CLEF 2025 JOKER Track: No Pun Left Behind

Notebook for the JOKER Lab at CLEF 2025

Igor Kuzmin[1,2]

[1]Universitat Pompeu Fabra
[2]Barcelona Supercomputing Center

Abstract

Humor processing remains a challenging problem for NLP due to linguistic ambiguity, language-specific nuances, and intricate wordplay. The CLEF JOKER 2025 lab tackles this with two tasks we participated in: humour-aware information retrieval in Portuguese and English (Task 1), and pun translation from English to French (Task 2). For Task 1 we developed a hybrid pipeline combining BM25, dense retrieval with `multilingual-e5-small`, and a cross-encoder reranker, achieving MAP 0.050 and NDCG@100 0.172 in English, and MAP 0.074 and NDCG@100 0.184 in Portuguese. For Task 2 we fine-tuned Lucie-7B-Instruct and CroissantLLMChat-v0.1 using supervised fine-tuning (SFT) and Adaptive Rejection Preference Optimization (ARPO), obtaining a best BLEU of 42.40 (Lucie + SFT) and demonstrating a modest overlap trade-off (41.32 BLEU) when integrating ARPO, while CroissantLLM variants scored 35.17 and 35.28 BLEU. Our experiments show the baseline IR setup underperforms compared to more advanced systems, while the LLMs-based pun translation achieves best results confirming the promise of their cross-lingual wordplay transfer.

Keywords

Humor Analysis, Humor Retrieval, Humor Translation, Information Retrieval, LLM, Machine Translation,

1. Introduction

While humour plays an essential role in human interaction, it remains a complex challenge for advanced natural-language-processing (NLP) systems—even for the latest large language models (LLMs). Cultural differences, implicit meanings, intricate wordplay, and the inherently subjective nature of humour all blur the clear indicators models rely on, making computational detection, translation and transfer of humour far from trivial.

The JOKER Lab at CLEF 2025 [1] introduces multiple tasks that are aimed to address this automatic humour analysis complexities. The following tasks are:

- Task 1: Humour-aware Information Retrieval [2]. The purpose of this task is to extract short humorous texts from a set of documents based on a given query. The received texts must meet two criteria: match the query and be an example of a wordplay. In the latest edition of this track was introduced a new Portuguese dataset to challenge multilingual and multicultural aspect of humour.
- Task 2: Wordplay Translation [3]. The objective of this task is to translate punning jokes from English to French preserving original idea and meaning. Latest version of the corpus comprise 1,405 English sources and 5,838 translations.
- Task 3: Onomastic Wordplay Translation [4]. This task is based on translation of name-related wordplay from English to French.

In this paper we participate Tasks 1 and 2. We investigate two research questions:

1. How well can a strong industry-grade hybrid retrieval baseline identify humorous passages in English and Portuguese?

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
igor.kuzmin@upf.edu (I. Kuzmin)
https://www.linkedin.com/in/igor-kuzmin-tech/ (I. Kuzmin)
0009-0001-1513-1834 (I. Kuzmin)

2. To what extent can latest LLMs succeed at translating English wordplay into French while retaining humour and intent?

Our contributions are threefold:

- A systematic evaluation of hybrid retrieval with reranking for humour-aware IR in English and Portuguese.
- An LLM-based pun translation pipeline that contrasts Supervised Fine-Tuning (SFT) with Adaptive Rejection Preference Optimization (ARPO).
- A quantitative evaluation of BLEU performance for SFT-only versus ARPO-augmented models.

The remainder of this report is organized as follows. Section 2 details our approaches, Section 3 presents experimental results, and Section 5 draws conclusion.

2. Approach

2.1. Task 1: Humour-Aware Information Retrieval

We follow the standard pipeline of (i) dense retrieval, (ii) lexical retrieval, and (iii) cross-encoder reranking.

2.1.1. Data

The official English and Portuguese corpora and query sets provided by the organizers [2] constitute our primary data. To enlarge training resources, we sampled 10% of the documents and generated synthetic queries with `gpt-4o-mini` wordplay classifier, prompting it to both label each passage and, for those marked as wordplay, generate a concise search-style query (see Listing 1).

Listing 1: System prompt for the wordplay classifier

```
You are an assistant that classifies short documents as wordplay (
    jokes) or not, and - if it is wordplay - generates a concise
    search-style query that would retrieve this joke.
For example:
- Text: "Why did the scarecrow win an award? Because he was
    outstanding in his field."
  is_wordplay: True
  generated_query: "scarecrow award"
- Text: "The mitochondria is the powerhouse of the cell."
  is_wordplay: False
  generated_query: ""
```

We combined the original (query, reference) pairs with these synthetic pairs to form our positive training set. Next, we performed hard-negative mining using a pretrained SentenceTransformer (`all-MiniLM-L12-v2`) with the following configuration:

- Score range: retrieve candidates ranked 8–100 by cosine similarity
- Maximum similarity: 0.8 (to avoid too-easy negatives).
- Relative margin: 0.05 (filter out near-duplicates).
- Negatives per positive: 5, sampled at random.

This yields triplets of the form (query, positive, negatives).

Finally, we split the mined triplets into train (90 %), validation (5 %), and test (5 %) sets by first holding out 10 % for evaluation and then equally splitting that hold-out into validation and test.

2.1.2. Models

Next we fine-tune `intfloat/multilingual-e5-small` [5] for one epoch using 16-sentence batches and a warm-up ratio of 0.1 on query–document pairs, to enhance humour-aware semantic search. We were interested to compare two contrastive objectives: the popular *Multiple-Negative Ranking* loss (MNRL) and our *Adaptive Margin* loss, inspired by SigLIP [6] and MNRL.

Humour often hinges on very subtle semantic shifts (puns, wordplays) where positives and hard negatives lie close in embedding space; MNRL is given in Equation (1), it's forces the model to pull true humour examples away from all negatives, while Adaptive Margin loss introduces a temperature t and bias b (Equation (2)), this dynamic penalty preserves learning signal for near-tie cases, helping the retriever tease apart genuinely funny hits from near misses. Preliminary experiments on `all-nli` dataset [7] indicated similar cosine distributions, with Adaptive Margin converging more stably.

$$\mathcal{L}_{\text{MultipleNegativeRanking}} = \frac{1}{K}\sum_{i=1}^{K}\Big[\underbrace{\log\sum_{j=1}^{K}\exp\big(\langle x_i, x_j\rangle\big)}_{\text{negative similarity}} - \underbrace{\langle x_i, x_i\rangle}_{\text{positive similarity}}\Big]. \tag{1}$$

$$\mathcal{L}_{\text{AdaptiveMargin}} = \frac{1}{K}\sum_{i=1}^{K}\Big[\log\sum_{j=1}^{K}\exp\big(t\,\langle x_i, x_j\rangle + b\big) - \big(t\,\langle x_i, x_i\rangle + b\big)\Big], \quad \text{where} \quad t = e^{T'}, \quad b \in \mathbb{R}. \tag{2}$$

As lexical retriever we used a BM25 index [8] which is built with Anserini, while dense vectors are stored in Qdrant [9].

Finally We train `cross-encoder/ms-marco-MiniLM-L12-v2` [10] for two epochs on the mined triplets, using batch size 16 and warm-up ratio 0.1.

For each query we retrieve the top-1000 documents from both dense and BM25 indices, merge by reciprocal rank fusion, and rerank the top-100 with the cross-encoder.

2.2. Task 2: Wordplay Translation

Our translation system follows a two-stage strategy: supervised fine-tuning (SFT) and ARPO preference optimization.

2.2.1. Data

We merge the JOKER Task 2 corpus [3] with parallel EN–FR sentences from X-ALMA[1] [11]. After formatting prompts with the template in Listing 2, the data are split $96 : 1.5 : 2.5$ for SFT and $90 : 2.5 : 7.5$ for ARPO preference tuning for train/val/test.

We applied the following prompt:

Listing 2: Translation prompt template

```
Translate the following text from English into French.
English: {source}
French: {target}
```

[1] https://huggingface.co/datasets/haoranxu/X-ALMA-Parallel-Data

2.2.2. Models

We experiment with `croissantllm/CroissantLLMChat-v0.1` [12] and `OpenLLM-France/Lucie-7B-Instruct-v1.1` [13] due to their bilingual capabilities. Both models are 8-bit-quantized and fine-tuned with LoRA [14]. At inference we use 3-beam search, temperature 0.3, top-p 0.9, and repetition penalty 1.3.

For both training and inference we employed random seed equal to 3407.

2.2.3. Supervised Fine-Tuning

The supervised fine-tuning (SFT) script uses a completions-only data collator that masks out prompt tokens and computes loss solely over the generated responses, which—given our instruction-style data—forces the model to focus on learning to produce high-quality completions rather than memorizing the prompts. Fine-tuning is done with the `trl` library [15], using an inverse-sqrt scheduler, a peak learning rate of 5×10^{-5}, batch size 32, and gradient accumulation steps of 4.

2.2.4. ARPO Optimization

To enhance humour retention, we apply ARPO[2] [11] after SFT stage, which combines behavior-cloning and preference losses. While Reinforcement learning from human feedback (RLHF) is known for out-of-domain improvement as well as better generalization within small amounts of training data compared to supervised fine-tuning only, we were particularly interested in bringing the latest state-of-the-art methods in Natural Machine Translation (NMT) to humour preservation tasks.

The ARPO loss has two components: a behavior cloning (BC) term to prevent the model from drifting too far from its original distribution, and an adaptive preference term that rejects low-quality candidates. Formally, the core ARPO loss is defined as:

$$\mathcal{L}_{\text{ARPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\big(\beta\log\pi_\theta(y_w|x) - \tau_\theta(y_w,y_l)\,\beta\log\pi_\theta(y_l|x)\big) + \log\pi_\theta(y_w|x)\Big]. \quad (3)$$

Here, the first term inside the expectation is the *preference loss* (with a temperature β), and the second term is the *BC* regularization.

The adaptive penalty weight $\tau_\theta(y_w, y_l) \in [0, 1]$ modulates how strongly we down-weight the likelihood of the worse translation y_l, based on its similarity to the preferred output y_w:

$$\tau_\theta(y_w, y_l) = \min\Big(\lambda\,\big(e^{\eta\cdot z_\theta(y_w,y_l)} - 1\big),\ 1\Big), \quad (4)$$

where λ is scale, η is a hyperparameter controlling penalty sensitivity, and $z_\theta(y_w, y_l)$ measures the distance between the two responses via their average log-likelihoods:

$$z_\theta(y_w, y_l) = \Big|\tfrac{1}{|y_w|}\log\pi_\theta(y_w|x) - \tfrac{1}{|y_l|}\log\pi_\theta(y_l|x)\Big|. \quad (5)$$

Preference pairs for ARPO stage are obtained by sampling negative (rejected) translations with an 8-bit version of X-ALMA model [3]. We used the library's default η (specified by `relax_coefficient_2`) and the scale of $\tau_\theta(y_w, y_l)$, λ (specified by `relax_coefficient_1`). The default values are $\eta = 0.4$ and $\lambda = 0.9$.

3. Results

3.1. Humour-Aware Information Retrieval

Table 1 summarizes our official submissions. For English, our reciprocal rank fusion (RRF) run achieves MAP 0.050 and NDCG@100 0.172; for Portuguese, MAP 0.074 and NDCG@100 0.184. These figures place our system in the middle of the leaderboard, indicating that hybrid search with lightweight reranking remains below state-of-the-art for humour-centric queries.

[2]Source code: https://github.com/fe1ixxu/ALMA/tree/xalma

[3]https://huggingface.co/mradermacher/X-ALMA-13B-Group4-GGUF

Table 1
Official Task 1 scores.

Language	MAP	RPrec	MRR	NDCG@100
English	0.050	0.030	0.065	0.172
Portuguese	0.074	0.057	0.119	0.184

Table 2
Examples of retrieved samples for Task 1.

Language	**Query**	**Top Retrieval**	**Score**
Portuguese	família	A família (do latim: familia) é um agrupamento humano formado por duas ou mais pessoas com ligações biológicas, ancestrais, legais ou afetivas que, geralmente, vivem ou viveram na mesma casa.	1.00
English	vision	Vision is the ability to think about or plan the future with imagination and wisdom.	1.00

Table 3
Official BLEU scores for Task 2.

Model	BLEU
Lucie-7B-Instruct-v1.1 + SFT	**42.40**
+ ARPO	41.32
CroissantLLMChat-v0.1 + SFT	35.17
+ ARPO	35.28

Although the pipeline returns high-confidence matches by retrieving lexically related definitions as shown in the Table 2, these passages lack any humorous content. This illustrates why our overall retrieval metrics remain low—the system fails to surface genuinely funny or wordplay.

3.2. Wordplay Translation

The official evaluation metric used by organizers for Task 2 is BLEU. Table 3 lists the four runs we submitted.

Surprisingly, the BLEU evaluation reveals that our straightforward SFT model still holds a slight edge over its ARPO-enhanced counterparts: the SFT baseline scored 42.40 BLEU, compared to 41.32 for the `Lucie-7B-Instruct-v1.1` + ARPO variant. This suggests that, although ARPO's adaptive preference loss can improve qualitative aspects—such as preserving humour or other nuanced translation properties—it may do so at the cost of n-gram overlap as measured by BLEU.

Table 4 provides a concrete example (source `en_83`) where ARPO better preserves the pun: the SFT-only translation *"poule en vitesse"* is a literal but awkward translation, whereas the ARPO output *"poule pressée"* more naturally mirrors the play on *"pullet"* and *"pressé."* In practical terms, if maximizing standard BLEU is the primary objective, the pure SFT approach remains the stronger choice. However, if downstream qualities that BLEU cannot fully capture—such as humour preservation—are important, integrating ARPO may still be worthwhile despite the slight BLEU trade-off.

Table 4
Examples of French translations for English source *en_83*, Task 2.

Model	Source (EN)	Translation (FR)
Lucie-7B-Instruct + SFT	A farmer wanting to kill a chicken for dinner has to move faster than a speeding pullet.	Un fermier qui veut tuer un poulet pour le dîner doit bouger plus vite qu'une poule en vitesse.
Lucie-7B-Instruct + SFT + ARPO	A farmer wanting to kill a chicken for dinner has to move faster than a speeding pullet.	Un fermier qui veut tuer un poulet pour le dîner doit se déplacer plus vite qu'une poule pressée.

4. Post-Competition Analysis

After the official competition deadline, we conducted a detailed post-competition analysis of Task 2. Our goal was to combine multiple datasets, explore alternative hyperparameters, and evaluate different model variants to improve upon our initial training data. In particular, we extended our experiments with the Lucie-Instruct model, motivated by some skepticism regarding our original ARPO loss results.

4.1. Extended SFT Experiments

We updated the SFT configuration by merging the JOKER and X-ALMA translation pairs and compared two training regimes:

1. Training exclusively on the JOKER Task 2 dataset.
2. Training on the combined JOKER Task 2 and X-ALMA parallel datasets to increase the diversity of translation examples.

First, we split the JOKER parallel dataset into training and validation sets (train size = 0.97, validation size = 0.03) and mixed the SFT training split with the X-ALMA parallel dataset. A grid search was performed over a range of hyperparameters (see Appendix A).

After evaluating SFT, we selected the two best models according to the COMET-22 [16] metric since its has high correlation with human judgment: one trained on the JOKER-only dataset (v4) and one on the combined dataset (v8) (see Appendix B.

4.2. Extended ARPO Experiments

For ARPO, negative samples were generated for the JOKER Task 2 dataset using the X-ALMA model 3 with the same train/validation split. We again compared two setups:

1. Training solely on the obtained JOKER preference dataset.
2. Training on a mixture of the JOKER preference dataset and X-ALMA EN-FR preference pairs dataset[4].

The ARPO hyperparameter grid is also detailed in Appendix A. Finally, we selected the best ARPO models in terms of COMET-22 for each dataset combination and each SFT model (see Appendix B).

4.3. Extended Results

As shown in Table 5, the SFT-only models achieve higher BLEU scores, consistent with the findings in Section 3. The ARPO-enhanced variants, however, fail to produce any significant gains on this metric. Notably, the configurations in Appendix B show that optimal performance required different η values: $\eta = 0.4$ for the JOKER-only dataset versus $\eta = 1.0$ for the combined dataset. This suggests that smaller,

[4]https://huggingface.co/datasets/haoranxu/X-ALMA-Preference

Table 5
BLEU scores for various `Lucie-7B-Instruct-v1.1` fine-tuning variants.

Configuration	BLEU
Lucie-7B-Instruct-v1.1	40.68
+ SFT (v4)	**42.12**
+ ARPO (v1)	40.65
+ ARPO (v5)	40.69
+ SFT (v8)	41.27
+ ARPO (v7)	40.69
+ ARPO (v11)	40.64

less diverse datasets benefit from weaker penalties that preserve more translation variants, while larger, more diverse datasets require stronger adaptive penalties to effectively filter translation quality while maintaining optimization stability.

5. Conclusion

In this paper, we presented two baseline systems for the CLEF 2025 JOKER Lab: a hybrid retrieval pipeline combining BM25, dense retrieval with `multilingual-e5-small`, and cross-encoder reranking for Task 1; and an LLM-based pun translation framework that combines SFT with ARPO preference optimization for Task 2. Our retrieval system achieved mid-tier performance (MAP 0.050/0.074, NDCG@100 0.172/0.184), revealing that purely lexical or semantic matches often miss true wordplay. In our translation experiments, SFT maximizes BLEU (42.40) but tends to produce overly literal translations, whereas ARPO trades a small drop in BLEU (41.32) for more idiomatic, pun-preserving outputs.

Future work will explore sophisticated retrieval methods. We also plan to explore larger bilingual LLMs and more diverse training corpora to improve humour translation in multilingual settings.

Declaration on Generative AI

By using the activity taxonomy in https://ceur-ws.org/genai-tax.html:
During the preparation of this work, the author(s) used **OpenAI ChatGPT (GPT-4)** in order to:

- **Formatting assistance**: ensuring adherence to the formatting guidelines required by journals or institutions.
- **Peer review simulation**: simulating peer review by providing feedback on the strengths and weaknesses of the manuscript.
- **Coherence enhancement**: improving the overall clarity and logical flow of the text.

After using this tool/service, the author(s) reviewed and edited the generated content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Clef 2025 joker lab: Humour in the machine, in: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2025, p. 389–397. URL: https://doi.org/10.1007/978-3-031-88720-8_59. doi:10.1007/978-3-031-88720-8_59.

[2] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the clef 2025 joker task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes

of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[3] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the clef 2025 joker task 2: Wordplay translation from english into french, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[4] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the clef 2025 joker task 3: Onomastic wordplay translation, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[5] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.
[6] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, 2023. URL: https://arxiv.org/abs/2303.15343. arXiv:2303.15343.
[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.
[8] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.
[9] Q. Team, Qdrant: Vector search engine for the next generation of ai, https://qdrant.tech/, 2025. Accessed: June 1, 2025.
[10] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.
[11] H. Xu, K. Murray, P. Koehn, H. Hoang, A. Eriguchi, H. Khayrallah, X-alma: Plug play modules and adaptive rejection for quality translation at scale, 2025. URL: https://arxiv.org/abs/2410.03115. arXiv:2410.03115.
[12] M. Faysse, P. Fernandes, N. M. Guerreiro, A. Loison, D. M. Alves, C. Corro, N. Boizard, J. Alves, R. Rei, P. H. Martins, A. B. Casademunt, F. Yvon, A. F. T. Martins, G. Viaud, C. Hudelot, P. Colombo, Croissantllm: A truly bilingual french-english language model, 2025. URL: https://arxiv.org/abs/2402.00786. arXiv:2402.00786.
[13] O. Gouvert, J. Hunter, J. Louradour, C. Cerisara, E. Dufraisse, Y. Sy, L. Rivière, J.-P. Lorré, O.-F. community, The lucie-7b llm and the lucie training dataset: Open resources for multilingual language generation, 2025. URL: https://arxiv.org/abs/2503.12294. arXiv:2503.12294.
[14] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.
[15] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, Q. Gallouédec, Trl: Transformer reinforcement learning, https://github.com/huggingface/trl, 2020.
[16] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, A. F. T. Martins, COMET-22: Unbabel-IST 2022 submission for the metrics shared task, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 578–585. URL: https://aclanthology.org/2022.wmt-1.52/.

A. Hyperparameter Grids

Table 6
SFT Hyperparameter Grid

Parameter	Values
Learning rates	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$
Epochs	{1, 2}
Batch size	16
Scheduler	inverse_sqrt
Warmup ratio	0.01
Weight decay	0.01

Table 7
ARPO Hyperparameter Grid

Parameter	Values
Learning rate	5×10^{-7}
Epochs	1
Batch size	8
Scheduler	inverse_sqrt
Warmup ratio	0.01
Weight decay	0.01
Loss type	sigmoid
CPO alpha	1.0
β	0.1
λ	0.9
η	{0.4, 1.0, 1.5}

B. Training Configurations

Table 8
SFT configurations: dataset, learning rate, and epochs.

Config	Dataset	LR	Epochs
v4	JOKER-only	1×10^{-4}	2
v8	JOKER + X-ALMA	1×10^{-4}	2

Table 9
ARPO configurations: dataset and η.

Config	Dataset	η
v1	JOKER-only	0.4
v7	JOKER-only	0.4
v5	JOKER + X-ALMA preference	1.0
v11	JOKER + X-ALMA preference	1.0