in the per-team summary of Table 2.

UAms [41] submitted 9 EN, 5 ES, and 4 FR runs to Task 4. The University of Amsterdam team adopted a single multi-task pipeline shared across Tasks 1–4: candidate humorous generations from instruction-tuned LLMs – including FLAN-T5-BASE/LARGE, MISTRAL-7B-INSTRUCT, QWEN2.5-14B/32B-INSTRUCT, LLAMA-3-NEURONA-8B, SALAMANDRA-7B-INSTRUCT and CHOCOLATINE-2-14B-INSTRUCT – are reranked through a humour-detection classifier and a neural pun detector that score each candidate's pun-word coverage, ambiguity, and distinctiveness, with the top-scoring generation per brief retained for submission.

SaLT [42] submitted 4 EN, 4 ES, and 4 FR runs to Task 4 (Codabench `jayi2525`). The Carnegie Mellon University Africa team adapted the *Cognitive Synergy Framework* (CSF) [19] – a *Mixture-of-Thought* approach grounded in psychological theories of humour, originally developed for English headline generation – to the domain of constrained cross-lingual pun generation in English, French, and Spanish. Given a brief `(pun_word, sense_a, sense_b)`, CSF generates K candidate jokes via K distinct cognitive personas and a pairwise ranking model selects the best candidate under four conditions: (i) use the given pun word (or a clear morphological variant), (ii) activate both intended senses simultaneously, (iii) read naturally without template-like phrasing, and (iv) be written entirely in the target language. SaLT submitted four systems designed as a pairwise ablation isolating where cognitive guidance contributes most – *test time* or *training time*: KIMI-CSF (the primary guided system, KIMI-K2.6 backbone running the full CSF pipeline at inference), KIMI-PLAIN (the same backbone with a direct zero-shot prompt, no guidance), and HUMORGEN-JOKER-14B / 32B (open-weight QWEN3-14B / QWEN3-32B student models LoRA-distilled on CSF-curated synthetic data, then prompted without explicit guidance). KIMI-CSF tops the Bradley-Terry leaderboard in all three language tracks (Section 6).

Table 2
Best run per team for Task 4, per language track. *# runs* counts distinct `run_id`s submitted by the team to that track; *BT* is the Bradley-Terry rating of the team's best run with 95% bootstrap CI; *Rank* is that run's position in the per-track full leaderboard (Section A).

	Team	#	BT	Rank	95% CI	run_id
EN	SaLT [42]	4	1,257	1	[1215, 1308]	`SaLT_KimiCSF`
	DUTH	80	1,127	5	[1055, 1184]	`DUTH_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_04`
	UBO [43]	5	1,125	6	[1084, 1168]	`UBO_gemini-3_pun`
	UAms [41]	9	983	74	[937, 1024]	`UAms_Qwen2.5-32B-Instruct`
ES	SaLT [42]	4	1,515	1	[1450, 1599]	`SaLT_KimiCSF`
	UBO [43]	5	1,351	4	[1310, 1396]	`UBO_gemini-3_pun`
	DUTH	120	1,314	6	[1259, 1367]	`DUTH_ARENAJUDGE_ES`
	UAms [41]	5	1,061	60	[1026, 1105]	`UAms_Llama-3-Instruct-Neurona-8b`
FR	SaLT [42]	4	1,395	1	[1342, 1449]	`SaLT_KimiCSF`
	DUTH	115	1,304	2	[1245, 1417]	`DUTH_FR_LEAK_NOR13_DROP_R04`
	UBO [43]	10	1,279	7	[1234, 1341]	`UBO_gemini-3-flash_pun`
	UAms [41]	4	1,191	41	[1149, 1229]	`UAms_Chocolatine-2-14B-Instruct`

UBO [43] submitted 5 EN, 5 ES, and 10 FR runs to Task 4 (Codabench `liana87`). Their approach was based on prompting four open instruction-tuned models – Qwen3-4B, Mistral-7B, Gemma-2-9B-it and Llama-3.1-8B – together with Gemini-3-flash. For each of the three languages, UBO submitted five runs using a different backbone prompted to generate a pun from the provided context (`pun_word`, `sense_a`, `sense_b`). For French, UBO also submitted five additional runs (one per backbone), prompting the models to generate, from the given `pun_word`, a *contrepèterie* – a common form of French wordplay in which swapping sounds, letters, or syllables between words reveals a hidden humorous meaning.

DUTH [44] submitted 80 EN, 120 ES, and 115 FR runs to Task 4 (Codabench `arampageos`), by far the largest sweep of any team. Their pool is dominated by ablation and reranking variants over a Gemma-2-9B-it-style backbone – dropout-substitution (DROP_SUB0x), per-language reweighting, leak-controlled training subsets, ARENA-judge reranking, and calibrated beam search – which is also what motivates the deduplication-by-`run_id` step in our canonicaliser. DUTH's CEUR notebooks document Tasks 1 and 2; the Task 4 working note was not available at the time of writing.

6. Results

6.1. Official Bradley-Terry leaderboards

Table 2 provides the best run per team per language track from the 3-judge Bradley-Terry leaderboard, together with the team's number of distinct runs (after `run_id` deduplication) and the global per-track rank of that best run. The full per-run leaderboards – 98 EN, 134 ES, and 133 FR distinct runs – appear in Section A. Ratings are on the canonical Elo scale anchored so that the BT origin lands near 1,000, with two-sided 95% bootstrap confidence intervals from 100 resamples.

SaLT's Kimi-CSF wins all three language tracks, and the internal ordering of the SaLT ablation is itself informative: the guided Kimi-CSF consistently beats the unguided Kimi-Plain on the same backbone (EN: 1,257 vs. 1,225; ES: 1,515 vs. 1,437; FR: 1,395 vs. 1,272), which in turn beats the distilled open-weight HumorGen-JOKER students – evidence that test-time cognitive guidance contributes more than distillation alone, consistent with the team's own ablation study [42]. Simple single-model prompting remains a strong baseline when the underlying model is strong: UBO's Gemini-3-flash run places in the top-7 of every track it entered, while the same prompting recipe over small open models lands mid-table or lower. The French *contrepèterie* runs occupy the bottom region of the FR leaderboard (Section A), reflecting both the intrinsic difficulty of the form and a judging protocol whose brief frames the target as a pun on the given word.

Table 3
Pun-word coverage (constraint-only metric) for each team's best run per track, alongside the Bradley-Terry rating (preference metric). The last rows give the track-level mean coverage over all runs and the Pearson correlation between coverage and BT rating across runs.

Track	Team (best run)	BT		Coverage
EN	SaLT (KimiCSF)	1,257		100.0%
	DUTH (EN_BEST984_PLUS_TRAINSEL_TOP1_B12_04)	1,127		98.7%
	UBO (gemini-3_pun)	1,125		88.7%
	UAms (Qwen2.5-32B-Instruct)	983		96.7%
	track mean over 98 runs		97.0%	$r(\mathrm{BT}, \mathrm{cov}) = 0.47$
ES	SaLT (KimiCSF)	1,515		97.3%
	UBO (gemini-3_pun)	1,351		88.7%
	DUTH (ARENAJUDGE_ES)	1,314		96.7%
	UAms (Llama-3-Instruct-Neurona-8b)	1,061		86.7%
	track mean over 134 runs		95.2%	$r(\mathrm{BT}, \mathrm{cov}) = 0.55$
FR	SaLT (KimiCSF)	1,395		97.4%
	DUTH (FR_LEAK_NOR13_DROP_R04)	1,304		91.7%
	UBO (gemini-3-flash_pun)	1,279		82.7%
	UAms (Chocolatine-2-14B-Instruct)	1,191		85.9%
	track mean over 133 runs		92.5%	$r(\mathrm{BT}, \mathrm{cov}) = 0.03$

6.2. Constraint compliance vs. preference: two axes of quality

The Arena rating measures *preference* – which generation the judges find funnier and better-formed. Task 4, however, is *constrained* generation, so we also report a constraint-only automatic metric: *pun-word coverage*, the fraction of a run's generations that contain the brief's pun word or a morphological variant of it (accent-insensitive substring match with a stemmed fallback). This is the generation-task analogue of the relevance-only / humour-only decomposition used in the retrieval task: coverage isolates instruction-following, the BT rating captures the holistic preference, and reading them together shows whether a run's rating comes from comedy or from mere compliance.

Three observations follow from Table 3. First, compliance is near-saturated: mean coverage is 92–97% per track, so virtually all systems have learned to include the pun word – the constraint that separates runs is the harder *both-senses* condition, which only the pairwise judges assess. Second, coverage correlates moderately with preference in English ($r = 0.47$) and Spanish ($r = 0.55$) but not at all in French ($r = 0.03$): the FR pool contains diagnostic runs with deliberately degenerate outputs at both coverage extremes, and French homophone briefs admit valid wordplay whose surface form drifts further from the quoted pun word. Third, high coverage is necessary but nowhere near sufficient – UAms runs at 96.7% coverage still rank mid-table, because compliance without a working double meaning does not win battles.

6.3. Judge agreement

Retaining per-judge verdicts lets us quantify how contested the humour judgments are. Table 4 reports pairwise Cohen's κ over the three-class verdicts on the $\approx$39k battles where all three judges returned a parseable verdict.

Agreement is moderate ($\kappa \approx 0.43$–0.53) – typical for subjective pairwise preference and the reason we ensemble rather than trust a single judge. Agreement is systematically highest on French and lowest on English, and 59.1% of battles are decided unanimously. The judges differ markedly in tie propensity (1.1% to 9.8%), which majority aggregation absorbs: a lone tie vote defers the decision to the two decisive judges. No judge approached the 92% one-side exclusion threshold; position bias at the ensemble level is negligible (48.2% A vs. 45.9% B wins, the remainder ties).

Table 4
Pairwise Cohen's κ between the ensemble judges (3-class verdicts: A / B / tie), with raw agreement in parentheses, per track and overall. The final column gives each judge's tie rate.

Judge pair	EN	ES	FR	Overall
GPT-OSS-20B VS. DEEPSEEK-CHAT	0.44	0.48	0.50	0.48 (72%)
GPT-OSS-20B VS. GEMINI-2.5-FLASH-LITE	0.44	0.43	0.50	0.47 (71%)
DEEPSEEK-CHAT VS. GEMINI-2.5-FLASH-LITE	0.46	0.51	0.53	0.52 (72%)
Tie rates: GPT-OSS-20B 1.1%, DEEPSEEK-CHAT 6.9%, GEMINI-2.5-FLASH-LITE 9.8%				
Unanimous verdicts: 59.1% of 39,039 fully-judged battles				

Table 5
English track: per-team ensemble win rate by brief provenance. n is the number of battle slots involving the team's runs on that subset.

	PunEval briefs		Mots-valises briefs	
Team	**Win rate**	n	**Win rate**	n
SaLT (`jayi2525`)	78.9%	530	69.1%	742
DUTH (`arampageos`)	53.5%	4,827	55.3%	6,718
UBO (`liana87`)	41.7%	666	41.4%	925
UAms (`ezrasmink`)	29.3%	666	24.8%	936
UAms (`zero0`)	15.3%	537	11.4%	735

6.4. Per-source breakdown (English track)

Unlike the retrieval and translation tasks, Task 4 has no train-phase leaderboard – systems generate only for the test briefs, so no separate results on train data are available; the train split serves for few-shot selection and fine-tuning only. What we can decompose is the *test* performance by brief provenance. The English test set mixes PunEval-derived briefs (WordNet sense pairs, `het_`/`hom_` ids) with JOKER Mots-valises portmanteau briefs (`mots_` ids). Table 5 reports each team's win rate (fraction of non-tie ensemble votes won) on the two subsets.

The ordering is stable across the two sources – no team's rank depends on the brief provenance – but the top system's margin narrows on portmanteau briefs (SaLT: 78.9%→69.1%), suggesting that blend-word briefs, whose "senses" are the two source words of the blend, are harder to satisfy convincingly than dictionary sense pairs.

6.5. Qualitative examples

To illustrate what the 3-judge ensemble actually saw and how it voted, Table 6 reproduces three real battles from the Arena – one per language track. Each block shows the structured brief shown to the systems, the two competing generations (A and B), the three per-judge verdicts under prompt, and the final majority-aggregated ensemble verdict. The first two examples are unanimous, and the third is split (2–1), illustrating both the agreement regime and the kinds of generations that draw disagreement among judges.

7. Discussion and Conclusions

We have described the first edition of JOKER Task 4 on Humour Generation: a constrained, brief-driven humour generation task in English, Spanish, and French, evaluated through an Arena-style pairwise protocol with a constraint-aware judge prompt, a cross-provider LLM-judge ensemble, and sparse Bradley-Terry ranking over 365 distinct runs from four teams.

The task attracted considerable attention, but due to the late availability of the data, only a subset of

Table 6
Three example head-to-head battles from the Arena (one per language track), with per-judge verdicts and the final majority-aggregated ensemble winner. A and B denote the order in which the two systems were shown to the judges; the per-judge verdicts use the same A/B/tie three-class scheme.

English brief — pun word: `lama` *Sense 1:* a Tibetan or Mongolian priest of Lamaism. *Sense 2:* a wild or domesticated South American cud-chewing animal related to camels but smaller and lacking a hump.
A (DUTH / `EN_BEST984_DROP_PATCH_BUCKET_05`): I followed a lama to the mountains, hoping to learn some spiritual enlightenment. He just kept asking me to spit out my gum. B (SaLT / `KimiCSF`): I joined what I thought was a Buddhist retreat—everyone followed our leader religiously. Then I noticed he spat when excited and ate exclusively hay. Turns out my lama was not what I expected. *Verdicts:* `gpt-oss-20b`=B, `deepseek-chat`=B, `gemini-2.5-flash-lite`=B. **Ensemble: B** (unanimous).
Spanish brief — pun word: `picazón` *Sentido 1:* picazón (*démangeaison*). *Sentido 2: itching* ∼ alergia; picazón; fiebre del heno.
A (SaLT / `KimiCSF`): Le dije al alérgico: "Relájate, es solo fiebre del heno." Él respondió: "¿Solo? Tengo picazón… ¡y ya no sé si es la alergia o el heno existencial!" B (DUTH / `ES_DIAG_REFERENCE_STYLE_2`): Nadie esperaba que picazón fuera el remate. *Verdicts:* `gpt-oss-20b`=A, `deepseek-chat`=A, `gemini-2.5-flash-lite`=A. **Ensemble: A** (unanimous).
French brief — pun word: `potentiel` *Sens 1:* potentiel (électricité). *Sens 2:* potentiel (ressources).
A (SaLT / `KimiCSF`): Panne électrique. Je sors mon facture, lumière de bougie: "Monsieur, vous avez du potentiel." Je tremble… est-ce qu'ils parlent de mes ressources, ou est-ce que le compteur se moque de moi? B (DUTH / `FR_LEAK_TOP21`): Il a un potentiel énorme, mais il préfère rester branché. *Verdicts:* `gpt-oss-20b`=A, `deepseek-chat`=B, `gemini-2.5-flash-lite`=B. **Ensemble: B** (2–1 split, brevity preferred).

registered participants could still submit runs. The results are nonetheless very promising: both the LLM Arena evaluation and the qualitative analysis suggest that explicitly prompting models with the pun word and its two senses is a viable route to wordplay generation, and sidesteps the memorisation-driven lack of diversity documented for unconstrained joke generation [17]. The winning system – SaLT's theory-guided, multi-persona Kimi-CSF [42, 19] – beat its own unguided backbone in every track, indicating that structured, test-time cognitive guidance still adds measurable comedic value on top of a strong foundation model; at the same time, the constraint-compliance analysis shows that simply following the brief is nearly saturated and is not what separates the leaders.

For future editions we plan to (i) enlarge and diversify the judge ensemble and study the sensitivity of the ranking to the judge prompt, (ii) extend the brief taxonomy beyond puns and portmanteaux (e.g., *contrepèteries* judged under form-specific criteria rather than the pun-centric prompt), and (iii) coordinate with the emerging humour-generation evaluation community around SemEval MWAHAHA [20] on shared judging protocols and vote-integrity defences [38, 39]. We keep the Codabench instances running after the deadline as a living test collection: post-competition submissions remain welcome.

For more information about the JOKER lab this year, please refer to the overview paper [1], and the Working Notes papers for Task 1: Humour-aware Information Retrieval [2], Task 2: Pun Translation [3], and Task 3: Onomastic Wordplay Translation [4]. Visit the JOKER website at https://joker-project.com for any other information related to the track.

Acknowledgments

We thank the Master's students in translation and technical writing from the University of Brest for participating in data annotation, and CodaBench for hosting the competition.

Liana Ermakova is partly funded by the National Research Agency (ANR-19-GURE-0001) under the program "*Investissements d'avenir*" integrated into France 2030. Jaap Kamps is partly funded by

the Netherlands Organization for Scientific Research (NWO NWA #1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT*, *Claude*, and *Grammarly* in order to: Grammar and spelling check, Paraphrase and reword. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, et al., Overview of the CLEF 2026 JOKER track: Humor detection, search, and translation, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[2] P. Vachharajani, A. Mukherjee, J. Singh, K. Tewari, S. Pal, J. Kamps, L. Ermakova, Overview of the CLEF 2026 JOKER Task 1: Humor-Aware Information Retrieval in English and Hinglish, in: [45], 2026.

[3] L. Ermakova, Y. Naud, L. Binet, B. Dumas, J. Kamps, Overview of the CLEF 2026 JOKER Task 2: Pun Translation from English to French, in: [45], 2026.

[4] L. Ermakova, Y. Naud, L. Binet, J. Kamps, Overview of the CLEF 2026 JOKER Task 3: Onomastic Wordplay Translation from English to French, in: [45], 2026.

[5] A. Nijholt, A. Niculescu, A. Valitutti, R. E. Banchs, Humor in human-computer interaction: a short survey, 2017.

[6] R. Oliveira, P. Arriaga, M. Axelsson, A. Paiva, Humor–robot interaction: a scoping review of the literature and future directions, International Journal of Social Robotics 13 (2021) 1369–1383.

[7] A.-M. Velentza, A.-G. Bosser, Humor style drives laughter, topic shapes acceptability: Evaluating bilingual personal and political robot-delivered AI jokes, 2026. arXiv:2606.13256, preprint.

[8] S. Petrović, D. Matthews, Unsupervised joke generation from big data, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 228–232. URL: https://aclanthology.org/P13-2041/.

[9] A. Valitutti, H. Toivonen, A. Doucet, J. M. Toivanen, "let everything turn well in your wife": Generation of adult humor using lexical constraints, ACL 2013, Association for Computational Linguistics, Sofia, Bulgaria, 2013, p. 243–248. URL: https://aclanthology.org/P13-2044, [Online; accessed 2021-10-14].

[10] B. A. Hong, E. Ong, Automatically extracting word relationships as templates for pun generation, Association for Computational Linguistics, Boulder, Colorado, 2009, p. 24–31. URL: https://aclanthology.org/W09-2004, [Online; accessed 2021-10-19].

[11] Z. Yu, J. Tan, X. Wan, A neural approach to pun generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1650–1660. URL: https://aclanthology.org/P18-1153/. doi:10.18653/v1/P18-1153.

[12] H. He, N. Peng, P. Liang, Pun generation with surprise, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1734–1744. URL: https://aclanthology.org/N19-1172/. doi:10.18653/v1/N19-1172.

[13] A. Mittal, Y. Tian, N. Peng, AmbiPun: Generating humorous puns with ambiguous context, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022. URL: https://aclanthology.org/2022.naacl-main.77/. doi:10.18653/v1/2022.naacl-main.77.

[14] J. Sun, A. Narayan-Chen, S. Oraby, S. Gao, T. Chung, J. Huang, Y. Liu, N. Peng, Context-situated pun generation, 2022. doi:10.48550/arXiv.2210.13522. arXiv:2210.13522, presented at EMNLP 2022.

[15] N. N. Joshi, Evaluating human perception and bias in ai-generated humor, in: Proceedings of the 1st Workshop on Computational Humor (CHum), Association for Computational Linguistics, 2025, pp. 79–87. URL: https://aclanthology.org/2025.chum-1.9/.

[16] Y. Cao, J. Cao, Y. Hou, L.-J. Ji, How humorous is ai? exploring chatgpt's role in humor generation and human-ai interaction, Computers in Human Behavior Reports 20 (2025) 100807.

[17] S. Jentzsch, K. Kersting, ChatGPT is fun, but it is not funny! Humor is still challenging large language models, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Toronto, Canada, 2023. URL: https://aclanthology.org/2023.wassa-1.29/. doi:10.18653/v1/2023.wassa-1.29.

[18] S. Zhong, Z. Huang, S. Gao, W. Wen, L. Lin, M. Zitnik, P. Zhou, Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. ArXiv:2312.02439.

[19] E. Ajayi, P. Mitra, HumorGen: Cognitive synergy for humor generation in large language models via persona-based distillation, 2026. doi:10.48550/arXiv.2604.09629. arXiv:2604.09629.

[20] S. Castro, L. Chiruzzo, N. Deng, J.-A. Meaney, S. Góngora, I. Sastre, V. Amoroso, G. Rey, S. Rahili, G. Moncecchi, J. J. Prada, A. Rosá, R. Mihalcea, SemEval-2026 task 1: MWAHAHA – models write automatic humor and humans annotate, https://pln-fing-udelar.github.io/semeval-2026-humor-gen/, 2026. Task overview to appear in Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026).

[21] A. Tikhonov, A. Ivanov, lmfaoooo at SemEval-2026 task 1: Humor is an audience. Preference modeling for constrained humor generation, 2026. doi:10.48550/arXiv.2606.00022. arXiv:2606.00022, semEval-2026 Task 1 system paper; ranked 1st on the English and Chinese subtasks.

[22] abateam, abateam at SemEval-2026 task 1: Plan2joke – humor policies for type-specific two-pass humor generation, in: Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026), Association for Computational Linguistics, 2026. URL: https://aclanthology.org/2026.semeval-1.295/.

[23] Z. Xu, S. Yuan, L. Chen, D. Yang, "a good pun is its own reword": Can large language models understand puns?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11766–11782. URL: https://aclanthology.org/2024.emnlp-main.657/. doi:10.18653/v1/2024.emnlp-main.657.

[24] T. Miller, C. Hempelmann, I. Gurevych, SemEval-2017 task 7: Detection and interpretation of English puns, in: S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, D. Jurgens (Eds.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 58–68. doi:10.18653/v1/S17-2005.

[25] J. Sun, A. Narayan-Chen, S. Oraby, A. Cervone, T. Chung, J. Huang, Y. Liu, N. Peng, ExPUNations: Augmenting puns with keywords and explanations, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4590–4605. URL: https://aclanthology.org/2022.emnlp-main.304/. doi:10.18653/v1/2022.emnlp-main.304.

[26] L. Ermakova, R. Campos, A. Bosser, T. Miller, Overview of the CLEF 2025 JOKER lab: Humour in machine, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 315–337. URL: https://doi.org/10.1007/978-3-032-04354-2_18. doi:10.1007/978-3-032-04354-2_18.

[27] L. Ermakova, A. Bosser, A. Jatowt, T. Miller, The JOKER corpus: English-french parallel data for multilingual wordplay recognition, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2796–2806. URL: https://doi.org/10.1145/3539618.3591885. doi:10.1145/3539618.3591885.

[28] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 1 - pun detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1785–1803. URL: https://ceur-ws.org/Vol-3497/paper-149.pdf.

[29] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 2 - pun location and interpretation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1804–1817. URL: https://ceur-ws.org/Vol-3497/paper-150.pdf.

[30] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 automatic wordplay analysis task 3 - pun translation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1818–1827. URL: https://ceur-ws.org/Vol-3497/paper-151.pdf.

[31] L. Ermakova, T. Miller, A. Bosser, V. M. Palma-Preciado, G. Sidorov, A. Jatowt, Overview of JOKER - CLEF-2023 track on automatic wordplay analysis, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 397–415. URL: https://doi.org/10.1007/978-3-031-42448-9_26. doi:10.1007/978-3-031-42448-9_26.

[32] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, I. Stoica, Chatbot Arena: an open platform for evaluating LLMs by human preference, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and Chatbot Arena, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023, pp. 46595–46623.

[34] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, I. Stoica, From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline, 2024. doi:10.48550/arXiv.2406.11939. arXiv:2406.11939.

[35] OpenAI, gpt-oss-120b & gpt-oss-20b model card, 2025. doi:10.48550/arXiv.2508.10925. arXiv:2508.10925.

[36] DeepSeek-AI, DeepSeek-V3 technical report, 2024. doi:10.48550/arXiv.2412.19437. arXiv:2412.19437.

[37] G. Comanici, Google Gemini Team, Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. doi:10.48550/arXiv.

2507.06261. arXiv:2507.06261.
[38] Y. Huang, M. Nasr, A. Angelopoulos, N. Carlini, W.-L. Chiang, C. A. Choquette-Choo, D. Ippolito, M. Jagielski, K. Lee, K. Z. Liu, I. Stoica, F. Tramèr, C. Zhang, Exploring and mitigating adversarial manipulation of voting-based leaderboards, in: Proceedings of the 42nd International Conference on Machine Learning (ICML), volume 267 of *Proceedings of Machine Learning Research*, PMLR, 2025. URL: https://proceedings.mlr.press/v267/huang25z.html, arXiv:2501.07493.
[39] R. Min, T. Pang, C. Du, Q. Liu, M. Cheng, M. Lin, Improving your model ranking on chatbot arena by vote rigging, 2025. doi:10.48550/arXiv.2501.17858. arXiv:2501.17858.
[40] Mistral AI, Mistral Small 3, https://mistral.ai/news/mistral-small-3/, 2025. Model release announcement; accessed June 2026.
[41] J. van Bakel, R. Flier, E. Smink, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2026 JOKER Track, in: [45], 2026.
[42] E. Ajayi, P. Mitra, Cross-Lingual Cognitive Synergy for Constrained Humor Generation in LLMs: SaLT Lab at CLEF 2026 JOKER Track, in: [45], 2026.
[43] L. Ermakova, Y. Naud, L. Binet, B. Dumas, Testing LLMs on JOKER Task 3: Onomastic Wordplay Translation and Task 4: Humour Generation. UBO at CLEF 2026, in: [45], 2026.
[44] G. Arampatzis, A. Arampatzis, DUTH at CLEF JOKER 2026: Conservative Hybridisation for Humor-Aware Retrieval, Pun Translation, Onomastic Wordplay Translation, and Multilingual Humour Generation, in: [45], 2026.
[45] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

A. Full per-run leaderboards

Tables 7, 8 and 9 list the complete Bradley-Terry leaderboards for the English (98 runs), Spanish (134 runs), and French (133 runs) tracks. Ratings, confidence intervals, and battle counts are as described in Section 4.

Table 7: Task 4 English track final Bradley-Terry leaderboard from the 3-judge LLM ensemble (Groq `openai/gpt-oss-20b` + `deepseek-chat` + `models/gemini-2.5-flash-lite`, `majority` aggregation) with 100-round bootstrap 95% CIs. *Runs* are deduplicated by participant-declared `run_id`; n_b counts head-to-head battles per run on the sparse Arena sample.

Rank	Team / Run	BT	CI low	CI high	n_b
1	`jayi2525` / `SaLT_task4_KimiCSF`	1257	1215	1308	344
2	`jayi2525` / `SaLT_task4_KimiPlain`	1225	1187	1286	296
3	`jayi2525` / `SaLT_task4_HumorGen14B`	1168	1128	1200	333
4	`jayi2525` / `SaLT_task4_HumorGen32B`	1135	1097	1174	325
5	`arampageos` / `DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_04`	1127	1055	1184	142
6	`liana87` / `UBO_task4_gemini-3_pun`	1125	1084	1168	326
7	`arampageos` / `DUTH_Task4_EN_BEST984_DROP_ID_het_1195`	1112	1049	1202	115
8	`arampageos` / `DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_05`	1107	1025	1186	110
9	`arampageos` / `DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_04`	1105	1034	1171	126
10	`arampageos` / `DUTH_Task4_EN_BEST984_DROP_ID_mots_1479`	1097	1020	1169	123
11	`arampageos` / `DUTH_Task4_EN_BEST984_PLUS_BASE_SAFE_B12_00`	1097	1035	1166	121

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
12	arampageos / DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_02	1092	1020	1147	117
13	arampageos / DUTH_Task4_EN_LEAK12_DROP_R01_hom_1991	1090	1030	1152	128
14	arampageos / DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_06	1088	1016	1153	111
15	arampageos / DUTH_Task4_EN_QWEN14_EN_XID_TOP8	1088	1018	1151	116
16	arampageos / DUTH_Task4_EN_BEST984_CONFIRM	1087	1026	1159	111
17	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_mots_0083	1084	1019	1159	123
18	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_00	1079	1019	1148	127
19	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB03	1076	1016	1145	124
20	arampageos / DUTH_Task4_EN_QWEN14_EN_XID_TOP1	1070	1004	1135	106
21	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB00	1069	1008	1153	121
22	arampageos / DUTH_Task4_EN_FROM8100_TOP5	1069	1003	1142	126
23	arampageos / DUTH_Task4_EN_LEAK_TOP13	1067	993	1146	104
24	arampageos / DUTH_Task4_EN_LEAK_NOR1_DROP_R02	1066	1004	1130	122
25	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_hom_1803	1066	1006	1147	102
26	arampageos / DUTH_Task4_EN_LEAK12_DROP_R02_hom_1803	1065	1007	1124	123
27	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_mots_0535	1063	1002	1123	120
28	arampageos / DUTH_Task4_EN_WEIGHTPOOL_TOP1	1063	994	1147	117
29	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_het_1780	1061	992	1153	106
30	arampageos / DUTH_Task4_EN_GOOGLE_XID_TOP3	1060	977	1139	114
31	arampageos / DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_07	1057	986	1123	117
32	arampageos / DUTH_Task4_EN_LEAK_NOR1_TOP14	1057	991	1148	108
33	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_06	1055	984	1118	117
34	arampageos / DUTH_Task4_EN_FROM8100_TOP1	1054	986	1109	126
35	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_mots_0027	1051	973	1141	104
36	arampageos / DUTH_Task4_EN_BEST1019_CONFIRM	1048	965	1110	127
37	arampageos / DUTH_Task4_EN_FROM8100_TOP2	1048	973	1124	114
38	arampageos / DUTH_Task4_gemma29bit_EN_balanced	1048	984	1112	115
39	arampageos / DUTH_Task4_EN_LEAK_NOR1_TOP20	1047	996	1111	141
40	arampageos / DUTH_Task4_EN_QWEN14_EN_XID_TOP2	1047	985	1108	149
41	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_00	1046	998	1101	140

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
42	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_mots_1263	1044	991	1125	106
43	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_01	1038	981	1107	141
44	arampageos / DUTH_Task4_EN_BEST984_DROP_ID_hom_1573	1037	965	1125	97
45	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_02	1036	982	1099	133
46	arampageos / DUTH_Task4_EN_LEAK12_DROP_R11_mots_1945	1035	972	1094	117
47	arampageos / DUTH_Task4_EN_QWEN14_EN_START_SAMPLE10	1032	999	1062	365
48	arampageos / DUTH_Task4_EN_BEST984_DROP_PATCH_BUCKET_00	1031	961	1118	121
49	arampageos / DUTH_Task4_EN_WEIGHTPOOL_TOP5	1026	965	1088	135
50	arampageos / DUTH_Task4_EN_GOOGLE_XID_TOP2	1026	961	1078	128
51	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB02	1023	949	1085	120
52	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_05	1023	956	1090	126
53	arampageos / DUTH_Task4_EN_QWEN14_EN_XID_TOP5	1017	935	1108	107
54	arampageos / DUTH_Task4_EN_LEAK_TOP8	1014	969	1062	156
55	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_04	1013	954	1083	138
56	arampageos / DUTH_Task4_EN_LEAK_TOP20	1013	947	1076	145
57	arampageos / DUTH_Task4_EN_LEAK12_DROP_R07_mots_1783	1013	919	1110	96
58	arampageos / DUTH_Task4_EN_LEAK_NOR1_TOP25	1012	952	1078	123
59	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_09	1012	952	1082	127
60	arampageos / DUTH_Task4_EN_QWEN14_EN_XID_TOP3	1012	957	1067	136
61	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB05	1012	944	1076	121
62	arampageos / DUTH_Task4_EN_GOOGLE_XID_TOP1	1007	938	1080	128
63	arampageos / DUTH_Task4_EN_QWEN14_EN_START_SAMPLE07	1003	964	1040	320
64	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_05	999	945	1057	156
65	arampageos / DUTH_Task4_EN_LEAK_TOP12	998	908	1062	105
66	arampageos / DUTH_Task4_EN_WEIGHTPOOL_TOP8	997	920	1061	120
67	arampageos / DUTH_Task4_EN_BUCKET_CAND_00	996	947	1047	149
68	arampageos / DUTH_Task4_EN_FROM8100_TOP12	996	942	1052	139
69	arampageos / DUTH_Task4_EN_LEAK_TOP1	996	938	1046	115
70	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_03	996	938	1060	153
71	liana87 / UBO_task4__qwen_pun	995	953	1039	299
72	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_10	989	934	1057	155

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
73	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB04	985	938	1045	134
74	ezrasmink / UAms_task4_Qwen2.5-32B-Instruct	984	938	1025	316
75	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_07	979	927	1031	119
76	arampageos / DUTH_Task4_EN_BEST984_PLUS_TRAINSEL_TOP1_B12_06	976	897	1044	85
77	arampageos / DUTH_Task4_EN_BEST1019_DROP_SUB01	964	888	1007	150
78	arampageos / DUTH_Task4_EN_BUCKET_CAND_08	942	898	993	164
79	arampageos / DUTH_Task4_EN_BEST1019_PLUS_TRAINSEL_TOP1_B12_08	942	884	993	158
80	arampageos / DUTH_Task4_EN_QWEN14_EN_START_SELECTOR_ALL	937	902	975	327
81	ezrasmink / UAms_task4_Qwen2.5-14B-Instruct	927	882	959	341
82	arampageos / DUTH_Task4_BASELINE_EN_SAFE	924	881	967	304
83	arampageos / DUTH_Task4_EN_TRAINSEL_TOP12	923	877	959	328
84	ezrasmink / UAms_task4_Mistral-7B-Instruct-v0.2	922	882	952	341
85	arampageos / DUTH_Task4_EN_TRAINSEL_TOP1	913	875	952	318
86	arampageos / DUTH_Task4_EN_TRAINSEL_TOP5	907	864	943	326
87	liana87 / UBO_task4_mistral_pun	903	869	937	343
88	arampageos / DUTH_Task4_EN_TRAINSEL_BUCKET_00	902	863	939	310
89	liana87 / UBO_task4_gemma_pun	895	851	929	360
90	arampageos / DUTH_Task4_EN_TRAINSEL_TOP40	891	844	928	267
91	arampageos / DUTH_Task4_BASELINE_EN_TOP80	851	808	891	302
92	liana87 / UBO_task4_llama_pun	816	765	858	316
93	zero0 / UAms_Task4_Gemma2_B	734	658	774	312
94	zero0 / UAms_Task4_Gemma2_A	727	678	773	301
95	zero0 / UAms_Task4_T5ptft	591	528	634	340
96	ezrasmink / UAms_task4_flan-t5-large	549	469	607	326
97	zero0 / UAms_Task4_T5ft	541	467	612	351
98	ezrasmink / UAms_task4_flan-t5-base	422	328	466	327

Table 8: Task 4 Spanish track final Bradley-Terry leaderboard from the 3-judge LLM ensemble (Groq `openai/gpt-oss-20b` + `deepseek-chat` + `models/gemini-2.5-flash-lite`, `majority` aggregation) with 100-round bootstrap 95% CIs. *Runs* are deduplicated by participant-declared `run_id`; n_b counts head-to-head battles per run on the sparse Arena sample.

Rank	Team / Run	BT	CI low	CI high	n_b
1	jayi2525 / SaLT_task4_KimiCSF	1515	1450	1599	366
2	jayi2525 / SaLT_task4_KimiPlain	1437	1382	1495	342
3	jayi2525 / SaLT_task4_HumorGen32B	1369	1326	1420	354
4	liana87 / UBO_task4_gemini-3_pun	1351	1310	1396	363
5	jayi2525 / SaLT_task4_HumorGen14B	1329	1301	1383	375
6	arampageos / DUTH_Task4_ARENAJUDGE_ES	1314	1259	1367	296
7	arampageos / DUTH_Task4_HYBRIDEXACT_ES	1273	1235	1326	346
8	arampageos / DUTH_Task4_ARENAJUDGE_V2_ES	1264	1232	1312	323
9	arampageos / DUTH_Task4_gemma29bit_ES_balanced	1263	1221	1319	305

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
10	arampageos / DUTH_Task4_ES_DEEP_EXACTPUN_TOP2	1211	1151	1291	165
11	arampageos / DUTH_Task4_SCRATCHPAPER_ES_ARENA	1171	1137	1211	333
12	arampageos / DUTH_Task4_ES_DEEP_EXACTPUN_TOP3	1171	1110	1240	148
13	arampageos / DUTH_Task4_QWEN14_BLEUMIMIC_ES_PATCH	1160	1127	1205	288
14	arampageos / DUTH_Task4_QWEN14_BLEUMIMIC_ES_ALL	1142	1089	1176	262
15	arampageos / DUTH_Task4_ES_XID_SALVAGE_TOP5	1135	1087	1200	178
16	arampageos / DUTH_Task4_ES_NOBLOCK_PATCH6	1135	1085	1183	231
17	arampageos / DUTH_Task4_QWEN14_BLEUMIMIC_ES_SAFE	1133	1086	1188	239
18	arampageos / DUTH_Task4_ES_QUALITY_V2_MEDALLON	1126	1070	1191	158
19	liana87 / UBO_task4_gemma_pun	1125	1084	1172	322
20	arampageos / DUTH_Task4_ES_TRAINSEL_TOP1	1122	1062	1174	162
21	arampageos / DUTH_Task4_QWEN14_BLEUMIMIC_V2_ES_ALL	1122	1090	1161	327
22	arampageos / DUTH_Task4_ES_QUALITY_V1_SAFE	1122	1067	1178	170
23	arampageos / DUTH_Task4_ES_DEEP_SINGLE_R05_es_0919_EXACT_PUN_TRAIN_es_0925	1120	1073	1184	177
24	arampageos / DUTH_Task4_DUPCONSIST_ES	1118	1078	1162	286
25	arampageos / DUTH_Task4_ES_BEST834_PAIR_SINGLE_R05_R06	1117	1053	1173	174
26	arampageos / DUTH_Task4_ES_SCOREENS2_PATCH7	1115	1070	1180	245
27	arampageos / DUTH_Task4_BLOCK_ES_BESTBLOCK	1114	1075	1157	278
28	arampageos / DUTH_Task4_ES_DEEP_SINGLE_R03_es_2303_EXACT_PUN_TRAIN_es_0762	1114	1058	1179	168
29	arampageos / DUTH_Task4_ES_QUALITY_V3_PLUS_BASE	1114	1061	1188	174
30	arampageos / DUTH_Task4_ES_DEEP_EXACTPUN_TOP1	1111	1041	1181	167
31	arampageos / DUTH_Task4_ES_DEEP_SINGLE_R01_es_0921_EXACT_PUN_TRAIN_es_0927	1110	1068	1158	198
32	arampageos / DUTH_Task4_ES_SCOREENS2_PATCH12	1107	1078	1153	254
33	arampageos / DUTH_Task4_ES_SINGLE_R03_es_0459_WIN_REPORT_R04	1107	1053	1175	134
34	arampageos / DUTH_Task4_ES_COMBO_SAFE	1106	1066	1156	291
35	arampageos / DUTH_Task4_ES_XID_SALVAGE_TOP3	1105	1050	1163	154
36	arampageos / DUTH_Task4_ES_BEST833_PLUS_TRAINSEL_TOP1_B12_11	1103	1056	1154	176
37	arampageos / DUTH_Task4_ES_BEST834_DEEP_SINGLE_R03_es_2303_EXACT_PUN_TRAIN_es_0762_CONFIRM	1103	1044	1164	160
38	liana87 / UBO_task4_qwen_pun	1100	1072	1144	366
39	arampageos / DUTH_Task4_ES_SINGLE_R12_es_0635_WIN_REPORT_R16	1099	1030	1160	154
40	arampageos / DUTH_Task4_ES_MLSEL_RF_ALL	1096	1061	1139	346

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
41	arampageos / DUTH_Task4_ES_MLSEL_GB_TOP	1092	1054	1127	323
42	arampageos / DUTH_Task4_ES_SINGLE_R33_es_1117_WIN_REPORT_R44	1088	1040	1161	175
43	arampageos / DUTH_Task4_ES_SCOREENS_TOP80	1085	1031	1129	194
44	arampageos / DUTH_Task4_ES_SCOREENS2_TOP90	1085	1040	1138	217
45	arampageos / DUTH_Task4_ES_TRAINSEL_TOP3	1084	1032	1139	164
46	arampageos / DUTH_Task4_ES_SCOREENS_TOP30	1079	1030	1125	242
47	arampageos / DUTH_Task4_ES_BEST833_PLUS_TRAINSEL_TOP1_B12_02	1075	1017	1129	152
48	arampageos / DUTH_Task4_ES_BEST833_PLUS_TRAINSEL_TOP1_B12_01	1075	1018	1136	137
49	arampageos / DUTH_Task4_ES_DEEP_SINGLE_R02_es_0605_EXACT_PUN_TRAIN_es_0627	1075	1016	1137	170
50	arampageos / DUTH_Task4_ES_BEST834_PAIR_SINGLE_R04_R05	1074	1015	1149	166
51	arampageos / DUTH_Task4_ES_SINGLE_TOP5	1071	1024	1133	152
52	arampageos / DUTH_Task4_ES_SCOREENS2_TOP130	1071	1016	1123	199
53	arampageos / DUTH_Task4_ES_SINGLE_R24_es_1283_WIN_REPORT_R35	1071	1017	1135	165
54	arampageos / DUTH_Task4_ES_NEUTRAL_R31_R33	1070	1025	1127	182
55	arampageos / DUTH_Task4_ES_DEEP_EXACTPUN_TOP5	1068	1011	1134	170
56	arampageos / DUTH_Task4_ES_NEUTRAL_R01_R31_R33	1066	1010	1129	188
57	arampageos / DUTH_Task4_ES_SINGLE_R01_es_0921_WIN_REPORT_R01	1065	983	1125	138
58	arampageos / DUTH_Task4_ES_BEST834_EXACTPUN_TOP2	1064	1006	1122	156
59	arampageos / DUTH_Task4_ES_DEEP_SINGLE_R04_es_0459_EXACT_PUN_TRAIN_es_2080	1064	1012	1117	184
60	ezrasmink / UAms_task4_Llama-3-Instruct-Neurona-8b	1062	1027	1105	315
61	arampageos / DUTH_Task4_DATAAWARE_bleu_ES	1062	1033	1110	314
62	arampageos / DUTH_Task4_ES_TRAINSEL_TOP2	1061	997	1121	144
63	arampageos / DUTH_Task4_ES_BEST833_PLUS_TRAINSEL_TOP1_B12_00	1061	1007	1131	142
64	arampageos / DUTH_Task4_ES_SCOREENS2_PATCH4	1061	1022	1112	195
65	arampageos / DUTH_Task4_SCRATCHPAPER_ES_MIX	1060	1023	1102	317
66	arampageos / DUTH_Task4_DATAAWARE_arena_ES	1059	1021	1092	370
67	arampageos / DUTH_Task4_SCRATCHPAPER_ES_REFSTYLE	1059	1028	1095	352
68	arampageos / DUTH_Task4_ES_NEUTRAL_R01_R33	1058	1011	1102	173
69	arampageos / DUTH_Task4_ES_MANUALPATCH_AGGRESSIVE11	1058	1015	1113	184
70	arampageos / DUTH_Task4_ES_BEST834_EXACTPUN_TOP3_CONFIRM	1058	997	1116	168
71	arampageos / DUTH_Task4_ES_MANUALPATCH	1058	1005	1107	167
72	arampageos / DUTH_Task4_ES_NOBLOCK_TOP120	1057	1006	1120	196

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
73	arampageos / DUTH_Task4_ES_SINGLE_R31_es_0606_WIN_REPORT_R42	1056	997	1115	150
74	arampageos / DUTH_Task4_ES_BAL_TOP3_PLUS_R4_R6	1055	1005	1109	186
75	arampageos / DUTH_Task4_ES_SCOREENS2_PATCH3	1051	1004	1105	160
76	arampageos / DUTH_Task4_ES_TRAINSEL_TOP5	1050	1008	1118	173
77	arampageos / DUTH_Task4_ES_NOBLOCK_PATCH2	1048	998	1094	197
78	arampageos / DUTH_Task4_ES_BEST833_CONFIRM	1048	995	1126	144
79	arampageos / DUTH_Task4_ES_SCOREENS2_PATCH1	1048	996	1095	165
80	arampageos / DUTH_Task4_ES_BAL_TOP3_DROP_es_1091	1047	987	1116	139
81	arampageos / DUTH_Task4_ES_DEEP_EXACTPUN_TOP6	1044	982	1099	174
82	arampageos / DUTH_Task4_ES_NOBLOCK_PATCH0	1043	998	1109	194
83	arampageos / DUTH_Task4_ES_SCOREENS_TOP50	1041	998	1082	254
84	arampageos / DUTH_Task4_ES_MANUALPATCH_MEDIUM7	1041	979	1110	174
85	arampageos / DUTH_Task4_QWEN14_BLEUMIMIC_ES_TOP80	1038	1000	1086	274
86	arampageos / DUTH_Task4_ES_SINGLE_TOP1	1036	981	1085	159
87	arampageos / DUTH_Task4_ES_NEUTRAL_R01_R31	1034	981	1103	173
88	arampageos / DUTH_Task4_ES_NOBLOCK_TOP115	1033	986	1094	174
89	arampageos / DUTH_Task4_ES_SINGLE_TOP2	1031	976	1087	163
90	arampageos / DUTH_Task4_ES_BAL_TOP3_PLUS_R8_es_1288	1029	983	1087	172
91	arampageos / DUTH_Task4_ES_SCOREENS_PATCH0	1027	947	1083	158
92	arampageos / DUTH_Task4_ES_SINGLE_R08_es_0319_WIN_REPORT_R10	1027	954	1092	140
93	liana87 / UBO_task4_llama_pun	1019	984	1063	360
94	arampageos / DUTH_Task4_ES_SCOREENS2_TOP120	1018	975	1070	207
95	arampageos / DUTH_Task4_ES_XID_SALVAGE_TOP8	1017	967	1074	178
96	arampageos / DUTH_Task4_BASELINE_ES_SAFE	1016	983	1061	269
97	arampageos / DUTH_Task4_ES_SCOREENS_ALL	1014	963	1071	187
98	arampageos / DUTH_Task4_ES_SCOREENS_PATCH2	1012	957	1069	198
99	arampageos / DUTH_Task4_ES_NOBLOCK_PATCHm0p5	1009	947	1071	152
100	arampageos / DUTH_Task4_GEMMA2_9B_MEGA_ES_TOP1	1009	983	1050	349
101	arampageos / DUTH_Task4_ES_SCOREABL_NO_BLOCK	1006	947	1069	160
102	arampageos / DUTH_Task4_ES_SCOREABL_ONLY_GEMMA_QWEN	1006	954	1048	248
103	arampageos / DUTH_Task4_GEMMA2_9B_MEGA_ES_TOP2	1002	965	1043	334
104	arampageos / DUTH_Task4_GEMMA2_9B_MEGA_ES_TOP5	988	959	1021	313
105	arampageos / DUTH_Task4_ES_SCOREENS2_TOP95	988	935	1040	199

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
106	liana87 / UBO_task4_mistral_pun	987	945	1025	332
107	arampageos / DUTH_Task4_ES_SCOREABL_ONLY_GEMMA_MISTRAL	975	923	1030	196
108	arampageos / DUTH_Task4_MISTRAL7B_MEGA_ES_TOP1	975	933	1017	301
109	arampageos / DUTH_Task4_ES_QUALITY_V4_SPANISH_LOCKET	975	908	1051	138
110	arampageos / DUTH_Task4_ES_SCOREABL_NO_DATAAWARE	948	908	994	202
111	arampageos / DUTH_Task4_BASELINE_ES_TOP80	935	885	990	294
112	arampageos / DUTH_Task4_ES_DIAG_REFERENCE_STYLE_2	903	866	950	355
113	ezrasmink / UAms_task4_Mistral-7B-Instruct-v0.2	903	858	936	339
114	arampageos / DUTH_Task4_ES_TRAINSEL_TOP30	902	823	952	216
115	arampageos / DUTH_Task4_ES_ANALYSIS_analysis_ES_rule_generation_TOP60	881	832	928	240
116	arampageos / DUTH_Task4_BASELINE_ES_PATCH	865	826	904	324
117	arampageos / DUTH_Task4_ES_RULE_ALL	855	825	899	375
118	arampageos / DUTH_Task4_QWENSIMPLE_ES_D	829	789	860	350
119	arampageos / DUTH_Task4_POLICYQWEN_ES	791	757	829	350
120	arampageos / DUTH_Task4_QWENSIMPLE_ES_C	788	747	835	330
121	arampageos / DUTH_Task4_QWENSIMPLE_ES_B	783	737	827	318
122	arampageos / DUTH_Task4_BASELINE_ES_ALLRULE	762	707	807	346
123	arampageos / DUTH_Task4_QWENSIMPLE_ES_BEST	756	703	800	309
124	arampageos / DUTH_Task4_BASELINE_ES_PUNWORD	701	650	740	342
125	arampageos / DUTH_Task4_ES_TRAINSEL_TOP120	611	520	670	275
126	ezrasmink / UAms_task4_flan-t5-large	393	276	479	303
127	ezrasmink / UAms_task4_flan-t5-base	344	272	406	340
128	arampageos / DUTH_Task4_ES_DIAG_FACT_SENT	335	261	405	327
129	arampageos / DUTH_Task4_ES_DIAG_BOW_STUFF	304	182	368	329
130	arampageos / DUTH_Task4_ES_DIAG_SHORT_PUN_CUE	303	186	386	354
131	arampageos / DUTH_Task4_ES_DIAG_FACT_SHORT	212	96	290	316
132	arampageos / DUTH_Task4_ES_DIAG_CUES_ONLY	92	-111	208	348
133	arampageos / DUTH_Task4_ES_DIAG_SENSE_B	90	-47	202	331
134	ezrasmink / UAms_task4_salamandra-7b-instruct	-270	-1734	-142	335

Table 9: Task 4 French track final Bradley-Terry leaderboard from the 3-judge LLM ensemble (Groq `openai/gpt-oss-20b` + `deepseek-chat` + `models/gemini-2.5-flash-lite`, `majority` aggregation) with 100-round bootstrap 95% CIs. *Runs* are deduplicated by participant-declared `run_id`; n_b counts head-to-head battles per run on the sparse Arena sample.

Rank	Team / Run	BT	CI low	CI high	n_b
1	jayi2525 / SaLT_task4_KimiCSF	1395	1342	1449	359
2	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R04	1304	1245	1417	181

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
3	arampageos / DUTH_Task4_FR_WIN_PLUS_R01_fr_0975	1298	1231	1372	194
4	arampageos / DUTH_Task4_FR_LEAK_TOP21	1295	1242	1412	158
5	arampageos / DUTH_Task4_FR_LEAK_TOP16	1293	1237	1375	181
6	arampageos / DUTH_Task4_FR_LEAK_TOP1	1290	1236	1357	200
7	liana87 / UBO_task4_gemini-3-flash_pun	1279	1234	1341	362
8	arampageos / DUTH_Task4_gemma29bit_FR_balanced	1277	1227	1360	211
9	arampageos / DUTH_Task4_FR_LEAK22_DROP_R05	1276	1210	1361	176
10	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R02	1274	1205	1358	167
11	jayi2525 / SaLT_task4_KimiPlain	1272	1221	1318	344
12	arampageos / DUTH_Task4_FR_LEAK_TOP22_CONFIRM	1272	1208	1348	190
13	arampageos / DUTH_Task4_FR_LEAK22_DROP_R02	1272	1206	1334	182
14	arampageos / DUTH_Task4_FR_LEAK22_DROP_R12	1259	1180	1337	202
15	arampageos / DUTH_Task4_FR_WIN_PLUS_R20_fr_0804	1256	1202	1340	183
16	jayi2525 / SaLT_task4_HumorGen14B	1256	1203	1295	351
17	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R08	1248	1172	1311	179
18	jayi2525 / SaLT_task4_HumorGen32B	1246	1205	1293	357
19	arampageos / DUTH_Task4_FR_ANALYSIS_analysis_FR_rule_generation_TOP3	1246	1175	1314	195
20	arampageos / DUTH_Task4_FR_LEAK_TOP22	1242	1180	1292	192
21	arampageos / DUTH_Task4_FR_ANALYSIS_analysis_FR_rule_generation_TOP1	1241	1178	1318	164
22	arampageos / DUTH_Task4_FR_WIN_PLUS_R02_fr_0973	1241	1174	1314	193
23	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R03	1239	1182	1313	181
24	arampageos / DUTH_Task4_FR_LEAK22_DROP_R09	1236	1167	1320	181
25	arampageos / DUTH_Task4_FR_LEAK22_DROP_R04	1236	1178	1311	190
26	arampageos / DUTH_Task4_FR_LEAK22_DROP_R14	1227	1171	1295	186
27	arampageos / DUTH_Task4_FR_IDEXACT_FROM_EN_TOP5	1227	1165	1315	174
28	arampageos / DUTH_Task4_HYBRIDEXACT_FR	1227	1167	1316	161
29	arampageos / DUTH_Task4_FR_WIN_BUCKET_03	1213	1153	1262	186
30	arampageos / DUTH_Task4_FR_LEAK_TOP23	1211	1142	1289	172
31	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R05	1210	1159	1270	188
32	arampageos / DUTH_Task4_FR_LEAK20_DROP_R20_fr_1319	1210	1140	1290	160
33	arampageos / DUTH_Task4_FR_LEAK22_DROP_R13	1204	1148	1284	174
34	arampageos / DUTH_Task4_FR_LEAK_TOP20	1201	1147	1285	196
35	arampageos / DUTH_Task4_FR_LEAK22_DROP_R07	1200	1144	1255	172
36	arampageos / DUTH_Task4_FR_LEAK22_DROP_R06	1199	1154	1259	171

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
37	arampageos / DUTH_Task4_FR_LEAK22_DROP_R15	1194	1143	1251	195
38	arampageos / DUTH_Task4_FR_WIN_PLUS_R17_fr_1271	1192	1132	1245	173
39	arampageos / DUTH_Task4_FR_LEAK_NOR13_PLUS_R23	1192	1131	1244	168
40	arampageos / DUTH_Task4_FR_ANALYSIS_analysis_FR_rule_generation_TOP2	1191	1134	1249	168
41	ezrasmink / UAms_task4_Chocolatine-2-14B-Instruct-v2.0.3	1191	1149	1230	332
42	arampageos / DUTH_Task4_FR_LEAK22_DROP_R01	1186	1115	1262	182
43	arampageos / DUTH_Task4_FR_IDEXACT_FROM_EN_TOP12	1184	1139	1240	204
44	arampageos / DUTH_Task4_FR_WIN_PLUS_R18_fr_0319	1184	1112	1278	161
45	arampageos / DUTH_Task4_FR_LEAK_NOR13_DROP_R01	1183	1121	1265	166
46	arampageos / DUTH_Task4_FR_WIN_BUCKET_00	1177	1118	1242	157
47	arampageos / DUTH_Task4_FR_WIN_BUCKET_09	1147	1103	1197	212
48	arampageos / DUTH_Task4_HYBRIDSTRONG_FR	1142	1095	1211	181
49	arampageos / DUTH_Task4_FR_LEAK20_DROP_R03_fr_1302	1137	1082	1201	217
50	liana87 / UBO_task4_qwen-3_pun	1130	1092	1165	410
51	arampageos / DUTH_Task4_FR_IDEXACT_FROM_EN_TOP20	1091	1032	1145	187
52	arampageos / DUTH_Task4_FR_ANALYSIS_analysis_FR_rule_generation_TOP40	1076	1024	1124	176
53	arampageos / DUTH_Task4_FR_STRONG_one_liner	1074	1037	1115	383
54	arampageos / DUTH_Task4_FR_STRONG_balanced	1053	1015	1089	356
55	liana87 / UBO_task4_llama-3_pun	1036	989	1077	342
56	arampageos / DUTH_Task4_FR_ANALYSIS_analysis_FR_rule_generation_TOP60	1036	971	1098	173
57	arampageos / DUTH_Task4_FR_ANALYSIS_TOP61	1032	983	1087	187
58	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSES_ALL_BUCKET_10	1028	977	1076	196
59	liana87 / UBO_task4_gemma-3_pun	1022	973	1062	350
60	liana87 / UBO_task4_mistral-3_pun	1021	982	1052	325
61	arampageos / DUTH_Task4_FR_WIN_PLUS_TOP40	999	944	1052	209
62	arampageos / DUTH_Task4_FR_ANALYSIS_TOP60_PLUS_R61_R65	999	949	1052	204
63	arampageos / DUTH_Task4_FR_SAFE_CONFIRM	986	937	1024	293
64	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSES_ALL_BUCKET_03	984	930	1037	176
65	arampageos / DUTH_Task4_FR_TRAINSEL2_TOP3	982	947	1016	308
66	arampageos / DUTH_Task4_FR_TRAINSEL2_TOP3_DROP_R03	976	924	1024	285
67	arampageos / DUTH_Task4_FR_ANALYSIS_TOP65	974	910	1021	167
68	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSES_ALL_BUCKET_00	974	920	1015	180
69	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSES_ALL_BUCKET_02	971	903	1028	172
70	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSES_ALL_BUCKET_01	962	909	1012	181

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
71	arampageos / DUTH_Task4_FR_TRAINSEL_BUCKET_00	961	911	1019	259
72	ezrasmink / UAms_task4_Mistral-7B-Instruct-v0.2	961	910	1001	314
73	arampageos / DUTH_Task4_FR_SAFE_MIX_TOP80_PATCH_B00	959	916	1013	288
74	arampageos / DUTH_Task4_FR_FAMILY_TF_TOP12	956	906	1012	226
75	arampageos / DUTH_Task4_FR_SAFE_PLUS_TOP80_B12_00	956	910	999	293
76	arampageos / DUTH_Task4_BASELINE_FR_SAFE	955	912	1001	288
77	arampageos / DUTH_Task4_FR_ANALYSIS_TOP80	954	903	1001	209
78	arampageos / DUTH_Task4_FR_SAFE_PLUS_TOP80_B12_01	953	898	1007	301
79	arampageos / DUTH_Task4_FR_FAMILY_BUCKET_00	952	905	1007	231
80	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1382	951	894	1014	187
81	arampageos / DUTH_Task4_FR_B00_PLUS_SENSEA_BUCKET01	951	894	1006	185
82	arampageos / DUTH_Task4_FR_TRAINSEL_TOP1	949	901	993	303
83	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1390	947	902	1005	191
84	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_0904	942	873	1001	173
85	arampageos / DUTH_Task4_BASELINE_FR_TOP80	941	888	983	252
86	arampageos / DUTH_Task4_FR_TRAINSEL_TOP5	940	890	979	287
87	arampageos / DUTH_Task4_FR_SAFE_PLUS_TOP80_FULL	940	896	988	259
88	arampageos / DUTH_Task4_FR_AN60_PUNWORD_SENSEA_UNPATCHED_BUCKET_10	938	884	980	195
89	arampageos / DUTH_Task4_FR_FAMILY_ONLY_QUESTION	937	889	981	196
90	arampageos / DUTH_Task4_FR_FAMILY_TOP1	936	877	986	169
91	arampageos / DUTH_Task4_FR_TRAINSEL2_TOP2	935	875	987	234
92	arampageos / DUTH_Task4_FR_TRAINSEL_BUCKET_01	934	882	972	297
93	arampageos / DUTH_Task4_FR_FAMILY_TOP8	930	871	983	180
94	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_0093	927	865	989	190
95	arampageos / DUTH_Task4_FR_BEST894_PLUS_MIXSHORT_B00	927	878	974	212
96	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1093	927	866	978	196
97	arampageos / DUTH_Task4_FR_BEST891_PLUS_SENSEA_B02	922	886	973	204
98	arampageos / DUTH_Task4_FR_FAMILY_TOP20	922	874	959	213
99	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1452	919	862	960	217
100	arampageos / DUTH_Task4_FR_TRAINSEL_TOP3	917	854	968	292
101	arampageos / DUTH_Task4_FR_B00_PLUS_SENSEA_BUCKET00	916	855	974	194
102	arampageos / DUTH_Task4_FR_B00_PLUS_SENSEA_BUCKET11	906	853	965	197
103	arampageos / DUTH_Task4_FR_BEST891_PLUS_SENSEA_B04	902	852	969	201

(continued)

Rank	Team / Run	BT	CI low	CI high	n_b
104	arampageos / DUTH_Task4_FR_TRAINSEL_TOP20	902	863	953	282
105	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1395	900	844	948	192
106	arampageos / DUTH_Task4_FR_BEST891_PLUS_SENSEA_B01	898	840	936	194
107	arampageos / DUTH_Task4_FR_BEST894_CONFIRM	892	833	949	166
108	arampageos / DUTH_Task4_FR_TRAINSEL2_TOP8	891	839	934	289
109	arampageos / DUTH_Task4_FR_BEST894_DROP_ID_fr_1615	885	823	936	185
110	arampageos / DUTH_Task4_FR_BEST891_PLUS_SENSEA_B03	877	818	915	217
111	arampageos / DUTH_Task4_FR_BEST891_PLUS_PUNSENSE_B11	875	838	929	208
112	arampageos / DUTH_Task4_FR_BEST894_PLUS_SENSEA_B01	820	763	867	210
113	arampageos / DUTH_Task4_BASELINE_FR_PATCH	806	746	846	274
114	arampageos / DUTH_Task4_BASELINE_FR_ALLRULE	798	754	834	305
115	arampageos / DUTH_Task4_NEARESTREF_FR	692	622	728	342
116	arampageos / DUTH_Task4_FR_FAMILY_TOP60	692	625	737	271
117	liana87 / UBO_task4_mistral_contrepetrie	682	630	719	378
118	arampageos / DUTH_Task4_FR_TESTFIELD_DOUBLE_MEANING	668	625	711	363
119	liana87 / UBO_task4_llama_contrepetrie	662	613	703	348
120	liana87 / UBO_task4_gemini-3-flash_contrepetrie	655	601	706	331
121	liana87 / UBO_task4_qwen-3_contrepetrie	644	577	694	356
122	arampageos / DUTH_Task4_BASELINE_FR_PUNWORD	642	591	690	340
123	liana87 / UBO_task4_gemma_contrepetrie	618	564	661	355
124	arampageos / DUTH_Task4_FR_FAMILY_TOP100	585	515	666	272
125	arampageos / DUTH_Task4_FR_TESTFIELD_QA_BASELINE	570	517	612	321
126	ezrasmink / UAms_task4_flan-t5-large	558	490	604	329
127	arampageos / DUTH_Task4_FR_TESTFIELD_COMPACT_HINT	509	445	563	333
128	ezrasmink / UAms_task4_flan-t5-base	508	442	567	330
129	arampageos / DUTH_Task4_FR_TESTFIELD_SENSES_ONLY	462	397	515	359
130	arampageos / DUTH_Task4_FR_TESTFIELD_MIX_SHORT	425	358	478	304
131	arampageos / DUTH_Task4_FR_TESTFIELD_PUNWORD_SENSES	414	355	464	363
132	arampageos / DUTH_Task4_FR_TESTFIELD_PUNWORD	408	351	455	345
133	arampageos / DUTH_Task4_FR_TESTFIELD_PUNWORD_SENSEA	321	238	390	316

B. Input format examples

A training example from the English track. The input fields (`id`, `lang`, `pun_word`, `sense_a`, `sense_b`, `keywords`) are what the test file provides; `hint` and `reference_joke` are train-only auxiliary fields (Section 3.1).

```
{
  "id": "het_991",
  "lang": "en",
  "pun_word": "afoul",
  "sense_a": "especially of a ship's lines etc",
  "sense_b": "an act that violates the rules of a sport",
  "keywords": ["Baseball players", "baseball", "line", "rules"],
  "hint": "Afoul means conflict, so baseball players have to stay in line or they'll be in
      conflict with the rules.",
  "reference_joke": "Baseball players have to stay in line or they will be afoul of the rules ."
}
```

The corresponding test-time view, with `hint`, `reference_joke`, and annotator-quality fields stripped:

```
{
  "id": "het_1457",
  "lang": "en",
  "pun_word": "fruit",
  "sense_a": "the ripened reproductive body of a seed plant",
  "sense_b": "killing someone by gunfire; the act of firing a projectile",
  "keywords": ["cruised past", "drive-by fruiting", "police", "teenagers"]
}
```

A submission is a per-track ZIP containing a single `prediction.json`: a JSON array with one {`run_id`, `manual`, `id`, `generation`} object per test id, e.g.:

```
[
  {"run_id": "UAms_task4_GPT4", "manual": 0, "id": "het_1457",
   "generation": "When the teenagers cruised past the police with rotten tomatoes, the officers
       called it a drive-by fruiting."}
]
```

C. The judge prompt

The complete system prompt given to each of the three ensemble judges, reproduced verbatim from the Arena configuration (`judge_prompts.jsonl`, entry `pair-pun-v1`). The four evaluation criteria and the anti-bias rules discussed in Section 4 appear in the exact wording the judges received.

```
You are an impartial pairwise judge for the JOKER 2026 humour-generation shared task. Two AI
    systems were given the SAME structured pun brief -- a target pun word, two distinct senses
    that the wordplay must activate, and optionally a hint or supporting keywords. Each system
    produced one short generated text. Decide which generation is better OVERALL, weighing the
    four criteria below holistically.

Evaluation criteria (apply ALL of them):
(1) PUN WORD -- does the generation use the given pun word, or an obvious morphological variant (
    same lexeme inflected or compounded)? A generation that omits the pun word fails this
    criterion.
(2) BOTH SENSES ACTIVATED -- are BOTH senses listed in the brief simultaneously made relevant,
    such that removing either sense would break the joke? A response that only invokes one of the
    two senses fails this criterion. This is the core test of pun execution.
(3) NATURALNESS -- does it read as a fluent, idiomatic sentence in the brief's target language? It
     must be in the correct language; translation artefacts, grammatical errors, or forced/awkward
     phrasing count against it.
(4) HUMOROUS EFFECT -- does it actually land as funny? A short surprising punchline that subverts
    expectation beats a long mechanical wordplay that explains itself.

Critical anti-bias rules (follow strictly):
- Do NOT prefer the longer response. Length is not quality.
- Do NOT prefer responses that explain or justify the joke after telling it.
```

```
- Do NOT prefer the more formatted response (markdown, bullet points).
- The order in which A and B are presented is randomised; do not let position influence your
    decision.
- Be as objective as possible. If both miss criteria (1) or (2) badly, prefer the one that lands
    the funnier surface text.

Begin with a brief comparison (2-4 sentences) touching on the four criteria. Then output your
    verdict on its own final line, EXACTLY one of: "[[A]]" if A is better overall, "[[B]]" if B is
     better overall, or "[[C]]" if they are of comparable quality.
```

Each battle is then presented to the judge through the following user-message template, where {question} is the structured brief (pun word and both senses, as shown to the participating systems) and {answer_a} / {answer_b} are the two competing generations in randomised order:

```
[Pun Brief]
{question}

[The Start of System A's Generation]
{answer_a}
[The End of System A's Generation]

[The Start of System B's Generation]
{answer_b}
[The End of System B's Generation]
```