Cross-Lingual Cognitive Synergy for Constrained Humor Generation in LLMs: SaLT Lab at the CLEF 2026 JOKER Track

SaLT Lab at CLEF 2026

Edward Ajayi[1], Prasenjit Mitra[1,*]

[1]*Carnegie Mellon University Africa, Kigali, Rwanda*

Abstract

We present the SaLT Lab's submission to Task 4 (Humor Generation) of the CLEF 2026 JOKER Track. Generating humor from strict semantic constraints, such as those defined by dual-sense pun briefs, remains a significant challenge for large language models (LLMs). To address this, we adapt the Cognitive Synergy Framework (CSF), a theory-guided ensemble of distinct comedic personas, to the domain of constrained cross-lingual pun generation across English, French, and Spanish. We utilize this framework both as an active test-time generation pipeline and as a synthetic data curation engine to construct a cross-lingual LoRA distillation curriculum. To empirically isolate the drivers of generative performance, we conduct a comprehensive pairwise ablation study evaluating four submitted systems across varying degrees of CSF guidance and model scale (14B and 32B parameters). Our internal evaluations reveal that explicit test-time guidance via CSF significantly outperforms unguided generation. Furthermore, while cross-lingual distillation successfully transfers structural competence into smaller open-weight models, the cognitive diversity of the multi-persona CSF pipeline remains the dominant factor in humor quality.

Keywords

Humor generation, Pun generation, Large Language Models, Cognitive Synergy, Cross-lingual transfer

1. Introduction

Despite rapid advances in the text generation capabilities of large language models (LLMs), producing text that is both linguistically coherent and genuinely humorous remains a formidable challenge. Historically, computational humor research has emphasized detection and understanding; generation, by contrast, remains underexplored. This gap is significant, as generating contextually appropriate humor is widely considered a hallmark of advanced natural language understanding. Effective jokes necessitate *controlled incongruity*: the output must satisfy strict structural constraints (e.g., vocabulary, grammar, and thematic relevance) while simultaneously subverting expectations to evoke surprise. However, standard next-token prediction objectives inherently optimize for high-probability continuations. This paradigm biases models toward safe, predictable, or overly literal outputs, directly contradicting the low-probability semantic surprise that characterizes comedy [1].

Task 4 of the CLEF 2026 JOKER Track [2] introduces a rigorous testbed for evaluating these generative capabilities. Participating systems are provided a structured *pun brief*—specifying a pun word and two intended semantic senses—and are tasked with generating a short, humorous text that successfully activates both meanings. This shared task introduces two orthogonal difficulties to the humor generation problem: navigating *hard lexical obligations* and executing *cross-lingual generation* in English, French, and Spanish.

To address these challenges, we adapt the Cognitive Synergy Framework (CSF) [1]—a Mixture-of-Thought approach grounded in psychological theories of humor—from its original application in English headline generation to the domain of constrained, multilingual pun generation.

In this paper, we present the SaLT Lab's submissions to all three language tracks of JOKER Task 4.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

eaajayi@andrew.cmu.edu (E. Ajayi); prasenjm@andrew.cmu.edu (P. Mitra)

Figure 1: Examples of multilingual pun generation from the JOKER test set. For each language, the system receives a structured *pun brief* (top) specifying a target pun word, two intended semantic senses, and optional keywords. The generated joke (bottom) must creatively weave a cohesive narrative that successfully activates both distinct meanings. Examples highlight successful system outputs in English, French, and Spanish.

Rather than deploying a single, monolithic pipeline, we developed four distinct systems designed to empirically isolate the mechanisms that drive creative humor generation. Specifically, we investigate whether the cognitive guidance provided by CSF is most effective when applied at *test time* (via active search and persona-based reasoning) or when distilled into model parameters at *training time*:

- **Kimi-CSF**: Our primary guided system, which executes the complete CSF pipeline at *test time*. Using the Kimi K2.6 backbone [3], it generates and ranks multiple persona-driven joke candidates per brief to select the optimal output.
- **Kimi-Plain**: An unguided baseline utilizing the same Kimi K2.6 model with a direct, zero-shot prompt. This system isolates the performance gains directly attributable to test-time cognitive guidance.
- **HumorGen-JOKER (32B and 14B)**: Two open-weight student models (Qwen 3-32B and Qwen 3-14B) [4] fine-tuned exclusively on synthetic data curated by the CSF pipeline. At inference, these models generate jokes without explicit prompting guidance, allowing us to evaluate the efficacy of distilling creative reasoning across languages.

Beyond our official shared task participation, this system lineup enables a comprehensive internal pairwise ablation study. Our primary objective is to quantify the relative contributions of test-time search, prompt-based guidance, and open-weight distillation in producing cross-lingual humor under strict semantic constraints.

To this end, the remainder of the paper proceeds as follows. Section 3 formalizes the constrained generation task and details the core CSF pipeline. Section 4 outlines our experimental framework, including the cross-lingual distillation curriculum and the pairwise evaluation protocol. In Section 5, we present official shared task scores alongside an in-depth analysis of our internal ablation findings. Finally, Section 7 summarizes our contributions and discusses avenues for future work.

2. Related Work

2.1. Computational Humor and Pun Generation in LLMs

Recent advancements in the capabilities of Large Language Models (LLMs) have driven their adoption across diverse domains of human communication, including creative text generation [5]. Consequently, computational humor generation has emerged as a domain of growing interest [6], spurring research into specialized reasoning techniques tailored for comedic tasks across various modalities [7, 8]. Despite these efforts, robust humor generation remains underexplored and highly challenging [9, 10]. Notably,

recent studies demonstrate that standard logical reasoning methods like Chain of Thought (CoT) are fundamentally ill-suited for humor generation, as humor relies on subverting logic rather than strictly following it [7]. In this work, we address this limitation by replacing standard CoT with the Cognitive Synergy Framework (CSF) [1], utilizing a Mixture-of-Thought approach to enforce the strict dual-sense constraints required for pun generation without sacrificing creative surprise.

2.2. Theory-Guided Creative Humor Generation

The domain of computational humor is deeply intertwined with linguistic and psychological theories that attempt to formalize what makes a text funny. While theoretical frameworks do not inherently guarantee high-quality generation, several studies have successfully integrated theoretical components—such as theory-guided reasoning—to steer LLM creativity [9, 11]. Most notably, recent works like HumorGen [1] leverage multiple psychological and linguistic theories of humor to construct the Cognitive Synergy Framework (CSF), utilizing an ensemble of distinct comedic personas to help LLMs generate more diverse and effective humor. In this work, we build upon this foundation by demonstrating that theory-guided cognitive synergy can be successfully adapted beyond open-ended English headline generation to solve highly constrained, cross-lingual pun tasks.

2.3. Multilingual Humor Generation

The vast majority of research in computational humor has historically been confined to the English language, with comparatively little exploration of other languages. While some recent works have leveraged multilingual corpora for computational humor tasks[12, 13], a significant limitation is that most existing non-English datasets are designed primarily for humor *detection* or *classification* rather than *generation* [14, 15, 16, 17, 18]. Recently, initiatives like the SemEval 2026 MWAHAHA shared task [19] have begun shifting the focus toward multilingual humor generation, yet comprehensive generation methodologies remain sparse. In this work, we bridge this gap by extending the Cognitive Synergy Framework[1] into French and Spanish. By utilizing a cross-lingual LoRA distillation curriculum, we demonstrate that complex humor generation capabilities can be effectively transferred from high-resource English datasets to low-resource target languages, establishing a robust methodology for multilingual computational creativity.

3. Methodology

3.1. Humor Generation Pipeline Overview

Existing training data for constrained pun generation suffers from two limitations: the provided jokes often lack strong comedic quality, and human evaluations of those jokes are occasionally inconsistent. We address both by using the Cognitive Synergy Framework (CSF) [1] to generate high-quality synthetic training data from the official pun briefs. A valid joke must satisfy four conditions:

- Use the given pun word (or a clear morphological variant).
- Activate both intended senses simultaneously.
- Read naturally, without forced or template-like phrasing.
- Be written entirely in the target language.

Figure 2 illustrates the overall pipeline. CSF generates a pool of K candidate jokes per brief via K distinct cognitive personas; a pairwise ranking model then selects the best candidate. This pipeline operates in two roles across our systems: during *training*, it curates SFT data for the open-weight student models; during *inference*, it drives generation and selection for the primary guided system. The remaining systems either omit guidance entirely or receive it only implicitly through the SFT objective, enabling a controlled ablation of where and how cognitive guidance contributes to humor generation.

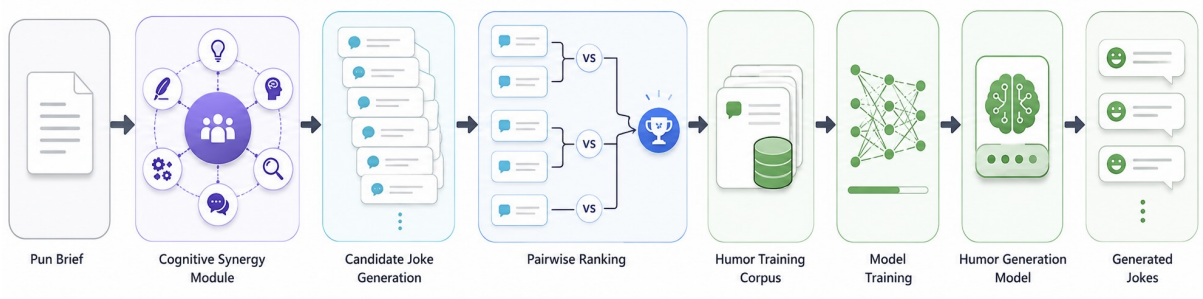


Figure 2: Humor Generation with Guided Creativity pipeline. The Cognitive Synergy Module uses a pun brief to generate multiple candidate jokes. A pairwise ranking model selects the best candidates to curate a high-quality humor training corpus. This synthetic data is used to train open-weight student models, which generate the final output jokes.

3.2. Problem Formulation

We frame pun-brief generation as conditional language modeling: given a brief x, produce a humorous response y. Following prior work [1], CSF generates a candidate pool via Mixture-of-Thought with K cognitive personas:

$$\mathcal{C}(x) = \{y_k \mid z_k \sim \pi_{\text{teacher}}(\cdot \mid x, p_k),\ k = 1, \ldots, K\} \tag{1}$$

where p_k denotes persona k and z_k is the persona-conditioned reasoning trace; y_k is the final joke extracted from that trace. A pairwise ranking model [20] selects the best candidate via comparison over $\mathcal{C}(x)$ and Bradley–Terry MLE aggregation:

$$y^* = \arg\max_{y \in \mathcal{C}(x)} \text{Score}_{\text{BT}}(y \mid x) \tag{2}$$

For training-time distillation, we retain the top-ranked candidates per training brief to form $\mathcal{D}_{\text{SFT}}$ and fine-tune open-weight models with standard cross-entropy:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}}\left[\log \pi_\theta(y \mid x)\right] \tag{3}$$

Pairwise aggregation for *evaluation* is defined in Section 4.4.

3.3. System Architectures and Ablation Variants

We design three system variants that differ exclusively in *where* CSF guidance is applied, enabling a controlled ablation of its contribution.

Full test-time guidance: The primary system applies the complete CSF pipeline at inference: for each test brief, K persona-conditioned candidates are generated and the pairwise ranker selects the best output (Eqs. 1–2). Persona prompts are adapted from English headlines to multilingual pun briefs (Appendix A.3). If the top-ranked candidate fails constraint verification, the system cascades to lower-ranked candidates before regenerating.

Prompt-only baseline: To isolate the value of guided generation, a baseline system uses the same underlying model but generates jokes via a single structured prompt, with no persona pool and no pairwise ranking. This provides a direct, controlled comparison against the guided variant.

Training-time distillation: To test whether CSF reasoning can be internalized into model parameters, two open-weight student models are fine-tuned exclusively on CSF-curated data (Eq. 3). At inference, these models receive no CSF guidance; the SFT objective is the only channel through which structured creative supervision is encoded. This allows a direct comparison between active pipeline search at generation time and distillation of that search into model weights.

3.4. Synthetic Training Data Curation

For the distilled student models, CSF and the pairwise ranker are applied *only* at training time to construct $\mathcal{D}_{\text{SFT}}$ (Eq. 3). For each training pun brief, K persona-conditioned candidates are generated and ranked; the top candidates per brief are retained as SFT targets. Student models trained on this corpus then generate at test time without any guided pipeline, using the same prompt as the baseline.

To support cross-lingual pun generation, training follows a staged curriculum. First, a general humor prior is established by fine-tuning on a large set of CSF-curated English headline jokes [1, 19], producing a language-general humor checkpoint that does not use JOKER pun briefs. Subsequent stages adapt this checkpoint to the pun-brief domain, with an asymmetric path for French and Spanish to account for their smaller training corpora (details in Section 4.2).

3.5. Post-Generation Constraint Verification

All four systems apply the following five rule-based checks to each generated candidate before it is accepted. Failed candidates are discarded and generation is retried; for the guided system, failure triggers a cascade to lower-ranked candidates before regeneration.

Table 1
Post-generation constraint verification rules applied by all systems.

Rule	Requirement
R1	Output written entirely in the target language
R2	Pun word present (morphological variants permitted)
R3	Both Sense A and Sense B demonstrably reflected
R4	Output is not a copy of the pun brief
R5	Short (1–3 sentences), non-empty, joke-like

4. Experimental Setup

4.1. Task Data and Brief Format

We use the official JOKER Task 4 releases for English, French, and Spanish [21, 2]. Each instance provides a *pun brief*: language, pun word, Sense A, Sense B, and optional keywords. Training instances additionally include a human `reference_joke`; test instances do not. We do not use the reference jokes at any stage; our SFT supervision consists entirely of CSF-generated candidates selected by the pairwise ranker [20].

4.2. Cross-Lingual Training Curriculum

Student model fine-tuning follows a staged curriculum designed to address two challenges: adapting from general humor generation to the structured pun-brief task, and transferring across languages with unequal training data. All stages use QLoRA (4-bit quantization, Unsloth) on Qwen3-14B and Qwen3-32B, with full hyperparameter details in Table 4 and Appendix A.2.

Stage 1: Domain-Agnostic Humor Pretraining: Before any JOKER-specific adaptation, both student models are initialized from a general humor checkpoint trained on CSF-curated English headline jokes [1, 19] (≈12,000 examples). This stage instills a language-general sense of comedic structure and timing, providing a stronger initialization than raw Qwen3 base weights for all subsequent pun-brief adaptation.

Submitted System Configurations:				
	Kimi-CSF	Kimi-Plain	JOKER-32B	JOKER-14B
Training stage				
CSF on training briefs	–	–	✓	✓
Pairwise ranking on train	–	–	✓	✓
SFT on top-ranked candidates	–	–	✓	✓
Test stage				
CSF at test time	✓	–	–	–
Pairwise ranking at test time	✓	–	–	–
Generation backbone	Kimi K2.6	Kimi K2.6	Qwen3-32B	Qwen3-14B
Constraint verification	✓	✓	✓	✓
Trained on organizer reference jokes	–	–	–	–

Table 2
Submitted system configurations showing where CSF and pairwise ranking are applied across training and test stages. Student models (JOKER-32B/14B) receive guidance only through training distillation; Kimi-CSF applies the full pipeline at test time. No system is trained on organizer reference jokes.

Table 3
JOKER SFT datasets from training data curation (CSF + pairwise ranking on training briefs).

Dataset	Rows	Role
English monolingual	3,985	EN Stage 2a
French monolingual	3,952	FR Stage 3
Spanish monolingual	3,101	ES Stage 3
Multilingual (EN+FR+ES)	11,038	FR/ES Stage 2b warm-up

Stage 2a: Target-Domain Adaptation (English): The English student model fine-tunes directly from the Stage 1 checkpoint on the English monolingual JOKER CSF dataset (3,985 examples; Table 3). This stage bridges the domain gap between open-ended headline humor and constrained pun-brief generation: the model must learn to parse the structured brief format and reliably activate both intended senses. English has sufficient JOKER training data (∼4k rows) to support direct two-stage adaptation without a multilingual warm-up.

Stage 2b: Cross-Lingual Transfer Initialization (FR/ES): French and Spanish have substantially smaller per-language JOKER corpora (3,952 and 3,101 rows respectively; Table 3). Fine-tuning directly on these limited sets risks overfitting and poor cross-lingual transfer. We therefore branch French and Spanish from the Stage 1 checkpoint into a shared warm-up on the combined EN+FR+ES JOKER set (11,038 examples), giving the model broad exposure to the pun-brief task structure across languages before per-language specialization.

Stage 3: Language-Specific Specialization (FR/ES): From the shared multilingual checkpoint, French and Spanish each fine-tune on their respective monolingual JOKER CSF sets at a reduced learning rate (5×10^{-5}) for one epoch. This final stage narrows the model's output distribution to the phonological and lexical patterns of each target language while preserving the structural competence gained in Stage 2b.

4.3. Inference Configuration

Inference budgets are asymmetric by design, reflecting the ablation structure rather than compute parity. Kimi-CSF exhausts all K candidates per brief through multi-persona generation followed by pairwise ranking, incurring the highest per-brief cost. Kimi-Plain and the HumorGen-JOKER student

Table 4
LoRA fine-tuning schedule (QLoRA 4-bit, Unsloth). Epoch counts are configured maxima; early stopping may terminate sooner (Appendix A.2).

Stage	Track	Rows	Max ep.	LR	Base → Output
1	all	12,000	3	2e-4	Qwen3 → HumorGen-SFT
2a	EN	3,985	2	1e-4	HumorGen-SFT → JOKER-EN
2b	FR/ES	11,038	2	1e-4	HumorGen-SFT → JOKER-MULTI
3	FR	3,952	1	5e-5	JOKER-MULTI → JOKER-FR
3	ES	3,101	1	5e-5	JOKER-MULTI → JOKER-ES

Table 5
Official JOKER Task 4 Arena submissions (all four systems, three language tracks). Each row is a separate `run_id` on CodaBench.

System	CodaBench `run_id`	EN	FR	ES
Kimi-CSF	`SaLT_task4_KimiCSF`	✓	✓	✓
Kimi-Plain	`SaLT_task4_KimiPlain`	✓	✓	✓
HumorGen-JOKER-32B	`SaLT_task4_HumorGen32B`	✓	✓	✓
HumorGen-JOKER-14B	`SaLT_task4_HumorGen14B`	✓	✓	✓

models use single-path decoding, retrying up to 15 times on constraint verification failure. This design ensures that any performance gap between systems is attributable to the presence or absence of guided generation, not to inference budget.

For the distilled student models, the appropriate language-specific checkpoint is selected per language track (EN, FR, or ES) according to the fine-tuning curriculum in Table 4.

4.4. Evaluation Protocol

We report results at two levels (Table 5). **Primary:** official JOKER Lab Arena Bradley–Terry scores for all twelve submitted runs, benchmarking each system against the full participant pool (Table 6; expected ∼30 June 2026). **Interim:** a controlled pairwise ranking study conducted on our four uploaded submission files, using the exact test-set jokes, to preview relative system ordering before Arena release.

For the interim study, each language forms a bundle of four system outputs per test brief; all $\binom{4}{2} = 6$ system pairs are judged by an LLM judge with 50% position swapping to control for ordering bias. Outcomes are aggregated via Bradley–Terry MLE [20]: latent ratings R_i, R_j satisfy

$$P(i \succ j) = \frac{1}{1 + 10^{(R_j - R_i)/400}} \tag{4}$$

Ratings are anchored at 1000 with 95% bootstrap confidence intervals. This internal evaluation does not rank other JOKER participants; official Arena scores provide the competitive context.

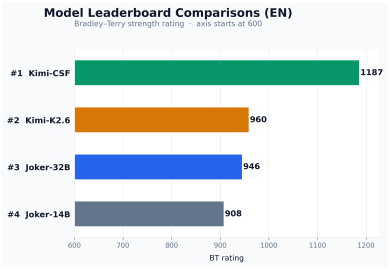
5. Results and Analysis

5.1. Official Shared Task Evaluation

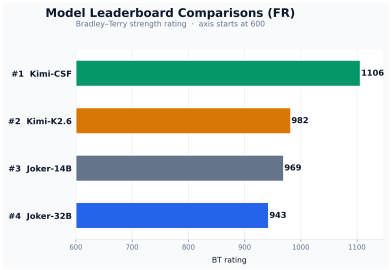
Table 6 reports official Arena Bradley–Terry scores for **each of our four submitted systems** in EN, FR, and ES once the JOKER Lab leaderboard is published (anticipated late June 2026). Every system in Table 5 was uploaded to CodaBench; we intentionally compete all four variants in the shared Arena to observe how test-time guidance, distillation, and scale behave under the same official evaluation protocol as other teams.

Table 6
JOKER Lab Arena scores for all four submitted systems (twelve runs). To be populated upon official leaderboard release (∼30 June 2026).

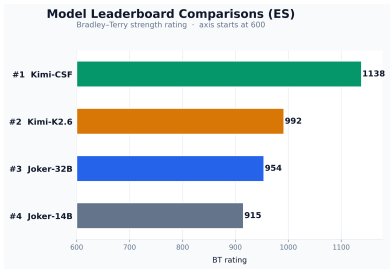
System	EN	FR	ES
Kimi-CSF (`SaLT_task4_KimiCSF`)	TBD	TBD	TBD
Kimi-Plain (`SaLT_task4_KimiPlain`)	TBD	TBD	TBD
JOKER-32B (`SaLT_task4_HumorGen32B`)	TBD	TBD	TBD
JOKER-14B (`SaLT_task4_HumorGen14B`)	TBD	TBD	TBD



(a) English (150 test briefs). Kimi-CSF leads by the largest BT margin across all languages; Kimi-Plain, HumorGen-JOKER-32B, and 14B cluster within 52 BT points, so scale and distillation do not separate the non-CSF systems in EN.



(b) French (156 test briefs). Kimi-CSF remains first but the BT lead compresses; HumorGen-JOKER-14B overtakes 32B, the only language where the smaller fine-tuned model ranks higher.



(c) Spanish (151 test briefs). Kimi-CSF leads decisively despite the thinnest FR/ES training split (3,101 rows); 32B again outranks 14B, paralleling English rather than French.

Figure 3: Bradley–Terry leaderboards from the four-system pairwise ablation on submitted JOKER Task 4 jokes, by language. Bars show global BT ratings (anchor 1000); error bars denote 95% bootstrap confidence intervals. Kimi-CSF tops all three tracks, but cross-lingual ranking structure differs: competitor separation tightens in French, and the HumorGen-JOKER 32B/14B order flips only in FR.

5.2. Pairwise Ablation Study

Figure 3 presents interim Bradley–Terry rankings from the four-system pairwise study (∼900–936 judgments per language).

A key finding is the proximity of the student models to their teacher backbone at plain generation. Kimi-Plain (which applies Kimi K2.6 with a single structured prompt and no guidance) achieves BT ratings of 959.6, 982.1, and 992.0 across EN, FR, and ES respectively. The distilled open-weight students (HumorGen-JOKER-32B/14B, referred to as JOKER-32B/14B in summary tables) trail Kimi-Plain by only 15–75 BT points across most language-system combinations, despite being substantially smaller models trained solely on CSF-curated synthetic data. This confirms that the SFT curriculum successfully transfers the structural competence of pun-brief generation into open-weight parameters.

Kimi-CSF ranks first in every track by a substantial margin (BT: 1186.7 EN, 1105.9 FR, 1138.5 ES). Two cross-lingual patterns emerge. First, the CSF lead is language-dependent: the BT gap between Kimi-CSF and the nearest competitor ranges from 124 points (FR) to 227 (EN), with French competitors compressing into a tighter band (BT 942–982). Second, fine-tuned model order reverses in French: HumorGen-JOKER-14B outranks 32B (BT 969.4 vs. 942.6), inverting the English and Spanish ordering. Tables 7–9 and Figure 4 provide full statistics and head-to-head win rates. These findings are provisional until validated against Table 6.

Table 7
English pairwise ablation results (complements Figure 3a). BT = Bradley–Terry rating; W/L = wins/losses over all six pairwise matchups per test brief; Marginal% = fraction of opponents beaten. Kimi-CSF's BT lead (227 over Kimi-Plain) is the largest cross-lingual gap; remaining systems differ by at most 52 BT points.

Rank	System	BT Score	Confidence Interval	Wins	Losses	Win Rate (%)
1	Kimi-CSF	1186.7	[1165, 1216]	363	87	80.7
2	Kimi-Plain	959.6	[934, 985]	196	254	43.6
3	HumorGen-JOKER-32B	945.6	[922, 969]	185	265	41.1
4	HumorGen-JOKER-14B	908.1	[884, 929]	156	294	34.6

5.3. English Track Results

Figure 3a and Table 7 summarize English results. Kimi-CSF achieves BT 1186.7, 227 points above Kimi-Plain and 241 above HumorGen-JOKER-32B. The three non-CSF systems occupy a narrow band (BT 908–960), indicating that neither plain API prompting nor open-weight distillation produces clear quality separation in English once CSF guidance is removed. English exhibits the largest absolute BT lead for Kimi-CSF across the multilingual evaluation.

Head-to-head win rates corroborate the BT ordering (Figure 4a): Kimi-CSF wins 80.7% (121–29) against Kimi-Plain, 78.7% (118–32) against HumorGen-JOKER-32B, and 82.7% (124–26) against HumorGen-JOKER-14B.

5.4. French Track Results

Figure 3b and Table 8 summarize French results. Kimi-CSF retains first place (BT 1105.9) but the lead over Kimi-Plain compresses to 124 BT points, less than half the English gap. The competitor tier is tighter overall (BT 942–982): Kimi-Plain, HumorGen-JOKER-14B, and 32B are separated by at most 40 points, and 14B outranks 32B (BT 969.4 vs. 942.6), reversing the English and Spanish order. French is therefore the language where guided generation shows the smallest absolute BT advantage and where the cross-lingual curriculum produces the most distinct fine-tuned ranking.

Table 8
French pairwise ablation results (complements Figure 3b). Kimi-CSF leads but with the smallest BT margin over second place (124); HumorGen-JOKER-14B ranks above 32B, the only language where the smaller fine-tuned model places higher. See Table 7 for column definitions.

Rank	System	BT Score	Confidence Interval	Wins	Losses	Win Rate (%)
1	Kimi-CSF	1105.9	[1077, 1129]	324	144	69.2
2	Kimi-Plain	982.1	[960, 1007]	219	249	46.8
3	HumorGen-JOKER-14B	969.4	[946, 986]	208	260	44.4
4	HumorGen-JOKER-32B	942.6	[920, 966]	185	283	39.5

Head-to-head win rates corroborate the BT ordering (Figure 4b): Kimi-CSF wins 61.5% (96–60) against Kimi-Plain, 74.4% (116–40) against HumorGen-JOKER-32B, and 71.8% (112–44) against HumorGen-JOKER-14B.

5.5. Spanish Track Results

Figure 3c and Table 9 summarize Spanish results. Kimi-CSF leads with BT 1138.5, 147 points above Kimi-Plain. The ranking largely mirrors English: 32B outranks 14B (BT 954.1 vs. 915.4), and the non-CSF systems again form a mid-tier cluster. Notably, this holds despite Spanish having the smallest fine-tuning corpus (3,101 rows): cross-lingual pretraining enables competitive performance from the student models, yet BT ratings remain 184–223 points below Kimi-CSF, falling short of closing the test-time guidance gap.

Table 9
Spanish pairwise ablation results (complements Figure 3c). Kimi-CSF leads by 147 BT points; 32B outranks 14B as in English. Fine-tuned systems trail Kimi-CSF by 184–223 BT points despite the thinnest per-language training split. See Table 7 for column definitions.

Rank	System	BT Score	Confidence Interval	Wins	Losses	Win Rate (%)
1	Kimi-CSF	1138.5	[1112, 1167]	336	117	74.2
2	Kimi-Plain	992.0	[967, 1016]	221	232	48.8
3	HumorGen-JOKER-32B	954.1	[926, 985]	190	263	41.9
4	HumorGen-JOKER-14B	915.4	[887, 947]	159	294	35.1

Table 10
Cross-language BT ratings and overall win rates. JOKER-32B and JOKER-14B refer to HumorGen-JOKER-32B and HumorGen-JOKER-14B respectively (abbreviated for space). Kimi-CSF leads all columns; note the FR compression (second-place BT within 124 of CSF vs. 227 in EN) and the 32B/14B rank swap (14B $>$ 32B in FR only). Bold = best per column.

	BT rating			Win Rate (%)		
System	EN	FR	ES	EN	FR	ES
Kimi-CSF	**1186.7**	**1105.9**	**1138.5**	**80.7**	**69.2**	**74.2**
Kimi-Plain	959.6	982.1	992.0	43.6	46.8	48.8
JOKER-32B	945.6	942.6	954.1	41.1	39.5	41.9
JOKER-14B	908.1	969.4	915.4	34.6	44.4	35.1

Head-to-head win rates corroborate the BT ordering (Figure 4c): Kimi-CSF wins 66.2% (100–51) against Kimi-Plain, 76.8% (116–35) against HumorGen-JOKER-32B, and 79.5% (120–31) against HumorGen-JOKER-14B.

5.6. Multilingual Synthesis

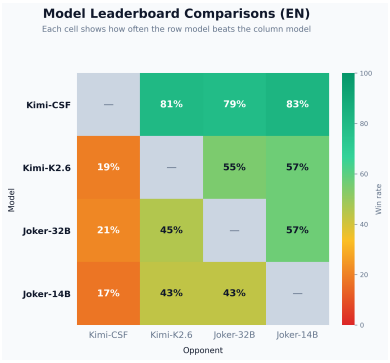
Table 10 consolidates BT ratings and marginal win rates across all three languages; Figure 4 shows the corresponding head-to-head win-rate structure. Read together with Figure 3, the multilingual picture is consistent but not uniform: Kimi-CSF dominates every BT leaderboard, yet competitor compression in French and the 32B/14B rank flip show that transfer of guided creativity and of distillation both depend on the target language.

5.7. Pairwise Win-Rate Analysis

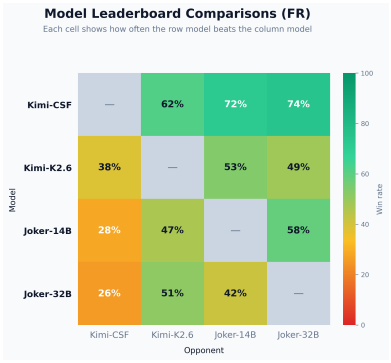
To better understand the performance gaps across models, we perform a direct analysis of all head-to-head matchups. Figure 4 visualizes these pairwise win rates across the three language tracks, highlighting exactly which matchups drive the overall Bradley–Terry rankings.

5.8. Key Experimental Findings

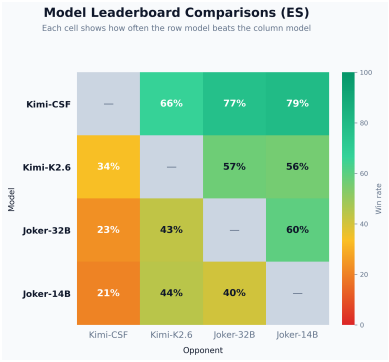
Guided creativity is the dominant factor: Figure 3 shows Kimi-CSF leading every language track by BT rating. The performance gap between Kimi-CSF and Kimi-Plain, both using Kimi K2.6, is the cleanest controlled comparison: BT separation ranges from 124 points (FR) to 227 (EN), while head-to-head win rates range from 61.5% to 80.7%. Since the backbone, constraint filtering, and test instances are identical, this gap is attributable entirely to CSF persona diversity and pairwise candidate selection. The result confirms that guided creativity transfers from open-ended English headline humor [1] to constrained cross-lingual pun-brief generation.



(a) English. Kimi-CSF wins ≥78% against every opponent; the three baselines split matchups near evenly among themselves (43–57%), matching the tight BT cluster in Figure 3a.



(b) French. Kimi-CSF's advantage vs. Kimi-Plain drops to 61.5%; HumorGen-JOKER-14B beats 32B head-to-head (58.3%), consistent with the BT rank reversal in Figure 3b.



(c) Spanish. Pattern resembles English: Kimi-CSF dominates all pairs; 32B beats 14B (59.6%) while both remain well below Kimi-CSF in BT (Figure 3c).

Figure 4: Head-to-head win-rate heatmaps by language (four-system ablation). Cell (i, j): percentage of test briefs where row system i is judged funnier than column system j (Llama 3.3 70B). Heatmaps complement the BT leaderboards (Figure 3): they expose *which* pairwise comparisons drive the multilingual ranking shifts, especially the tighter French competitor field and the 14B/32B inversion.

Distillation nearly matches the teacher backbone at plain generation, but cannot replicate guided search: The fine-tuned students were trained on CSF-curated, pairwise-ranked data and therefore received the pipeline's guidance at training time. Critically, the student models (Qwen3-14B/32B) trail Kimi-Plain (the same Kimi K2.6 backbone with no guidance) by only 15–75 BT points across most language-system pairs, demonstrating that the SFT objective successfully encodes backbone-level joke generation competence into substantially smaller open-weight models. Nevertheless, neither HumorGen-JOKER-32B nor 14B closes the gap to Kimi-CSF in any language (BT deficit: 147–241 points), and they do not consistently outperform Kimi-Plain, which received no task-specific training. This suggests that distillation transfers the structural form of pun-brief generation (parsing the brief, activating both senses) but not the test-time search breadth that CSF with pairwise ranking achieves through multi-candidate evaluation.

6. Limitations

We note the following three limitations of this study:

- **Automated Evaluation**: Our pairwise evaluation was conducted automatically. We did not conduct a secondary large-scale human evaluation to corroborate the automated rankings.
- **Text-Only Modality**: Our humor generation framework and the JOKER task itself are strictly text-based. This does not capture multimodal elements, such as visual or auditory cues, which are often integral to real-world humor.
- **Language Scope**: Our experiments are restricted to English, French, and Spanish. Extending the pipeline to a broader, more linguistically diverse set of languages remains an area for future work.

7. Conclusions

We submitted all four systems to the JOKER Lab Arena in English, French, and Spanish (Table 5), ablating where CSF enters the pipeline: test time (Kimi-CSF), training curation only (HumorGen-JOKER), or not at all (Kimi-Plain). Stage 1 fine-tuning follows HumorGen-SFT on MWAHAHA train headlines; subsequent JOKER stages apply an asymmetric cross-lingual LoRA curriculum.

Interim internal evaluation (Figure 3) ranks Kimi-CSF first in all three languages, with distilled student models trailing Kimi-Plain by a modest margin, indicating that the SFT curriculum transfers generation-level competence into open-weight parameters. The decisive performance gap lies between CSF-guided and unguided systems, not between backbone models. Official Arena scores (Table 6) will situate all four variants against the full participant pool; we anticipate these results will further validate the contribution of guided generation under the competitive multilingual evaluation.

Declaration on Generative AI

Gemini 3.0 was used in fixing grammar corrections in this work, however the ideas and experimentations are intellectual contributions of the authors.

References

[1] E. Ajayi, P. Mitra, Humorgen: Cognitive synergy for humor generation in large language models via persona-based distillation, arXiv preprint arXiv:2604.09629 (2026).

[2] I. Kuzmin, et al., Overview of the CLEF 2026 JOKER task 4: Humor generation, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026. To appear.

[3] K. Team, Y. Bai, Y. Bao, Y. Charles, C. Chen, G. Chen, H. Chen, H. Chen, J. Chen, N. Chen, et al., Kimi k2: Open agentic intelligence, arXiv preprint arXiv:2507.20534 (2025).

[4] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[5] D. Yang, Human-ai interaction in the age of large language models, in: Proceedings of the AAAI Symposium Series, volume 3, 2024, pp. 66–67.

[6] 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.

[7] 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, 2024, pp. 13246–13257.

[8] H. Wang, Y. Zhao, D. Li, X. Wang, G. Liu, X. Lan, H. Wang, Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps, arXiv preprint arXiv:2410.10370 (2024).

[9] J. Zhang, S. Luo, R. Zhang, Q. Su, Humorchain: Theory-guided multi-stage reasoning for interpretable multimodal humor generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 19176–19185.

[10] M. Goel, P. Krishnamurthy, R. Mamidi, Automating humor: A novel approach to joke generation using template extraction and infilling, in: Proceedings of the 21st International Conference on Natural Language Processing (ICON), 2024, pp. 442–448.

[11] J. Lemmens, V. De Marez, Computational humor modeling: A survey on the state of the art, ACM Computing Surveys 58 (2026) 1–37.

[12] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the clef 2025 joker lab: Humour in machine, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 315–337.

[13] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The joker corpus: English-french parallel data for multilingual wordplay recognition, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2796–2806.

[14] R. He, Y. He, L. Bai, J. Liu, Z. Sun, Z. Tang, H. Wang, H. Xia, N. Deng, Chumor 1.0: A truly funny and challenging chinese humor understanding dataset from ruo zhi ba, arXiv preprint arXiv:2406.12754 (2024).

[15] R. He, Y. He, L. Bai, J. Liu, Z. Sun, Z. Tang, H. Wang, H. Xia, R. Mihalcea, N. Deng, Chumor 2.0:

Towards better benchmarking chinese humor understanding from (ruo zhi ba), in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 21799–21818.
[16] R. Ortega-Bueno, C. E. Muniz-Cuza, J. E. M. Pagola, P. Rosso, Uo upv: Deep linguistic humor detection in spanish social media, in: Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish society for natural language processing (SEPLN 2018), 2018, pp. 204–213.
[17] M. Yatsu, K. Araki, Comparison of pun detection methods using japanese pun corpus, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.
[18] E. Ajayi, P. Mitra, Automatic humor detection: A comprehensive survey from theoretical foundations to large language models, ResearchGate Preprint (2026). URL: https://doi.org/10.13140/RG.2.2.24393.61288. doi:10.13140/RG.2.2.24393.61288.
[19] S. Castro, L. Chiruzzo, S. Góngora, S. Rahili, N. Deng, I. Sastre, V. Amoroso, G. Rey, A. Rosá, G. Moncecchi, J. A. Meaney, J. J. Prada, R. Mihalcea, SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate, in: Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026), 2026.
[20] E. Ajayi, P. Mitra, Humorrank: A tournament-based leaderboard for evaluating humor generation in large language models, arXiv preprint arXiv:2604.19786 (2026).
[21] L. Ermakova, et al., Overview of CLEF 2026 JOKER track: Humor detection, search, and translation, in: M. Hagen, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), LNCS, Springer-Verlag, 2026. To appear.

A. System and training details

A.1. Model and Judge Configuration

We employ **Kimi K2.6** (MoonshotAI) as the core generation backbone for both Kimi-CSF and Kimi-Plain. Our distilled open-weight student models, HumorGen-JOKER-32B and HumorGen-JOKER-14B, are initialized from the **Qwen3-32B** and **Qwen3-14B** base models respectively. Both students are fine-tuned using QLoRA via the Unsloth library in 4-bit precision. Finally, we utilize **Llama 3.3 70B Instruct** as the pairwise judge. This model was accessed via the Groq API for the final post-submission ablation, and run locally on NVIDIA H100 80 GB GPUs for candidate ranking during training data curation.

A.2. Fine-Tuning Hyperparameters

All LoRA fine-tuning stages share the following configuration: bf16 precision, linear learning-rate decay, a gradient accumulation factor of 4 (with a per-device batch size of 4, yielding an effective batch size of 16), and evaluation every 50 gradient steps. Early stopping patience was set to 2 for Stage 2 and 1 for Stage 3. Additionally, thinking mode was explicitly disabled at inference time to evaluate pure plain generation. All fine-tuning and local inference experiments were conducted on NVIDIA H100 80 GB GPUs (allocating one GPU per job). The resulting adapter checkpoints are highly lightweight, requiring approximately 500 MB for the 32B models and 250 MB for the 14B models.

Table 3 summarizes the training convergence, reporting the lowest evaluation loss achieved on a 5% held-out validation split for each stage. Stage 1 follows the methodology of [1] on the SemEval-2026 MWAHAHA train headlines (CSF generation + pairwise ranking, yielding ≈12k examples). Stages 2a and 2b use the JOKER CSF datasets detailed earlier in Table 3.

A.3. Pun-Brief Prompt Template

All systems use the following base pun-brief instruction at both training and test time:

Stage	Model	Data	Epochs	Best eval ↓	Δ
1	14B	12k headlines	1.26	1.964	–
1	32B	12k headlines	1.26	1.890	–
2a	14B	3,985 EN rows	2.0	1.846	−0.118
2a	32B	3,985 EN rows	2.0	1.789	−0.101
2b	14B	11,038 multi	1.14	1.696	−0.180
2b	32B	11,038 multi	1.14	1.622	−0.168

Table 11
Training convergence summary. Δ is change from Stage 1 init.

```
Write a humorous text for this pun brief.

Language: {language}
Pun word: {pun_word}
Sense A: {sense_a}
Sense B: {sense_b}
Keywords: {keywords}   [omitted when absent]
```

Figure 5: Pun-brief instruction template.

A.4. Cognitive Synergy Personas

For Kimi-CSF, we adopt the six cognitive synergy personas exactly as defined in [**?**]. This ensemble maximizes the diversity of the generated candidates:

1. **The Observer** (Style: Jerry Seinfeld) – Finds the mundane absurdity using "The Relatable Truth."
2. **The Wordsmith** – Master of wordplay, focusing on double meanings and linguistic twists.
3. **The Optimist** (Style: Ted Lasso) – Employs joyful misdirection and wholesome enthusiasm.
4. **The Absurdist** – Breaks causal expectations using surreal leaps of logic.
5. **The Cynic** – Applies ironic social critique with deadpan delivery.
6. **The Neurotic** (Style: Larry David) – Focuses on petty grievances and hyper-fixation.

Each persona prepends a system prompt that encodes its specific humor mechanism and requires a <THOUGHT> / <JOKE> structured output format.

A.5. Constraint Verification Prompt

To ensure candidates adhere to the JOKER task parameters before entering the pairwise ranking tournament, Kimi-CSF filters candidates using an LLM-as-a-judge (Llama 3.3 70B Instruct). The judge evaluates five strict constraints (R1–R5) and outputs a JSON boolean dictionary:

```
R1. Is the joke entirely in {language}?
R2. Is "{pun_word}" (or a clear inflection) present in the joke?
R3. {sense_verification_check}
R4. Does the joke NOT start with meta-commentary ("Here is a joke", etc.)?
R5. Is the joke short, punchy, and non-empty (not a wall of text)?

Return JSON exactly:
{"R1":true/false, "R2":true/false, "R3":true/false, "R4":true/false,
"R5":true/false, "all_pass":true/false, "fail_reason":"..."}
```

Figure 6: LLM-as-a-judge verification prompt (R1–R5 constraints).