From LOL to LLM: Measuring Multilingual Multi-Turn Humor Understanding in AI

Atrey Desai adesai10@umd.edu
Leo Du ldu0040@umd.edu
James van Doorn jvand@umd.edu
Kamala Sreepada kamala@umd.edu

1 Problem statement

The ability of Large Language Models (LLMs) to recognize and understand humor has long been a subject of study, with many benchmarks created for this purpose. Humor is an interesting topic to study because it is cognitively intense and requires a nuanced understanding of language to be properly appreciated. Additionally, many forms of comedy rely on implied meaning and skipping logic, asking the audience to bridge the gap to understand the humor. As such, it provides a useful method for testing LLMs' natural language understanding.

We have identified several gaps in the existing literature, specifically regarding the lack of benchmarks that study LLMs' ability to understand long-context jokes across cultures. As such, the goal of this project is to develop a humor-focused benchmark for models that spans cultures and assesses their ability to make connections and understand implicature.

2 What you proposed vs. what you accomplished

- ~~Compile multi-turn humor dataset and assess multi-turn humor understanding in both English and Spanish~~
- ~~Create perturbed multi-turn humor dataset and study adversarial robustness in LLMs~~
- ~~Perform error analysis to figure out what kinds of jokes LLMs struggle with~~
- *Study spatiotemporal multimodal humor understanding*: Following feedback from the proposal to reduce our scope due to limited time, we decided to drop this task set.
- *Assess cross-cultural joke competency*: Following feedback from the proposal to reduce our scope due to limited time, we decided to adapt this into multilingual understanding.

3 Related work

Early work in recognizing humor using computational methods was based on binary classification. Rayz (2004) developed an algorithm to recognize humor through statistical language recognition techniques. Subsequently, Mihalcea and Strapparava (2005) used supervised learning to determine if sentences were humorous or non-humorous.

Later work recognized the importance of subjectivity in humor, focusing on demographic factors such as age, gender, and socioeconomic status (Meaney et al., 2021).

3.1 Multilingual and Cross-Cultural Humor Understanding

Humor understanding in multimodal contexts remains a relatively understudied area despite a huge amount research done in computational humor. Foundational work in this direction started from the area of vision-language modeling.

Cross-cultural humor understanding represents a critical evaluation frontier for Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), which require systems to comprehend intersections of linguistic knowledge and cultural contexts. While current benchmarks like HumorBench (Narad et al., 2025) demonstrate strong correlations with reasoning capabilities, they do not significantly consider cross-cultural and multilingual competencies of the models, which could result in potential

Our GitHub repository can be found here: https://github.com/atreydesai/humorbench cultural biases and misinterpretations when deployed across cultural boundaries. Recent multimodal and multilingual benchmarks have attempted to address these challenges. For example, YESBUT (Nandy et al., 2024a) offers a multilingual comics dataset for narrative contradiction tasks, PunMemeCN (Anonymous, 2025) focuses on Chinese pun memes that combine wordplay and imagery, FanChuan (Zheng et al., 2025) examines multilingual parody detection in social contexts, StandUp4AI (Barriere et al., 2025) captures laughter-tagged stand-up routines across seven languages, and the MuSe-Humor sub-challenge (Amiriparian et al., 2024) evaluated real-world, spontaneous humor transfer from German to English press conferences under cross-cultural conditions. Despite these advances, existing datasets remain skewed toward a narrow set of languages and genres, most tasks emphasize binary humor detection rather than translation, localization, or explanation, and model architectures rarely integrate explicit cultural context representations in multimodal humor understanding.

3.2 Multi-Turn Humor

Recent work on computational humor has focused on detecting humor in isolated sentences or single-line jokes in English, without examining how humor develops over multiple turns of dialogue. Current benchmarks focus on single-line and single-speaker humor. Several works have benchmarked LLMs' ability to understand (Trott et al., 2025) and generate (Quan et al., 2025) single-turn humor in conversational settings, demonstrating that state-of-the-art language models generally perform well, but fail to understand more subtle and nuanced humor. Additionally, work has shown that state-of-the-art LLMs are capable of understanding conversational implicature (Yue et al., 2024), and that models have innate knowledge for social language tasks (Choi et al., 2023). However, there has been little work to gauge LLMs' ability to understand and analyze humor that evolves over the course of multiple turns of dialogue.

4 Dataset

Our dataset assesses two tasks in both English and Spanish:

- **Overall Joke Classification:** Classify a whole multi-line joke into exactly one of the following categories: satire, parody, irony, aggressive, dry, self-deprecating, surreal/absurdism, wordplay, witty, topical, observational, anecdotal, dark.

- **Line Purpose Identification:** Identify the purpose of a line in developing the joke. Valid purposes are: establishing context, escalation, subversion, callback, misdirection, timing, meta-humor, punchline, redirection, non-line, wrap-up, repetition. In this task, the model will be presented the entire joke and asked to identify all purposes for a single multi-line joke at once. Newline characters are used to denote when a new sentence begins.

These tasks are challenging for models since it requires that they retain and correctly relate information across multiple lines, sometimes with deliberate ambiguity or delayed resolution. It also requires understanding implied meaning, as jokes rarely mean what they literally say and frequently violate conversational norms, or ask LLMs to catch implicit reasoning as jokes omit reasoning steps. LLMs must also possess world knowledge and cultural grounding, as many jokes use cultural references or stereotypes. Finally, it also requires the LLM to manage uncertainty. Where LLMs usually optimize for coherence or collapse multiple meanings into a single most-likely interpretation, these tasks ask the model to maintain uncertainty while it is being built, before being resolved by the punchline later.

The English data was scraped from transcripts of stand-up comedy shows stored on https://scrapsfromtheloft.com/stand-up-comedy-scripts/. For the overall joke classification, there are 286 jokes. For line purpose identification, there are 214 jokes, containing a total of 3106 individually annotated lines.

The Spanish data was gathered from the StandUp4AI (Barriere et al., 2025) https://github.com/Standup4AI/dataset using their URL extraction scripts, which filter for Spanish stand-up-style titles and discard clips shorter than one minute. The selected videos were then downloaded and transcribed using the StandUp4AI ASR pipeline, using WhisperX configured for Spanish. Whisper produces sentence-level transcripts, while WhisperX additionally generated word-level timestamps. The resulting ASR outputs formed the raw Spanish text used for joke selection and annotation in our benchmark.

Our English + Spanish combined dataset consisted of 499 jokes, with English being 326 and Spanish being 173 jokes.

Here are a few abbreviated examples of labeled jokes from the English dataset:

- Joke:

  ```
  So theres this gorilla named
      Koko.\nHas everyone heard
      of Koko the gorilla?\nYes
      .\nSome for yeses.\nKoko
      the gorilla, for those of
      you that dont know, is a
      gorilla that spoke fluent
      sign language.\n
  ...
  \nAnd theyre like, Koko!  And
      Kokos like, What else do
      you know And theyre like,
      Koko, no. And Kokos like
      And theyre like, Koko, no.
       And Kokos like, Who is
      the president right now?
      Those are the kind of
      jokes I write.
  ```

  Overall Joke Classification: witty

  Line Purpose Identification Labels:

  ```
  establishing context\
      nestablishing context\
      ntiming\ntiming\
      nestablishing context
  ...
  \nmeta-humor
  ```

- Joke:

  ```
  So, I got a dog recently.\
      nThats a big thing in my
      life.\nI went down to the
      pound and I got a free dog
      .\nThats how I say it.
  ...
  \ nIve  been mugged, repeatedly
      ! but the second they see
      that four-legged P90x body
       coming down the street
      everybody scatters.\nThe
      greatest dog you could
      ever have is a pit bull.\
      nIts like having a gun you
       can pet!\nIts tremendous!
  ```

  Overall Joke Classification: Observational Line Purpose Identification Labels:

  ```
  establishing context\
      nestablishing context\
      nestablishing context\
      ntiming
  ...
  \npunchline\nsubversion\
      npunchline\nwrap-up
  ```

One final challenge with this specific formatting of data is that it requires the ability to understand that newline characters represent sentence breaks, and also requires greater instruction following capability to output responses that can be parsed. The exact prompt is described in detail in 5.2.

The perturbed data used to study adversarial robustness was computed using all the English and Spanish data.

4.1 Data preprocessing

We first removed any transcripts containing inappropriate content, as most models will abstain from answering prompts containing inappropriate content. Refusing to answer would negatively affect the performance metrics, despite refusing to answer being the correct action in that situation. We also removed any artifacts that were left behind from the scraping, such as website banner text or missed HTML tags.

To finish preprocessing, we went through the remaining transcripts and segmented them into individual jokes to get them ready for annotation, with each joke being pasted into its own spreadsheet cell.

4.2 Data annotation

During our pilot annotation experiment, the major issue that came up was strictly defining what each category meant. Once we all agreed on what each category meant, we all based our annotation off of them. After annotations were finished, each person's annotations were double checked by one other person.

5 Approach

5.1 Models Studied

We focused our experiments on a set of small to mid-scale language models to balance computational constraints, architectural diversity, and relevance to current open-model research.

The selected models formed two ranges: 7-10 billion parameters, and 30-32 billion parameters.

- **7-10B class:**

  Qwen3-8B, Olmo-3-7B-Instruct (referred to as Olmo3-7B in this paper), Falcon3-10B-Instruct (referred to as Falcon3-10B in this paper), Apertus-8B-Instruct-2509, (referred to as Apertus-8B in this paper) and Ministral-8B-Instruct-2410 (referred to as Ministral-8B in this paper)

- **32B class:**

  Qwen3-32B and Olmo-3.1-32B-Instruct (referred to as Olmo3.1-32B in this paper)

The Qwen3 family emphasizes large-scale multilingual pretraining and strong reasoning performance, while the Olmo models represent a fully open scientific approach, and are relatively new. Falcon3-10B reflects a data-efficient, instruction-tuned design optimized for deployment constraints. Apertus-8B provides a European-led open model trained with a strong emphasis on openness and governance, and Ministral-8B serves as a compact, high-quality alternative designed to retain strong reasoning and instruction-following from its larger siblings.

All models were evaluated with sampling parameters suggested by their developers on HuggingFace. The most common set of parameters was a temperature of 0.6, and a top-p of 0.95. If no parameters were suggested, then we would default to the most common set of parameters. In addition, a stop sequence, as defined by the string “$\#\#\#END$” in the prompt, is used.

To prevent infinite repetitions, level the playing field between models, and speed up inference, models were limited to 2048 tokens of output. No other sampling hyperparameters were modified.

5.2 Multi-Turn Humor

5.2.1 Prompting and Model Loading

To enable model evaluation, we added task-specific prompts to each joke. For the first task of overall joke classification, we used this prompt that will provide the entire joke at once, and expect an answer in JSON form:

```
'Classify the following joke into one of
    these types: satire, parody, irony,
    aggressive, dry, self-deprecating,
    surreal/absurdism, wordplay, witty,
    topical, observational, anecdotal,
    dark. Output valid JSON of the form
    {"category": "<one type>", "
    reasoning": "<1-2 sentence
    explanation>"}### END for the joke:
    <JOKE IS PUT HERE> Your final answer
    should take the form {"category":
    "<one type>", "reasoning": "<1-2
    sentence explanation>"}### END If
    you do not output your final answer
    in this format, you will not receive
    any credit.'
```

For the second task of sentence-by-sentence purpose classification, we opted to still provide the entire joke at once to the model to provide as much context as possible, but included new-line characters to serve as line-breaks. As such, we also expect the model to provide classifications to all lines in a joke at once. This is the prompt:

```
'Here is a joke: <JOKE IS PUT HERE> END
    OF JOKE. Classify each newline-
    separated line of a multi-line joke
    by its role, assigning exactly one
    label from establishing context,
    setup, escalation, subversion,
    callback, misdirection, timing, meta
    -humor, punchline, redirection, non-
    line, wrap-up, repetition. Your
    final answer should take the form {"
    ANSWER":["label1", "label2",...]}###
    END. If you do not output your
    final answer in this format, you
    will not receive any credit.'
```

For jokes in Spanish, we altered the prompt slightly, including the phrase “in spanish” after the first occurrence of the word “joke”. So, the phrase “Here is a joke” becomes “Here is a joke in spanish”.

We found that putting the task or reminders on output format at the very end appears to improve adherence to the task in the smaller models.

Models were loaded offline using the `vLLM` python package in the UMD Nexus Cluster. For the 8-10B parameter models, a single Nvidia RTX A5000 or A6000 GPU is sufficient. For the 30-32B models, two A6000 GPUs were used simultaneously due to a single GPU not having sufficient VRAM to load the model.

5.2.2 Model Evaluation

We calculated three metrics to gauge model performance: accuracy, f1 score, and area under the ROC curve (AUC). Each metric was evaluated with Pass@1 and Pass@5. As such, each task and model combination resulted in five sets of completions. The Pass@1 metrics were calculated using the first set of completions. Under Pass@5, f1 score and AUC are calculated for all responses to a single prompt. If a model either fails to provide an completion that can be parsed, its answer will be considered “NA”. The viability of using AUC for a task like this is discussed in 6.

5.3 Adversarial Robustness

5.3.1 Adversarial Perturbation Framework

To support adversarial robustness evaluation, we developed a perturbation framework for generating controlled variants of labeled humor data. Prior work has shown that large language models are highly brittle under minimal lexical and orthographic changes, particularly in humor-related tasks where surface form plays a disproportionate role (Zangari et al., 2025), (Zhu et al., 2023). Our framework operationalizes these findings by systematically varying input form while preserving overall dataset structure.

Perturbations were applied selectively to punchline spans only, as identified by sentence-level annotations. This choice preserves joke setup and narrative context while targeting the most semantically critical component of each example. Recent work suggests that humor comprehension is asymmetric, with failures often localized to punchlines rather than distributed across the entire input (Horvitz et al., 2024). By holding surrounding context constant, this design enables downstream analyses to isolate punchline-level robustness failures.

The perturbation pipeline was implemented in Python and produces minimal pairs consisting of original jokes and perturbed counterparts, while maintaining alignment with the original task labels. Minimal-pair constructions have been shown to support more fine-grained robustness analysis than aggregate accuracy alone (Trott et al., 2025).

We generated four categories of perturbations. Semantic-preserving perturbations apply low-probability synonym substitutions and light orthographic noise designed to preserve meaning (Yang et al., 2023). Semantic drift perturbations introduce more aggressive synonym replacement, word-order swaps, and typographical noise. Orthographic perturbations consist of character-level typos and deletions (Zhu et al., 2023). Cultural or dialectal perturbations substitute region-specific lexical items to test cultural grounding in humor interpretation (Meaney et al., 2021), (Nandy et al., 2024b).

Each perturbation category was exported as a separate TSV file containing the perturbed joke text alongside the original task labels. These datasets were provided to downstream evaluation pipelines for robustness testing and were not evaluated directly within this component.

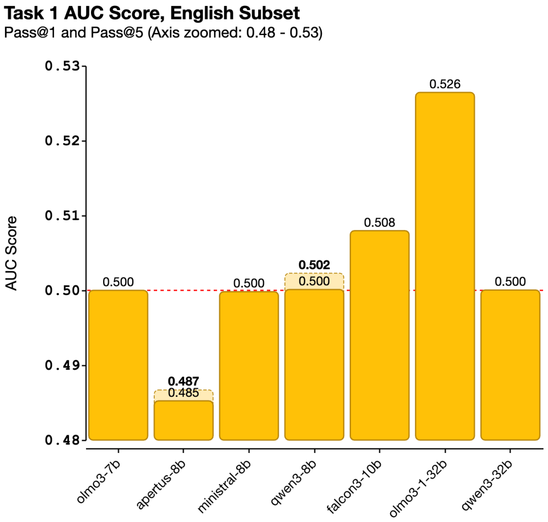

Figure 1: Area Under ROC Curve (AUC) for the overall joke classification task on the English dataset

6 Results

6.1 Area Under ROC Curve (AUC)

As seen in 1, the AUC appears to hold relatively steady at around 0.5 for all models, regardless of accuracy. This may be indicative of the fact that AUC is likely not extremely applicable to a task like this, as the LLMs do not return the probabilities of their answers. As such, there is no direct translation from the output we collect to an ROC curve.

6.2 English

6.2.1 Overall Joke Classification

For the overall joke classification task, the 32 billion parameter models drastically outperformed the smaller 6-10 billion parameter models. As seen in 2, Qwen3-32B reached 31.8% accuracy for Pass@5, while Falcon3-10b can only reach 8% accuracy, and the other models fall below 5% accuracy, with Olmo3-7b and Ministral-8b failing to answer a single prompt correctly.

While the degree to which the small models are worse than the large models is surprising, this difference matches our intuition behind what is required to succeed at this task, as the larger models are expected to contain greater language understanding and better instruction-following capability.

6.2.2 Line Purpose Identification

For the line purpose identification task, the 32 billion parameter models once again, perhaps un-

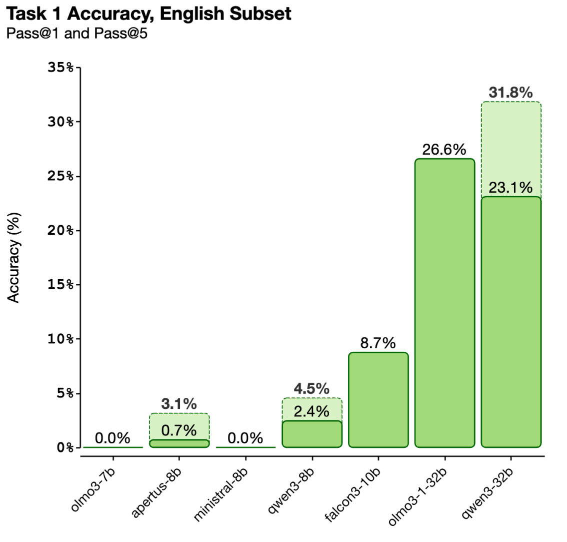

Figure 2: Accuracy for the overall joke classification task on the English dataset

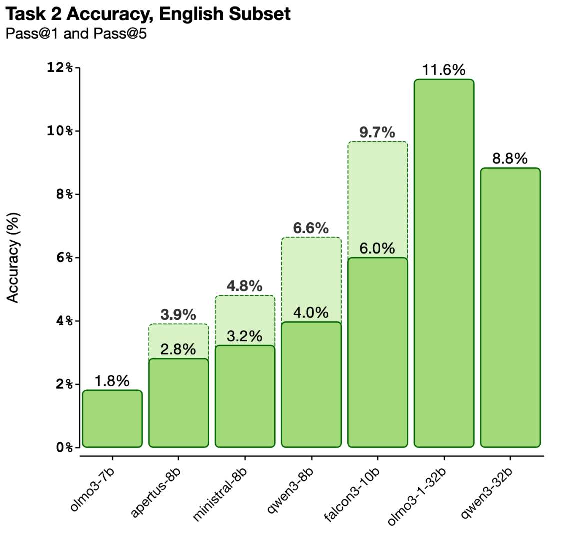

Figure 3: Accuracy for the line purpose identification task on the English dataset

surprisingly, outperform the 6-10 billion models. Here, Olmo3.1-32B is clearly the best model of the group, with an accuracy of 11.6% in both Pass@1 and Pass@5. Falcon3-10B performs surprisingly well on Pass@5 accuracy, having the second highest at 9.6%. Interestingly, the smaller 7-8B models all see increased accuracy compared to their accuracies for the overall joke classification task in both Pass@1 and Pass@5, while Falcon3-10B sees an increase in its pass@5 accuracy, while Qwen3-32B sees a substantial hit to its accuracy.

The decreases in accuracy for the 32B models are not surprising, as the line purpose identification task is expected to be more difficult, requiring more subtle language understanding. However, the increase in performance from the smaller models is unexpected. This suggests that while the small models are capable of understanding small parts of the jokes, they are not as capable in understanding the entire joke as a whole.

6.2.3 Adversarial Robustness

To evaluate adversarial robustness, we re-ran the same Task 1 and Task 2 prompts on four perturbation sets: cultural or dialect shifts, orthographic typos, semantic drift, and semantic-preserving perturbations. Perturbations were generated using sentence-level annotations and applied selectively to lines labeled as punchlines, while all other lines were left unchanged regardless of their role. This design preserves the overall structure and narrative progression of each joke while directly perturbing the most semantically critical span. We report Pass@1 and Pass@5 metrics, where Pass@1 corresponds to a single completion set and Pass@5 aggregates results across five completion sets.

For Task 1 (overall joke type classification), the strongest models were relatively stable under perturbations. Qwen3-32B has 30.8% Pass@5 accuracy on the unperturbed English set and ranges from 30.8% to 33.2% across perturbations (maximum change of +2.4 percentage points under orthographic typos). OLMo3.1-32B shows similarly small movement, from 26.6% at baseline to 25.9% under semantic drift (a change of -0.7 percentage points). Taken together, these results suggest that for the best-performing models, coarse humor category prediction is not consistently disrupted by localized punchline perturbations in our setup. In contrast, Falcon3-10B shows much larger swings (8.0% baseline versus 20.3% under semantic-preserving perturbations). Given Falcon3-10B's weaker baseline, these swings are more consistent with higher variance across conditions than with a reliable robustness improvement.

For Task 2 (line purpose identification), overall performance is low even at baseline, which makes small absolute shifts more meaningful in relative terms. Qwen3-32B degrades under all perturbations at Pass@1 accuracy, dropping from 8.8% baseline to 6.6% under orthographic typos and 6.7% under semantic-preserving perturbations (both about -2.2 percentage points, roughly a

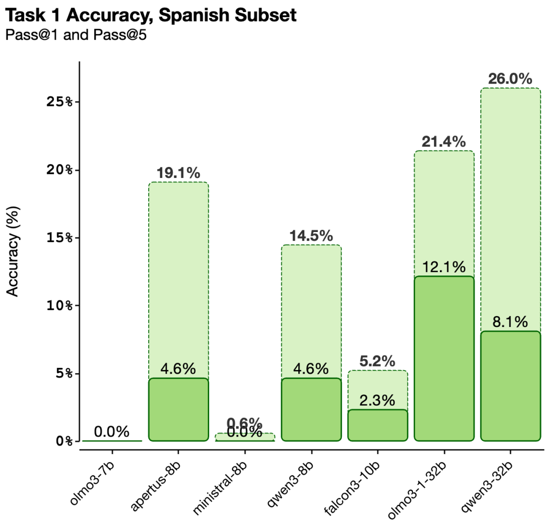

Figure 4: Accuracy for the overall joke classification task on the Spanish dataset

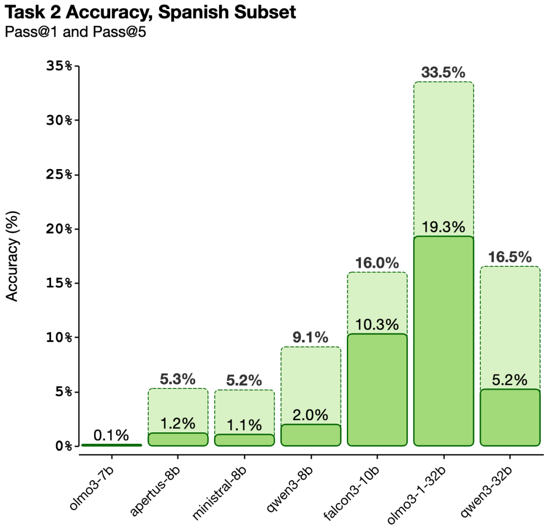

Figure 5: Accuracy for the line purpose identification task on the Spanish dataset

25% relative drop). The same pattern is reflected in Pass@1 F1 for Qwen3-32B, which falls from 6.65% at baseline to 4.31% under orthographic typos and 5.08% under semantic-preserving perturbations, indicating that the degradation is not limited to a single metric. Aggregating across five completion sets sometimes changes the picture. Under cultural or dialect shifts, Qwen3-32B decreases from 8.8% to 7.2% at Pass@1, but increases from 8.8% to 11.9% at Pass@5 (+3.1 percentage points), indicating substantial variability across completion sets for this structured labeling task. OLMo3.1-32B is comparatively stable on Task 2 at Pass@1 (within about 1 percentage point of baseline across perturbations), and improves at Pass@5 under semantic-preserving perturbations from 11.6% to 14.9% (+3.3 percentage points). Averaged across the evaluated English models, orthographic typos produce the largest mean drop in Task 2 Pass@1 F1 (about -0.46 percentage points), while semantic drift produces the largest mean drop in Task 2 Pass@5 accuracy (about -1.14 percentage points). Overall, Task 2 appears more sensitive to localized punchline edits than Task 1, consistent with Task 2 requiring structured multi-line outputs rather than a single global label.

6.3 Spanish

6.3.1 Overall Joke Classification

For Task 1 (overall joke classification task) on the Spanish subset, the 32 billion parameter models again outperform the 8-10 billion parameter models. As shown in 4 and Table 19, OLMo3.1-32B achieves the strongest results with 12.1% Pass@1 accuracy and 21.4% Pass@5 accuracy, while Qwen3-32B reaches 8.1% and 26.0% respectively. Smaller models rarely exceed 5% Pass@1 accuracy, and OLMo-3-7B and Ministral-8B fail to produce rarely any predictions at Pass@1, indicating very weak Spanish humor categorization for this size range.

Despite these differences, F1 scores for all models stay below 0.11 and AUC values remain near 0.5, which suggests that models often default to a small set of frequent label rather than capturing fine-grained joke types in Spanish. This pattern contrasts with English, where the same 32B models reach substantially higher accuracies, highlighting a gap in multilingual coverage even for otherwise strong LLMs.

6.3.2 Line Purpose Identification

For Task 2 (line purpose identification task), performance on the Spanish subset is uniformly low and lags behind the English subset. As shown 5 and Table 20, OLMo3.1-32B again performs best, but only reaches 19.3% Pass@1 accuracy and 33.5% Pass@5 accuracy, while Qwen3-32B attains 5.2% and 16.5% respectively. Smaller models such as Qwen3-8B, Apertus-8B, and Ministral-8B remain below 10% Pass@1 accuracy, while OLMo-3-7B performs well.

F1 score is slightly lower in Spanish than in English and AUC values hover close to 0.5, similar to English, which may indicate that models frequently misidentified key roles such as punchlines, escalations, and callback in Spanish jokes. Compared to English, in general, the results suggest that the multi-line understanding is particularly brittle in Spanish, likely due to weaker Spanish pretraining and the added difficulty of mapping role labels across languages and cultural references.

6.3.3 Adversarial Robustness

To evaluate adversarial robustness in Spanish, we ran (with slight alterations) the same Task 1 and Task 2 prompts on three perturbation sets: orthographic typos, semantic-drift, and semantic-preserving perturbations. Perturbations were generated using line-level annotations and applied selectively to lines labeled as punchlines, while all other lines were left unchanged regardless of their role. (Unlike English, we did not include a Spanish cultural/dialect shift condition.) We report Pass@1 and Pass@5 metrics, where Pass@1 corresponds to a single completion set and Pass@5 aggregates results across five completion sets.

For Task 1 (overall joke type classification), the strongest models remained relatively stable under Spanish perturbations, with changes on the order of a few percentage points and no consistent pattern of degradation. Qwen3-32B achieves 26.0% Pass@5 accuracy on the unperturbed Spanish set and ranges from 27.2% (semantic drift) to 28.3% (orthographic typos) under perturbations. OLMo3.1-32B is similarly stable overall, with 21.4% Pass@5 at baseline and a range of 19.7% (semantic drift) to 23.7% (semantic-preserving). These results suggest that, for Spanish Task 1, coarse humor category prediction is not strongly disrupted by localized punchline perturbations in our setup, and observed fluctuations are comparable to typical sampling variance across conditions. Lower-performing models again exhibit larger swings; for example, Falcon3-10B increases from 5.2% Pass@5 at baseline to 8.7% under semantic-preserving perturbations, which is more consistent with higher variance and weaker baseline reliability than with a robust improvement.

For Task 2 (line purpose identification), perturbations produce clearer brittleness for some models. Qwen3-32B shows a large degradation under all three Spanish perturbation types: Pass@1 accuracy drops from 5.2% at baseline to 1.9% under orthographic typos and 1.5% under semantic-preserving perturbations (absolute changes of -3.3 to -3.8 percentage points). The effect is even more pronounced at Pass@5: Qwen3-32B decreases from 16.5% at baseline to 5.8% under orthographic typos (a -10.7 percentage point drop, roughly a 65% relative reduction), with similar drops under semantic drift (6.3%) and semantic-preserving (6.8%). In contrast, OLMo3.1-32B is comparatively robust on Spanish Task 2, improving at Pass@1 from 19.3% baseline to approximately 23.1–23.7% across perturbations, while remaining essentially unchanged at Pass@5 (33.5% baseline versus 34.0% under all three perturbations). Averaged across evaluated Spanish models, all perturbations reduce Task 2 performance at Pass@5 accuracy by roughly 3 percentage points on average (with orthographic typos producing the largest mean drop), reinforcing the overall pattern that structured multi-line labeling in Task 2 is more sensitive to localized punchline noise than the single-label prediction required by Task 1.

6.4 English and Spanish

6.4.1 Overall Joke Classification

On the combined English and Spanish set, we observe the same general pattern as in the single-language splits: the 32B models clearly outperform the 8–10b models. Qwen3-32B achieves about 30% accuracy on Pass@5, with OLMo3.1-32B next at about 24%. Among the smaller models, performance remains in the single digits to low teens on Pass@5, with Apertus and Falcon generally performing best within the 8–10B group.

6.4.2 Line Purpose Identification

For line purpose identification on the combined set, OLMo3.1-32B is again the strongest model, reaching about 19% Pass@5 accuracy. Qwen3-32 and Falcon3-10B form a second tier at roughly 11% Pass@5, while the remaining 8B models lag behind. Overall, the combined-set results align with the English and Spanish subsets: this task remains challenging for all models, but larger models retain a consistent advantage.

6.4.3 Adversarial Robustness

Robustness trends on the combined set largely mirror what we observed in the individual language analyses. For overall joke classification, perturbations substantially degrade performance, especially for Qwen3-32B: Pass@5 accuracy drops from about 30% in the unperturbed setting to around 6-7% under orthographic and semantic drift perturbations, and remains similarly low under semantic-preserving perturbations. In contrast, OLMo3.1-32B is noticeably more resilient, maintaining roughly 20-21% Pass@5 accuracy across perturbation types. This suggests that while both 32B models lead on clean data, OLMo3.1-32B is the more robust choice under input noise that targets punchlines.

6.5 Error analysis

6.5.1 English

In the small 6-10 billion parameter models, the primary method of failure in both tasks is failure to produce an answer that could be parsed, or failure to produce any kind of answer at all.

Olmo3-7b and Ministral-8b were particularly notable, as they rarely produced any text. Olmo3-7b only produced outputs for 1% of the prompts in the overall joke classification task, and in those answers, none of them appeared to attempt to answer the question. Instead, Olmo3-7b appeared to attempt to complete the sentence, despite being instruction-tuned. For example, one output was "Don't forget to output your final answer in the specified format!". Ministral-8b only produced a single answer out of the 286 prompts in its first attempt. It did follow instructions, answering with ""category": "satire", "reasoning": "The joke uses exaggeration and absurdity to satirize the dynamics between employees and customers in retail stores, particularly the lack of responsibility and accountability of the employee."". However, it answered incorrectly, as the correct answer was "anecdotal".

Apertus-8b expresses a fairly unique failure mode, as while it produces valid, sometimes extremely long answers for 72% prompts, it only produced an answer containing a valid JSON for 30% of prompts. This indicates that apertus-8b may possess natural language ability greater than what its performance metrics suggest, and is instead restricted by its poorer instruction-following ability.

In the larger 32 billion parameter models, they successfully produced valid outputs for all prompts the overall joke classification task. Here, the models tended to guess "observational/anecdotal" too much, as well as "self-deprecating" and "dry".

For the line purpose identification task, the smaller models' primary failure most is once again failure to produce answers. Olmo3-7B again seems to demonstrate a reversal in behavior to sentence completion, with outputs like "Do not include keyword it's in the response". However, Ministral-8B demonstrates a significant improvement compared to its performance in the joke classification task, answering more prompts while following the prompt and outputting valid JSONs.

The larger 32 billion parameter models started to produce invalid outputs, as they get caught in infinite repetitions or think for too many tokens, hitting the output token limit. For example, in one output, Olmo3.1-32B answered "escalation" repeatedly, until it hit the output token limit, despite being run on developer-recommended sampling parameters.

Beyond this failure mode, the models tended to guess "escalation" or "establishing context" for many of the labels, including for many punchlines. "Punchline" was the label that was incorrectly guessed the most, with Olmo3.1-32B answering incorrectly 97% of the time, and Qwen3-32B answering incorrectly 96% of the time.

6.5.2 Spanish

For Task 1, across the models, many different labels are predicted as "witty", "surreal/absurdism", or "satire/parody/irony", especially at Pass@5. This could mean that when the model is unsure, it tends to classify the joke as something "witty" or "absurd" instead of committing to a more specific joke type.

For the smallest model (OLMo-3-7B), the Pass@1 and especially Pass@5 matrices are almost entirely in the NA column, which could show that the model often fails to output a valid category at all. Larger models like Qwen3-32B and OLMo3.1-32B on Pass@5 show slightly more variation, as they classify more "witty", "surreal/absurdism" and "satire/parody/irony" jokes, and spread predictions across more labels instead of simply resorting to NA. However, even these stronger models still mix up neighboring categories (for example, "surreal" vs. "witty", "satire" vs. "topical"), which helps explain why overall accuracy tops out around 30%.

For Task 2, across models and Pass@1, most lines are predicted as establishing context or escalation, regardless of their true role. Lines that are really "punchlines", "callbacks", "redirections", or “wrap-ups” are often put into “establishing context” and “escalation” columns, which indicates that models recognize what is happening in the joke but default to generic roles instead of more specific roles. This effect is especially visible in smaller models like OLMo-3-7B, where almost every line is mapped to a single column, showing that it often fails to learn more fine-grained label distinctions.

In terms of recognizing punchlines, almost all models under-identify punchline lines. True punchlines are often mislabeled as “escalation”, “timing”, or “establishing context”, and only a small fraction land on the punchline column, even for the strongest models. Similarly, “wrap-up” and “repetition” lines are frequently predicted as “escalation” or “context” rather than “closing”, suggesting that models are poor at locating where a bit of a joke actually belongs.

At Pass@5, larger models such as Qwen3-32B and OLMo3.1-32B spread probability mass over more late-stage roles like “punchline”, “wrap-up”, or “subversion”, but the diagonal for these labels is still relatively weak, indicating persistent uncertainty about where the joke necessarily peaks.

Labels like “callback”, “meta-humor”, “misdirection”, and “redirection” have very low true-positive counts in all matrices. Their rows typically show small numbers scattered across many columns, with no dominant predicted class, meaning models rarely resort to this choice.

Increasing the model size from 7-8B to 32B and moving from Pass@1 to Pass@5 noticeably sharpens the diagonals but does not fundamentally change the error profile. The larger models produce more correct ”punchline”, ”timing”, and ”subversion” predictions and reduce the extreme over-use of ”establishing context”, but still the confusion between neighboring similar choices (like ”escalation” vs ”punchline”, ”timing” vs ”escalation”, ”wrap-up” vs ”punchline”) remains common.

Across both tasks, the confusion matrices show that the models often collapse many fine-grained humor types and categories into a small set of broad labels, which could mean that they capture more coarse tone jokes, but struggle to distinguish more nuanced joke types.

7 Contributions of group members

- Atrey Desai: Created initial idea for project, found relevant dataset and citations for proposed and final tasks. Annotated all of the Spanish data used in the project (including re-annotating Kamala’s subset for IAA purposes). Set up code repository, tooling, and initial vLLM inference scripts. Fixed bugs in prompt generation and multi-GPU model output scripts for large models. Generated model outputs for all normal and perturbed tasks in Spanish, bugs in English metrics, and all metrics for English and Spanish normal and perturbed supersets. Created all graphics & figures and contributed to writing.

- Leo Du: Scraped all transcripts in English. Preprocessed and annotated half of the English datapoints. Tested prompts. Wrote all prompt generation and model output evaluation scripts. Generated model output for both tasks in English and Spanish. Also generated model output for English perturbations. Evaluated model performance on English tasks. Lots of writing.

- James van Doorn: Preprocessed and annotated half of the English datapoints. Designed and implemented the adversarial robustness framework and data perturbation pipelines, producing perturbed English and Spanish joke datasets aligned with original task labels. Conducted and interpreted robustness analysis for English and Spanish, and analyzed overall joke classification, line purpose identification, and robustness results on the combined English+Spanish dataset. Contributed to writing.

- Kamala Sreepada: Worked on extracting the transcriptions for the StandUp4AI dataset, using ASR. Annotated part of the Spanish dataset. Looked for the dataset initially and researched more about multilingual and cross-cultural humor recognition. Contributed to writing various sections of the paper.

8 Conclusion

Small, open-source LLMs have seen improvement in recent years, but they appear to still be a ways off in understanding more nuanced, subtle language.

Of note, this project is focused on benchmark creation and evaluation, rather than model training. As such, we did not use a traditional train/-validation/test split. All models were evaluated in zero-shot capacity to measure underlying reasoning and linguistic capabilities.

We expected these tasks to be challenging for the models, and the 32 billion parameter models performed as expected, but it was surprising how the smaller 7-10 billion failed to produce answers in a large portion of the prompts.

In the future, it would be interesting to investigate if increasing the number of output tokens would affect performance, especially for models that can utilize chain-of-thought reasoning or models that are trained to ”think” using reasoning tokens, as increasing output tokens should theoretically improve performance by allowing for more in-depth reasoning behavior. Assessing large, state-of-the-art LLMs like GPT-5.2, Claude Opus 4.5, and Gemini 3 Pro in this benchmark would also be of interest.

Additionally, the benchmark could be expanded with more examples overall, increased representation of lesser-used labels in both tasks, and richer cultural and cross-linguistic perturbation schemes to better assess robustness across languages and cultural contexts.

9 AI Disclosure

- Did you use any AI assistance to complete this project (report and/or code)? If so, please also specify what AI(s) you used.

  Yes, we used AI assistance for code and report purposes. Specifically:

  - AI assistance was used to complete the initial proposal, primarily in the initial literature review. We used the AI2 Asta research agent to conduct a broad sweep of the field and to narrow down our topic. LLM chat platforms such as ChatGPT and Gemini were used for general subject matter queries and summarizing research papers.
  - For code, we primarily used Cursor (model: auto) for scripting, particularly for downloading YouTube videos, organizing the repository into its current structure, augmenting the bash script to run inference in various methods, augmenting existing human-written evaluation scripts for combining subsets, adding redundancy in the vLLM inference, and creating the README.md.
  - For the report, AI assistance was also used to check grammar and LaTeX issues in this proposal. For the report figures, AI was used to format data to the human-written Vega-Lite template (this was thoroughly checked for every figure to prevent any inaccuracies). GitHub Copilot was used to summarize file changes in git commit messages. ChatGPT was also used to help generate a name for this project.

If you answered yes to the above question, please complete the following as well:

- **Free response:** Describe your overall experience with the AI. Did you use it to generate new text, write code, generate research ideas, check your own ideas, or rewrite text? How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant?

  - AI tools were used primarily for support and refinement. AI2’s Asta research agent was used for an initial literature survey, and chat-based LLMs were used for clarification questions, summarizing papers, and assisting with drafting and revising portions of code and text.
  - AI assistance was especially helpful for debugging and implementation details (e.g., scripting, repository boilerplate, and diagnosing Python/model errors), including suggestions related to parallelism for larger model runs. We also used AI to help label the numerous tables in the appendix.
  - Overall, all AI outputs required human review and revision, and final decisions, implementations, and conclusions were made by the authors. Generally, though better than earlier AI tools, present-day AI models still require a degree of suspicion and guidance.

References

Amiriparian, S., Christ, L., Kathan, A., Gerczuk, M., Muller, N., Klug, S., Stappen, L., Konig, A., Cambria, E., Schuller, B. W., and Eulitz, S. (2024). The muse 2024 multimodal sentiment analysis challenge: Social perception and humor recognition. *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*.

Anonymous (2025). PunmemeCN: A benchmark to explore vision-language models' understanding of chinese pun memes. In *Submitted to ACL Rolling Review - May 2025*. under review.

Barriere, V., Gomez, N., Hemamou, L., Callejas, S., and Ravenet, B. (2025). Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. *ArXiv*, abs/2505.18903.

Choi, M., Pei, J., Kumar, S., Shu, C., and Jurgens, D. (2023). Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *Conference on Empirical Methods in Natural Language Processing*.

Horvitz, Z., Chen, J., Aditya, R., Srivastava, H., West, R., Yu, Z., and McKeown, K. (2024). Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., and Magdy, W. (2021). Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Mihalcea, R. and Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In Mooney, R., Brew, C., Chien, L.-F., and Kirchhoff, K., editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nandy, A., Agarwal, Y., Patwa, A., Das, M. M., Bansal, A., Raj, A., Goyal, P., and Ganguly, N. (2024a). Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models.

Nandy, A., Agarwal, Y., Patwa, A., Das, M. M., Bansal, A., Raj, A., Goyal, P., and Ganguly, N. (2024b). Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models.

Narad, R., Suresh, S., Chen, J., Dysart-Bricken, P. S., Mankoff, B., Nowak, R. D., Zhang, J., and Jain, L. (2025). Which llms get the joke? probing non-stem reasoning abilities with humorbench. *ArXiv*, abs/2507.21476.

Quan, K., Ramakrishnan, P., and Chin, J. (2025). Can ai take a joke—or make one? a study of humor generation and recognition in llms. *Proceedings of the 2025 Conference on Creativity and Cognition*.

Rayz, J. (2004). Computationally recognizing wordplay in jokes. *Cognitive Science - COGSCI*.

Trott, S., Walker, D. E., Taylor, S. M., and Coulson, S. (2025). Turing jest: Distributional semantics and one-line jokes. *Cognitive Science*, 49.

Yang, X., Liu, W., Bailey, J., Tao, D., and Liu, W. (2023). Semantic-preserving adversarial text attacks. In *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. X, NO. X, X X*.

Yue, S., Song, S., Cheng, X., and Hu, H. (2024). Do large language models understand conversational implicature- a case study with a chinese sitcom. In *China National Conference on Chinese Computational Linguistics*.

Zangari, A., Marcuzzo, M., Albarelli, A., Pilehvar, M. T., and Camacho-Collados, J. (2025). Pun unintended: Llms and the illusion of humor understanding. arXiv. arXiv:2509.12158v1 [cs.CL].

Zheng, Y., Li, S., Wu, F., Ziyi, Y., Lin, H., Hu, Z., Cai, X., Wang, Z., Chen, J., Luan, S., Xu, J., and Chen, L. (2025). Fanchuan: A multilingual and graph-structured benchmark for parody detection and analysis. *ArXiv*, abs/2502.16503.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Gong, N., Zhang, Y., and Xie, X. (2023). Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*.

A Appendix

This appendix contains the performance tables for all tasks, subsets, and adversarial perturbation categories across the seven evaluated models.

Note on Language Proficiency: The Spanish subset was annotated by experienced but non-native speakers. Annotations were cross-checked with translation tools to ensure accuracy, but this may introduce slight biases compared to native intuition and understanding of humor.

Error Analysis: A detailed error analysis of the English and Spanish tasks was conducted by analyzing confusion matrices.

- **English Confusion Matrices:** Available at https://github.com/atreydesai/humorbench/tree/main/results/en/confusion_matrices.
- **Spanish Confusion Matrices:** Available at https://github.com/atreydesai/humorbench/tree/main/results/es/confusion_matrices.

A.1 Full Set (English and Spanish)

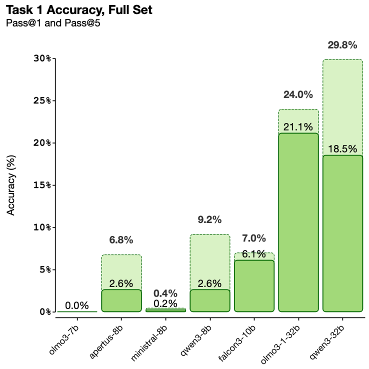

Figure 6: Task 1 Accuracy

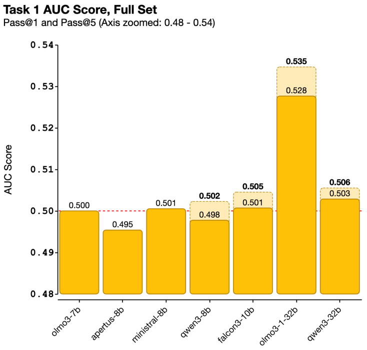

Figure 7: Task 1 AUC Score

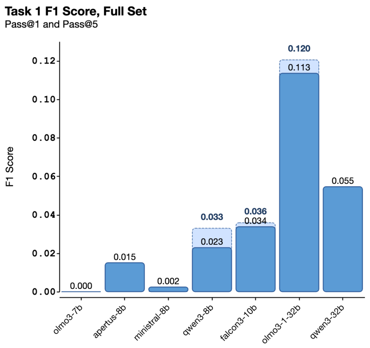

Figure 8: Task 1 F1 Score

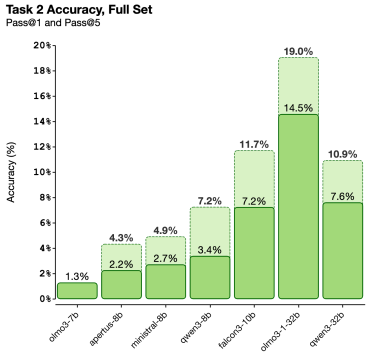

Figure 9: Task 2 Accuracy

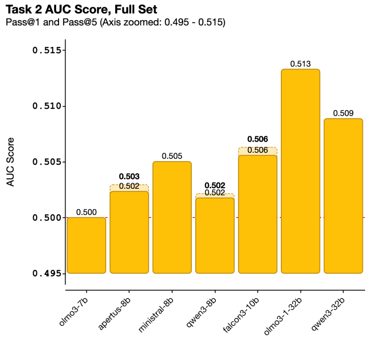

Figure 10: Task 2 AUC Score

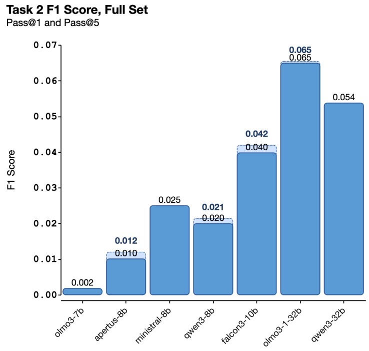

Figure 11: Task 2 F1 Score

Appendix Figure 2: Comprehensive performance metrics for Full Set (English and Spanish) across Task 1 (Top) and Task 2 (Bottom).

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.026	0.092	0.023	0.033	0.498	0.502
Qwen3-32B	0.185	0.298	0.055	0.047	0.503	0.506
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.211	0.240	0.113	0.120	0.528	0.535
Falcon3-10B	0.061	0.070	0.034	0.036	0.501	0.505
Apertus-8B	0.026	0.068	0.015	0.007	0.495	0.493
Ministral-8B	0.002	0.004	0.002	0.001	0.501	0.500

Table 1: Performance metrics for Full Set of Task 1: Overall Joke Classification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.034	0.072	0.020	0.021	0.502	0.502
Qwen3-32B	0.076	0.109	0.054	0.053	0.509	0.509
OLMo3-7B	0.013	0.013	0.002	0.002	0.500	0.500
OLMo3.1-32B	0.145	0.190	0.065	0.065	0.513	0.513
Falcon3-10B	0.072	0.117	0.040	0.042	0.506	0.506
Apertus-8B	0.022	0.043	0.010	0.012	0.502	0.503
Ministral-8B	0.027	0.049	0.025	0.018	0.505	0.503

Table 2: Performance metrics for Full Set of Task 2: Line Purpose Identification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.039	0.081	0.033	0.036	0.501	0.503
Qwen3-32B	0.190	0.301	0.065	0.052	0.513	0.508
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.216	0.240	0.103	0.102	0.527	0.525
Falcon3-10B	0.065	0.078	0.023	0.027	0.496	0.499
Apertus-8B	0.052	0.109	0.020	0.011	0.509	0.510
Ministral-8B	0.013	0.065	0.023	0.015	0.502	0.501

Table 3: Full Set performance for Task 1 (Joke Classification) under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.028	0.059	0.014	0.016	0.502	0.502
Qwen3-32B	0.050	0.066	0.034	0.034	0.504	0.504
OLMo3-7B	0.013	0.013	0.002	0.002	0.500	0.500
OLMo3.1-32B	0.165	0.198	0.070	0.070	0.517	0.517
Falcon3-10B	0.056	0.097	0.033	0.035	0.501	0.501
Apertus-8B	0.022	0.038	0.009	0.009	0.501	0.501
Ministral-8B	0.023	0.031	0.011	0.008	0.503	0.501

Table 4: Full Set performance for Task 2 (Line Purpose) under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.028	0.072	0.020	0.026	0.495	0.498
Qwen3-32B	0.196	0.298	0.084	0.077	0.520	0.519
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.200	0.233	0.094	0.094	0.520	0.518
Falcon3-10B	0.087	0.102	0.043	0.046	0.507	0.510
Apertus-8B	0.059	0.120	0.019	0.009	0.512	0.509
Ministral-8B	0.028	0.054	0.027	0.014	0.505	0.501

Table 5: Full Set performance for Task 1 (Joke Classification) under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.030	0.040	0.018	0.019	0.501	0.501
Qwen3-32B	0.055	0.071	0.035	0.036	0.501	0.501
OLMo3-7B	0.013	0.013	0.002	0.002	0.500	0.500
OLMo3.1-32B	0.160	0.196	0.067	0.068	0.513	0.513
Falcon3-10B	0.058	0.082	0.036	0.038	0.500	0.502
Apertus-8B	0.024	0.041	0.013	0.014	0.503	0.503
Ministral-8B	0.021	0.031	0.009	0.009	0.500	0.499

Table 6: Full Set performance for Task 2 (Line Purpose) under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.037	0.087	0.038	0.037	0.504	0.504
Qwen3-32B	0.185	0.283	0.068	0.056	0.514	0.510
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.214	0.244	0.105	0.104	0.528	0.525
Falcon3-10B	0.102	0.148	0.043	0.046	0.504	0.511
Apertus-8B	0.063	0.122	0.016	0.009	0.503	0.505
Ministral-8B	0.026	0.054	0.023	0.021	0.502	0.503

Table 7: Full Set performance for Task 1 (Joke Classification) under Semantic-Preserving perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.035	0.047	0.020	0.020	0.499	0.500
Qwen3-32B	0.049	0.064	0.035	0.035	0.503	0.503
OLMo3-7B	0.013	0.013	0.002	0.002	0.500	0.500
OLMo3.1-32B	0.150	0.214	0.069	0.068	0.518	0.516
Falcon3-10B	0.052	0.109	0.032	0.033	0.500	0.501
Apertus-8B	0.022	0.038	0.013	0.012	0.502	0.502
Ministral-8B	0.021	0.038	0.011	0.011	0.501	0.501

Table 8: Full Set performance for Task 2 (Line Purpose) under Semantic-Preserving perturbations.

A.2 English Subset

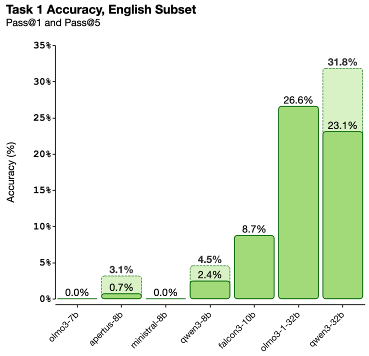

Figure 12: Task 1 Accuracy

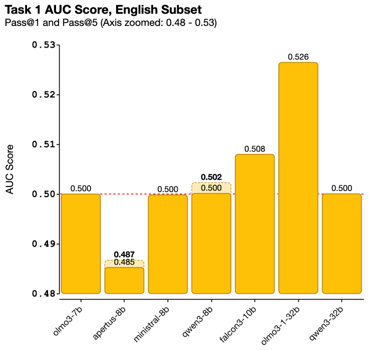

Figure 13: Task 1 AUC Score

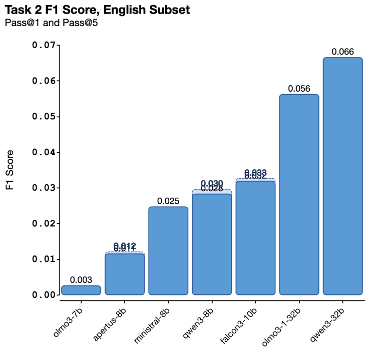

Figure 14: Task 1 F1 Score

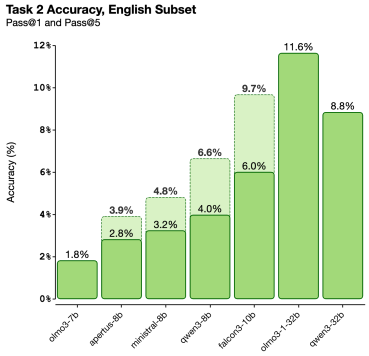

Figure 15: Task 2 Accuracy

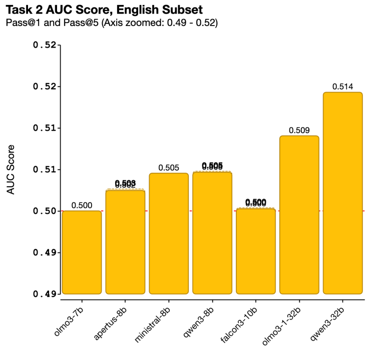

Figure 16: Task 2 AUC Score

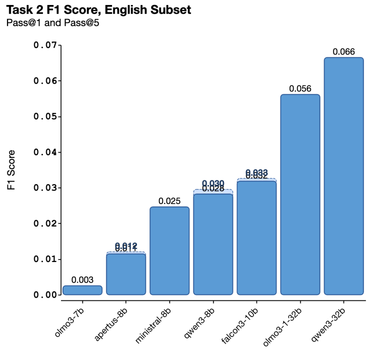

Figure 17: Task 2 F1 Score

Appendix Figure 1: Comprehensive performance metrics for English Subset across Task 1 (Top) and Task 2 (Bottom).

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.024	0.045	0.027	0.035	0.501	0.503
Qwen3-32B	0.231	0.308	0.052	0.048	0.501	0.500
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.266	0.266	0.090	0.090	0.525	0.525
Falcon3-10B	0.080	0.080	0.031	0.031	0.506	0.506
Apertus-8B	0.007	0.031	0.001	0.003	0.487	0.489
Ministral-8B	0.000	0.000	0.000	0.000	0.500	0.500

Table 9: Performance metrics for English Subset of Task 1: Overall Joke Classification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.040	0.066	0.028	0.030	0.505	0.505
Qwen3-32B	0.088	0.088	0.066	0.066	0.514	0.514
OLMo3-7B	0.018	0.018	0.003	0.003	0.500	0.500
OLMo3.1-32B	0.116	0.116	0.056	0.056	0.509	0.509
Falcon3-10B	0.060	0.097	0.032	0.033	0.500	0.500
Apertus-8B	0.028	0.038	0.011	0.012	0.502	0.503
Ministral-8B	0.031	0.047	0.024	0.023	0.505	0.504

Table 10: Performance metrics for English Subset of Task 2: Line Purpose Identification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.021	0.038	0.029	0.024	0.506	0.501
Qwen3-32B	0.234	0.311	0.055	0.055	0.515	0.511
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.269	0.269	0.086	0.086	0.519	0.519
Falcon3-10B	0.143	0.143	0.042	0.042	0.507	0.507
Apertus-8B	0.063	0.063	0.038	0.038	0.511	0.511
Ministral-8B	0.003	0.003	0.006	0.006	0.501	0.501

Table 11: English Subset of Task 1 (Joke Classification) performance under Cultural Shift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.039	0.039	0.029	0.029	0.506	0.506
Qwen3-32B	0.072	0.119	0.057	0.056	0.511	0.512
OLMo3-7B	0.018	0.018	0.003	0.003	0.500	0.500
OLMo3.1-32B	0.125	0.125	0.057	0.057	0.509	0.509
Falcon3-10B	0.066	0.096	0.041	0.041	0.502	0.502
Apertus-8B	0.023	0.023	0.010	0.010	0.501	0.501
Ministral-8B	0.026	0.056	0.011	0.013	0.500	0.501

Table 12: English Subset of Task 2 (Line Purpose) performance under Cultural Shift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.031	0.031	0.028	0.028	0.505	0.505
Qwen3-32B	0.224	0.332	0.048	0.051	0.498	0.500
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.276	0.276	0.082	0.082	0.517	0.517
Falcon3-10B	0.094	0.094	0.031	0.031	0.506	0.506
Apertus-8B	0.042	0.042	0.022	0.022	0.511	0.511
Ministral-8B	0.000	0.000	0.000	0.000	0.500	0.500

Table 13: English Subset of Task 1 (Joke Classification) performance under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.038	0.065	0.020	0.021	0.503	0.503
Qwen3-32B	0.066	0.066	0.043	0.043	0.508	0.508
OLMo3-7B	0.018	0.018	0.003	0.003	0.500	0.500
OLMo3.1-32B	0.126	0.126	0.059	0.059	0.512	0.512
Falcon3-10B	0.068	0.097	0.044	0.042	0.504	0.502
Apertus-8B	0.021	0.024	0.006	0.006	0.500	0.500
Ministral-8B	0.032	0.038	0.013	0.011	0.503	0.502

Table 14: English Subset of Task 2 (Line Purpose) performance under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.021	0.021	0.010	0.010	0.496	0.496
Qwen3-32B	0.245	0.318	0.081	0.074	0.520	0.517
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.259	0.259	0.087	0.087	0.513	0.513
Falcon3-10B	0.119	0.119	0.040	0.040	0.510	0.510
Apertus-8B	0.035	0.066	0.015	0.013	0.500	0.501
Ministral-8B	0.000	0.000	0.000	0.000	0.500	0.500

Table 15: English Subset of Task 1 (Joke Classification) performance under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.041	0.041	0.027	0.027	0.504	0.504
Qwen3-32B	0.074	0.074	0.053	0.053	0.513	0.513
OLMo3-7B	0.018	0.018	0.003	0.003	0.500	0.500
OLMo3.1-32B	0.122	0.122	0.058	0.058	0.511	0.511
Falcon3-10B	0.069	0.069	0.048	0.048	0.509	0.509
Apertus-8B	0.028	0.031	0.015	0.013	0.504	0.503
Ministral-8B	0.028	0.036	0.012	0.011	0.500	0.500

Table 16: English Subset of Task 2 (Line Purpose) performance under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.024	0.031	0.019	0.025	0.500	0.504
Qwen3-32B	0.231	0.308	0.059	0.053	0.505	0.505
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.266	0.266	0.084	0.084	0.517	0.517
Falcon3-10B	0.143	0.203	0.058	0.051	0.518	0.517
Apertus-8B	0.059	0.059	0.022	0.022	0.506	0.506
Ministral-8B	0.000	0.000	0.000	0.000	0.499	0.499

Table 17: English Subset of Task 1 (Joke Classification) performance under Semantic-Preserving perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.046	0.046	0.031	0.031	0.504	0.504
Qwen3-32B	0.067	0.067	0.051	0.051	0.513	0.513
OLMo3-7B	0.018	0.018	0.003	0.003	0.500	0.500
OLMo3.1-32B	0.113	0.149	0.061	0.060	0.513	0.512
Falcon3-10B	0.059	0.111	0.040	0.038	0.502	0.501
Apertus-8B	0.025	0.025	0.010	0.010	0.501	0.501
Ministral-8B	0.028	0.047	0.014	0.015	0.501	0.501

Table 18: English Subset of Task 2 (Line Purpose) performance under Semantic-Preserving perturbations.

A.3 Spanish Subset

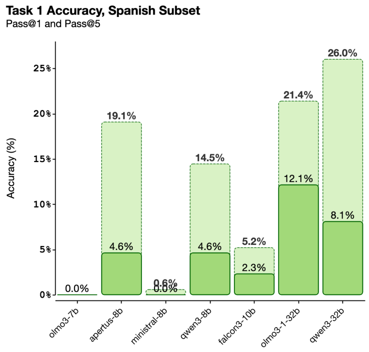

Figure 18: Task 1 Accuracy

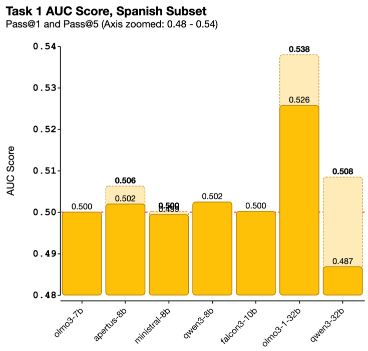

Figure 19: Task 1 AUC Score

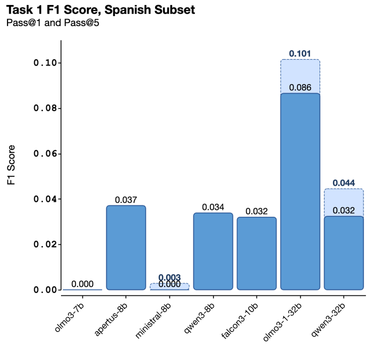

Figure 20: Task 1 F1 Score

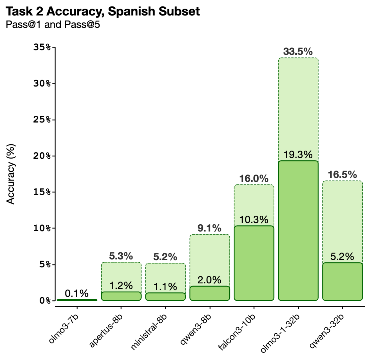

Figure 21: Task 2 Accuracy

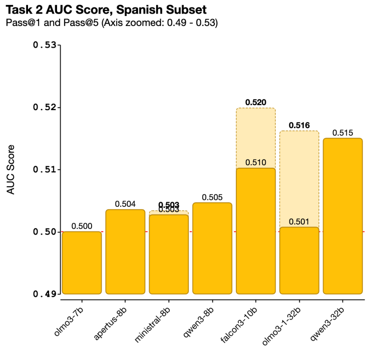

Figure 22: Task 2 AUC Score

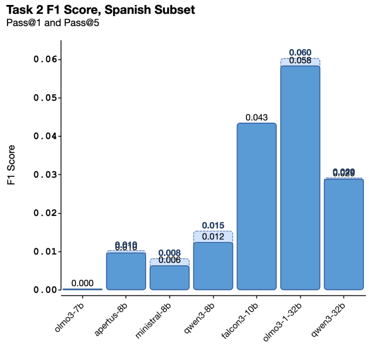

Figure 23: Task 2 F1 Score

Appendix Figure 2: Comprehensive performance metrics for Spanish Subset across Task 1 (Top) and Task 2 (Bottom).

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.046	0.145	0.034	0.033	0.502	0.502
Qwen3-32B	0.081	0.260	0.032	0.044	0.487	0.508
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.121	0.214	0.086	0.101	0.526	0.538
Falcon3-10B	0.023	0.052	0.032	0.022	0.500	0.496
Apertus-8B	0.046	0.191	0.037	0.025	0.502	0.506
Ministral-8B	0.000	0.006	0.000	0.003	0.499	0.500

Table 19: Performance metrics for Spanish Subset of Task 1: Overall Joke Classification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.020	0.091	0.012	0.015	0.505	0.502
Qwen3-32B	0.052	0.165	0.029	0.029	0.515	0.507
OLMo3-7B	0.001	0.001	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.193	0.335	0.058	0.060	0.501	0.516
Falcon3-10B	0.103	0.160	0.043	0.040	0.510	0.520
Apertus-8B	0.012	0.053	0.010	0.010	0.504	0.496
Ministral-8B	0.011	0.052	0.006	0.008	0.503	0.503

Table 20: Performance metrics for Spanish Subset of Task 2: Line Purpose Identification.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.075	0.156	0.079	0.053	0.517	0.505
Qwen3-32B	0.104	0.283	0.055	0.053	0.505	0.516
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.133	0.231	0.086	0.085	0.523	0.525
Falcon3-10B	0.040	0.064	0.033	0.025	0.502	0.499
Apertus-8B	0.058	0.249	0.020	0.012	0.501	0.506
Ministral-8B	0.029	0.156	0.023	0.024	0.492	0.499

Table 21: Spanish Subset of Task 1 (Joke Classification) performance under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.009	0.046	0.004	0.007	0.502	0.499
Qwen3-32B	0.019	0.058	0.017	0.009	0.506	0.496
OLMo3-7B	0.001	0.001	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.236	0.340	0.068	0.063	0.508	0.507
Falcon3-10B	0.033	0.097	0.021	0.020	0.515	0.512
Apertus-8B	0.022	0.062	0.015	0.010	0.506	0.504
Ministral-8B	0.007	0.019	0.006	0.003	0.502	0.501

Table 22: Spanish Subset of Task 2 (Line Purpose) performance under Orthographic perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.081	0.162	0.063	0.049	0.509	0.505
Qwen3-32B	0.116	0.272	0.063	0.061	0.509	0.514
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.139	0.197	0.088	0.071	0.525	0.517
Falcon3-10B	0.029	0.075	0.031	0.034	0.496	0.504
Apertus-8B	0.110	0.254	0.036	0.012	0.526	0.510
Ministral-8B	0.040	0.139	0.032	0.021	0.500	0.497

Table 23: Spanish Subset of Task 1 (Joke Classification) performance under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.009	0.045	0.003	0.007	0.501	0.503
Qwen3-32B	0.018	0.063	0.008	0.008	0.484	0.496
OLMo3-7B	0.001	0.001	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.237	0.340	0.076	0.069	0.530	0.514
Falcon3-10B	0.032	0.113	0.021	0.021	0.516	0.512
Apertus-8B	0.013	0.057	0.008	0.009	0.503	0.496
Ministral-8B	0.009	0.025	0.004	0.004	0.502	0.502

Table 24: Spanish Subset of Task 2 (Line Purpose) performance under Semantic Drift perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.081	0.185	0.065	0.066	0.512	0.511
Qwen3-32B	0.139	0.277	0.076	0.064	0.522	0.516
OLMo3-7B	0.000	0.000	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.139	0.237	0.093	0.092	0.549	0.535
Falcon3-10B	0.052	0.087	0.049	0.040	0.508	0.507
Apertus-8B	0.064	0.249	0.016	0.010	0.497	0.499
Ministral-8B	0.052	0.145	0.040	0.028	0.504	0.499

Table 25: Spanish Subset of Task 1 (Joke Classification) performance under Semantic-Preserving perturbations.

Model	Pass@1 Acc	Pass@5 Acc	Pass@1 F1	Pass@5 F1	Pass@1 AUC	Pass@5 AUC
Qwen3-8B	0.013	0.059	0.007	0.008	0.503	0.503
Qwen3-32B	0.015	0.068	0.006	0.009	0.503	0.500
OLMo3-7B	0.001	0.001	0.000	0.000	0.500	0.500
OLMo3.1-32B	0.231	0.340	0.074	0.066	0.531	0.515
Falcon3-10B	0.040	0.105	0.021	0.020	0.517	0.515
Apertus-8B	0.015	0.065	0.010	0.008	0.504	0.500
Ministral-8B	0.007	0.020	0.004	0.002	0.501	0.501

Table 26: Spanish Subset of Task 2 (Line Purpose) performance under Semantic-Preserving perturbations.