

Designing a Conservative Humor Filter

Can a Model Tell If an Image Caption Is Funny?

Stanford CS224N Custom Project

Michael Roger

Department of Computer Science

Stanford University

michael.roger@stanford.edu

Abstract

Humor judgment is difficult because it is subjective, audience-dependent, and often requires both linguistic and visual context. Generic caption-evaluation signals do not directly model funniness, and text-only image proxies may miss the visual cue that makes a joke land. I study this problem on weekly Columbia humor studies from a 180+ student class I teach, where students vote on AI-generated captions for uploaded images. I compare a base Qwen setup against two stronger cascades that append a frozen GPT-generated image-caption alignment score and image-caption alignment score reasoning. Across text-only and vision-language Qwen models, the reasoning-augmented cascade is strongest in broad development experiments. I also compare against prompt-only GPT, Gemini, and Claude baselines to test whether fine-tuning adds value over frontier multimodal judging.

The final system is evaluated as a conservative filter that should surface likely-funny captions while abstaining on weaker cases. Under that objective, the best precision-first development-time model is text-only and uses the caption, simple image description, image-caption alignment score, and alignment reasoning, while the strongest balanced- F_1 model is a vision-language variant. On a later untouched holdout study round, recall drops sharply and the strongest learned holdout model shifts to a simpler vision-language variant. The current judge is therefore overly selective but still tends to surface funny captions when it makes a positive prediction, which is ideal for the end-user.

Overall, the paper supports a narrow claim: a Qwen-based cascade that consumes frozen teacher-generated image-caption alignment reasoning improves downstream humor filtering on this task, but threshold calibration across study rounds remains the main weakness.

1 Key Information

Project type: Custom. **Mentor:** Fang Wu (fangwu97@stanford.edu). **External collaborators/-mentor:** None. **Sharing project with another class:** No.

2 Introduction

My goal is to build a humor filter that decides whether an image-caption pair is funny. In this project, “funny” is operationalized as receiving more upvotes than downvotes in weekly humor studies from a class of 180+ students that I currently teach at Columbia University. This is a useful NLP problem because it combines multimodal reasoning, subjective human preference, and a deployment constraint that is stricter than ordinary binary classification: in practice, I would rather reject some funny captions than show unfunny ones.

This task is difficult for two reasons. First, humor is audience-specific: the same caption may land differently across communities, even for the same image. Second, text-only descriptions of an image may miss precisely the visual detail that makes a caption succeed or fail. For example, a caption like “woke up to find all seven alarms went on strike” only works if the model connects a stern child pose to mock indignation about oversleeping. Standard caption-evaluation signals mostly measure semantic fit or broad preference, not whether people will actually find a caption funny.

These observations suggest two interventions. One is to provide richer structured context about image-caption alignment. The other is to compare text-only image proxies against true vision-language models that see the image itself. They also motivate a practical benchmark: if strong GPT, Gemini, or Claude models can already judge these pairs well with prompting alone, then task-specific fine-tuning may not add much value.

I therefore designed three suites of controlled fine-tuning experiments: CaptionAndImageOnly, PlusAlignScore, and PlusAlignScoreAndReasoning. These correspond to the internal experiment suites test1, test8, and test9 in the codebase. CaptionAndImageOnly is the base condition: text models receive only the caption and simple image description, VL models receive only the caption and downloaded image bytes, and the supervision target is the binary funny/not funny label. PlusAlignScore adds a scalar image-caption alignment score. PlusAlignScoreAndReasoning further adds free-text reasoning for that alignment score. Across these suites, I compare text-only and VL Qwen backbones, then narrow the search through finalist sweeps and matched-seed confirmation. The stronger PlusAlignScore/PlusAlignScoreAndReasoning systems are therefore cascades rather than standalone humor judges: at inference time they depend on frozen GPT-generated alignment fields in addition to the Qwen classifier itself. The final stage aligns model selection with the deployment objective by choosing checkpoints and thresholds for precision under a validation recall floor.

Contributions.

- I construct a custom humor-judging benchmark with study-level train/validation/test splits that reduce leakage across study rounds while preserving the human-preference nature of the task.
- I perform a systematic comparison of three input suites and both text-only and VL Qwen families through broad suite sweeps, finalist hyperparameter search, and matched-seed confirmation.
- I benchmark the learned models against prompt-only GPT, Gemini, and Claude frontier baselines to test whether task-specific fine-tuning improves over strong non-fine-tuned multimodal judges.
- I show that frozen teacher-generated alignment reasoning is more useful than a scalar teacher score alone for downstream conservative filtering, and that metric alignment changes the final answer: the best balanced- F_1 model is not the best model for a conservative “show only likely-funny captions” policy.

3 Related Work

Zhang et al. [1] show that humor captioning is a difficult benchmark even at very large scale, and that creative tasks require dedicated evaluation rather than assuming that standard alignment methods transfer cleanly. Hasan et al. [2] make a complementary point from multimodal humor detection: humor often depends on interactions between text and visual or prosodic cues rather than on text alone. My setting is narrower than both of these lines of work, but it asks a related question in a more deployment-oriented form: can a model conservatively surface captions that human raters will actually reward in a realistic humor-voting setting?

The project is also related to learning from human preferences. Preference learning and reward modeling often use pairwise human-preference supervision to approximate a human judgment function [3]. My task is simpler than full reward modeling for generation, but it shares the same core challenge: map messy human signals to a useful scoring function. The final system is therefore best interpreted as a conservative humor filter rather than a universal estimator of funniness.

A third relevant thread is selective prediction and calibration. SelectiveNet [4] frames prediction with abstention as a risk-coverage tradeoff rather than a forced-decision problem, which matches my filtering objective directly. Guo et al. [5] show that modern neural networks can be poorly calibrated even when standard accuracy metrics look acceptable, which helps explain why thresholds tuned on validation transfer poorly to the later study 16 shift.

On the modeling side, I rely on parameter-efficient fine-tuning with LoRA [6] and QLoRA-style 4-bit training [7], implemented with the Hugging Face Transformers stack [8]. I compare released Qwen3.5 and Qwen3-VL checkpoints that I fine-tuned in text-only and VL settings [9, 10], and I also use external multimodal foundation models as frozen teachers or prompt-only baselines, including GPT-4o mini, GPT-5-family models, Gemini Flash models, and Claude Sonnet models [11, 12, 13, 14]. This makes the central comparison more precise than “text versus vision” in the abstract: it is direct image access versus strong textualized proxies from external models, with optional teacher reasoning.

Suite	Allowed non-label fields	Text prompt fields	VL prompt fields
1	caption + simple image description (text) / caption + image bytes (VL)	simple image description, caption	caption + downloaded image bytes
2	CaptionAndImageOnly inputs + image alignment score	simple image description, caption, alignment score	caption, alignment score + downloaded image bytes
3	PlusAlignScore inputs + alignment score reasoning	simple image description, caption, alignment score, reasoning	caption, alignment score, reasoning + downloaded image bytes

Table 1: Controlled suite design. Suite 1 = CaptionAndImageOnly, Suite 2 = PlusAlignScore, and Suite 3 = PlusAlignScoreAndReasoning. The label `is_funny` is always the supervision target and is never passed as prompt input.

4 Approach

4.1 Task Formulation

Each example contains a caption, image context, and a binary label $y \in \{\text{funny}, \text{not funny}\}$. The label is derived from vote aggregation: an example is `funny` if it received more upvotes than downvotes. I use a binary label rather than vote-share regression because the downstream product decision is itself binary: either surface the caption or abstain. The training objective is next-token prediction over the label string, but inference is score-based rather than decode-only.

For an input x , I score the two candidate labels by their conditional log probabilities:

$$\ell_f = \log p_\theta(\text{funny} \mid x), \quad \ell_n = \log p_\theta(\text{not funny} \mid x).$$

I then define

$$p_{\text{funny}}(x) = \frac{\exp(\ell_f)}{\exp(\ell_f) + \exp(\ell_n)}.$$

Because inference compares only these two fixed label strings rather than open-ended generations, any label-length bias is constant across examples and the final operating point is still tuned on validation. This score supports a deployment policy with abstention: show the caption only if $p_{\text{funny}}(x) \geq \tau$.

4.2 Suite Design

The three suites differ only in which non-label fields are exposed to the model; all three predict the same binary `is_funny` target. This isolates the effect of richer alignment context. For all suites, I enforce strict field control: no extra metadata is passed beyond the declared fields, and VL models receive downloaded image bytes rather than the raw `image_url` string.

Feature provenance and leakage control.

The auxiliary fields in `PlusAlignScore` and `PlusAlignScoreAndReasoning` were generated offline and frozen before fine-tuning. `simple_image_description` was produced once per image by GPT-4o mini [11]. `image_alignment_score` and `image_alignment_score_reasoning` were then produced by a GPT-5-family model from the pair (`simple_image_description`, `caption`) using a fixed JSON-only prompt about emotional match, reaction exploitation, mismatch type, humor mechanism, and possible improvements [12]. For example, for the caption “woke up to find all seven alarms went on strike” paired with a stern child image, the `humor_mechanism` field described “personification of alarms paired with a sassy, unimpressed reaction to oversleeping chaos.” These fields did not consume vote totals, `is_funny`, split membership, or downstream model predictions. The suite comparison should therefore be interpreted as an ablation on frozen external LLM-generated auxiliary judgments, not only as hand-crafted metadata engineering.

Inference dependency and cascade cost.

The stronger systems are cascades, not standalone classifiers. For text models, `CaptionAndImageOnly` needs one cached GPT-4o mini description per image, while `PlusAlignScore`/`PlusAlignScoreAndReasoning` also need one cached GPT-5 alignment pass per image-caption pair; VL `CaptionAndImageOnly` can score from raw image bytes alone, but VL `PlusAlignScore`/`PlusAlignScoreAndReasoning` still require the GPT fields. This extra stage is still worth comparing because the learned filters beat directly thresholding the raw teacher score on development-time precision and on holdout AUPRC/AUROC. The raw alignment baseline remains the cheaper fallback.

4.3 Models and Training

I ran broad sweeps over 13 model families per suite: 9 text-only models and 4 VL models. The text families included Qwen2.5-7B-Instruct, Qwen3 4B/8B variants, and Qwen3.5 4B/9B variants. The

Split	Rows	Funny	Not funny	Positive rate	Studies
Train	1581	364	1217	0.230	2, 3, 4, 5, 6, 7, 8
Validation	220	36	184	0.164	9, 10
Development test	190	50	140	0.263	11, 15
Reserved holdout	100	57	43	0.570	16

Table 2: Study-level data split and reserved holdout. Broad suite comparisons and baseline calibration use the train/validation/development-test split; final frozen-finalist evaluation uses study 16.

VL families included Qwen3-VL 4B/8B Instruct and Thinking variants. For the open-weight Qwen checkpoints, the parameter scale is reported directly in the model name (4B, 7B, 8B, 9B); for the commercial GPT/Gemini/Claude baselines, the providers do not publish parameter counts, so I report the exact provider model IDs instead. All models were fine-tuned with LoRA adapters on Modal using NVIDIA L40S GPUs, 4-bit NF4 quantization, bfloat16 compute, gradient checkpointing, and paged AdamW 8-bit optimization.

The default text setting used 3 epochs, learning rate 2×10^{-4} , LoRA rank 16, LoRA alpha 32, dropout 0.05, batch size 2, and gradient accumulation 8. The VL setting used the same optimizer and LoRA defaults with batch size 1 and gradient accumulation 16. Finalist sweeps varied learning rate, epoch count, and LoRA dropout.

4.4 Model Selection

Model selection evolved in three stages. Initially, selecting checkpoints by validation loss produced many collapsed “always not funny” models, because loss and accuracy can look acceptable under class imbalance even when positive-class performance is poor. I therefore moved broad comparison to validation F_1 on the positive class.

Once the deployment goal became clearer, I further aligned selection with the product objective. For each checkpoint, I tune a validation threshold

$$\tau^* = \arg \max_{\tau} \text{Precision}_{\text{val}}(\tau) \quad \text{s.t.} \quad \text{Recall}_{\text{val}}(\tau) \geq 0.15.$$

If no threshold satisfies the recall floor, the fallback is the highest-precision threshold. This gives a conservative classifier whose operating point is chosen on validation only.

5 Experiments

5.1 Data

The dataset comes from weekly humor studies run in a class of 180+ students that I currently teach at Columbia University, through The Humor Project course website and the associated caption-voting platform Humor Study [15, 16]. Each record contains an image-caption pair plus derived metadata such as a simple image description and image-caption alignment features. The weekly studies were intentionally heterogeneous rather than generated from one fixed prompt recipe: some rounds emphasized generic memes, some leaned on Columbia-specific jokes or pop-culture references, and some image sets were chosen precisely because they seemed awkward or unlikely to produce strong captions. The final supervised splits are grouped by `study_id`, not by random row, to reduce leakage across related rounds of captioning and voting.

The main reported split is intentionally challenging: validation and the earlier development test come from later study rounds with different image sets and slightly different label balances. Study IDs are sparse database identifiers rather than contiguous experiment numbers, so there were no studies 1 or 12–14 to omit. Images repeat within a study because multiple captions are voted on for the same underlying photo, so the split is by `study_id` rather than by row. This prevents a single weekly study round from being split across train/validation/test, but it does not make validation image-disjoint from train: validation shares 11 image IDs with training (99/220 rows), whereas studies 11/15 share no training images. Each row also has substantial human feedback: there are no zero-vote examples, mean vote counts are 140.5/125.0/108.0/66.3 for train/validation/development-test/holdout, and ties are labeled not funny because funny requires a strict upvote majority.

After the main iteration phase, I reserved study 16 as a one-shot untouched holdout and evaluated the frozen finalists on it with validation-selected thresholds fixed. It is fully image-disjoint from train and validation, but it is not fully image-disjoint from the earlier development test: 14 image

IDs overlap, covering 70/100 holdout rows. I therefore treat study 16 as a cleaner caption-level holdout than studies 11/15, but not as a perfectly image-blind final benchmark. The holdout is still distributionally different, with a 0.57 positive rate; study 16 also included especially timely jokes sourced from Sidechat, which likely increased both its freshness and its mismatch with earlier rounds. That shift remains important later: thresholds tuned on a 0.164-positive validation split remain too conservative on study 16.

5.2 Evaluation Method

The early broad sweeps used positive-class F_1 as the primary model-selection metric because accuracy was misleading under imbalance. For the deployment-oriented stage, I report five metrics on the development test and on the reserved study 16 holdout: **precision_funny**, **recall_funny**, F_1 , **false-positive rate (FPR)**, and **coverage**. Coverage is the fraction of examples predicted funny; for a conservative filtering system, it is the operational cost of abstention. The recall floor of 0.15 is therefore a deployment heuristic rather than a strong statistical guarantee; on this validation split it corresponds to only 6 of 36 positive examples. Because study 16 is 0.57 positive, positive-class F_1 remains sensitive to class balance on that split; precision, coverage, and raw confusion counts are more informative for the abstaining-filter objective. Because studies 11/15 were inspected repeatedly during development, I use them as model-selection evidence rather than as the final benchmark. The study 16 evaluation should therefore be read as a one-shot untouched holdout evaluation of frozen finalists, with the caveat that it still has residual image overlap with studies 11/15.

5.3 Comparison Protocol

The experiments answer four main questions. First, does richer alignment context help at all? To answer that, I compare 39 suite/model combinations (CaptionAndImageOnly, PlusAlignScore, PlusAlignScoreAndReasoning crossed with 13 model families) on the development split. Second, which configurations remain worth trusting after robustness checks? For that, I narrow to a small finalist set and confirm the top 5 configurations on matched seeds 13, 17, and 23 under the final precision-first threshold policy. Third, do fine-tuned Qwen judges add value beyond strong off-the-shelf multimodal models? For that, I evaluate prompt-only GPT, Gemini, and Claude baselines on the same development split with the same validation-tuned threshold policy. Finally, do the frozen finalists transfer beyond development? To answer that, I reserve study 16 as a one-shot untouched holdout and evaluate the frozen systems there with their validation-selected thresholds fixed. The development-test tables below are model-selection evidence. The final finalist table reports the untouched holdout evaluation.

Prompt-only frontier baselines used the actual image plus caption, zero-shot and 4-shot prompting, temperature 0, and the provider model IDs `gpt-5.4`, `gemini-3-flash-preview`, and `claude-sonnet-4-6`. I include these frontier baselines specifically to test whether the fine-tuned Qwen judges add value over strong off-the-shelf multimodal models with no task-specific training. Appendix D gives the exact preserved prompt templates, and Appendix E adds extra study 16 examples beyond the main-paper table.

The finalists entering the last stage were: PlusAlignScore + Qwen3-8B (epochs 4), PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), and PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0).

5.4 Results

Development-time suite ablation.

Table 3 summarizes the broad matched-seed comparison on the development test that I used to shortlist models before the final deployment-oriented sweep. At that stage, PlusAlignScoreAndReasoning had the strongest mean development-test F_1 across all 13 families, suggesting that alignment-score reasoning was more useful than either the base fields or the raw alignment score alone.

Lightweight and prompt-only baselines.

Two simple baselines help calibrate how much of the final performance comes from lightweight signals rather than Qwen fine-tuning. The TF-IDF + logistic-regression baseline is clearly weaker, but directly thresholding the raw GPT-5.4 `image_alignment_score` is genuinely competitive. At the validation-tuned operating point, that raw alignment baseline reaches development-test precision

Suite	Added context	Mean test F_1	Best model on shared seed	Best F_1
CaptionAndImageOnly	Caption + description/image only	0.5300	Qwen3.5-9B	0.5641
PlusAlignScore	+ image alignment score	0.5025	Qwen3-8B-Base	0.5895
PlusAlignScoreAndReasoning	+ alignment-score reasoning	0.5553	Qwen3-4B-Instruct-2507	0.6087

Table 3: Development-time suite comparison on studies 11/15, where means are over all 13 model families per suite.

Final configuration	Mean precision	Mean recall	Mean F_1	Mean FPR	Mean coverage
PlusAlignScore + Qwen3-VL-8B-Instruct + dropout 0.1	0.700 $\pm$ 0.100	0.129 $\pm$ 0.079	0.213 $\pm$ 0.118	0.062 $\pm$ 0.013	0.100 $\pm$ 0.050
PlusAlignScoreAndReasoning + Qwen3.5-4B + dropout 0.1	0.686 $\pm$ 0.038	0.123 $\pm$ 0.035	0.206 $\pm$ 0.049	0.078 $\pm$ 0.036	0.103 $\pm$ 0.035
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking + dropout 0	0.702 $\pm$ 0.129	0.094 $\pm$ 0.044	0.163 $\pm$ 0.071	0.054 $\pm$ 0.036	0.077 $\pm$ 0.032
PlusAlignScore + Qwen3-8B + epochs 4	0.587 $\pm$ 0.138	0.088 $\pm$ 0.046	0.151 $\pm$ 0.072	0.078 $\pm$ 0.027	0.083 $\pm$ 0.032
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking + baseline	0.556 $\pm$ 0.096	0.076 $\pm$ 0.027	0.132 $\pm$ 0.040	0.085 $\pm$ 0.048	0.080 $\pm$ 0.035

Table 4: Study 16 holdout evaluation of the five frozen finalists under thresholds selected on validation during development. Each row is averaged over seeds 13, 17, and 23; $\pm$ denotes standard deviation across seeds. Only 2 of the 15 seed-specific holdout runs still satisfy the original 0.15 recall floor, which exposes calibration shift rather than a ranking-only change.

0.517, recall 0.600, F_1 0.556, and coverage 0.305. That confirms that the auxiliary teacher signal is informative on its own, even before a learned downstream filter is added.

I also evaluated prompt-only multimodal frontier baselines with GPT-5-family, Gemini Flash, and Claude Sonnet models using the actual image plus caption and the same validation-tuned precision-first threshold policy [12, 13, 14]. This comparison asks directly whether task-specific Qwen fine-tuning adds anything over strong non-fine-tuned frontier models. Among those baselines, Claude Sonnet 4.6 zero-shot was strongest overall with precision 0.430, recall 0.740, F_1 0.544, and coverage 0.453. GPT-5.4 was best in 4-shot mode at 0.382/0.840/0.525/0.579 precision/recall/ F_1 /coverage, while Gemini 3 Flash was best zero-shot at 0.344/0.880/0.494/0.674. These frontier prompting baselines are meaningful, but they still underperform both the raw alignment-score baseline and the best fine-tuned Qwen finalists on this humor-filtering objective, so the fine-tuned models do add value beyond prompt-only frontier judging here.

Frozen learned finalists on the untouched holdout.

However, the final deployment objective was not “maximize balanced F_1 ” but “show only likely-funny captions.” On the earlier development test, that precision-first criterion favored PlusAlignScoreAndReasoning + Qwen3.5-4B + dropout 0.1 for precision, while PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking remained the strongest balanced- F_1 model.

Table 4 asks the holdout-transfer question: how do those same frozen finalists perform on reserved study 16 when their validation-selected thresholds are left unchanged? All five learned models remain highly conservative. Among learned finalists, the strongest holdout result is the simpler VL configuration PlusAlignScore + Qwen3-VL-8B-Instruct, with the text-only PlusAlignScoreAndReasoning + Qwen3.5-4B model close behind. From a product perspective, this is still the behavior of a strict comedy gate: it throws away many jokes that the audience would have liked, but the captions it does pass through are funny more often than not.

Why the holdout table needs context.

Table 5 contextualizes why positive-class F_1 can be misleading on study 16. An always-funny classifier gets $F_1 = 0.726$ simply because the holdout is 57% positive, despite predicting every row as funny. On this holdout, directly thresholding the raw alignment score at the frozen validation threshold actually exceeds every learned finalist on positive-class F_1 by trading away precision for recall. The best learned model still improves precision over both the always-funny baseline and the raw alignment baseline, but it does so with much lower recall. The same table also shows that performance is materially stronger on overlap rows than on novel-image rows.

Threshold-free metrics help separate ranking quality from threshold transfer. Across the three frozen seeds, the best learned finalist reaches mean study 16 AUPRC 0.687 $\pm$ 0.027 and mean AUROC 0.616 $\pm$ 0.025, compared with 0.568 AUPRC and 0.518 AUROC for the raw alignment-score baseline. These are meaningful ranking gains, even though the fixed validation thresholds still

System	Slice	TP	FP	FN	TN	Precision	Recall	F_1
Always funny	Overall	57	43	0	0	0.570	1.000	0.726
Raw alignment score	Overall	11	7	46	36	0.611	0.193	0.293
Best learned model	Overall (pooled)	22	8	149	121	0.733	0.129	0.219
Best learned model	Overlap images	20	8	103	79	0.714	0.163	0.265
Best learned model	Novel images	2	0	46	42	1.000	0.042	0.080

Table 5: Study 16 contextual baselines and overlap split. The best learned model pools the three saved holdout runs of PlusAlignScore + Qwen3-VL-8B-Instruct + dropout 0.1, making the coarse rate changes on a 100-row holdout explicit.

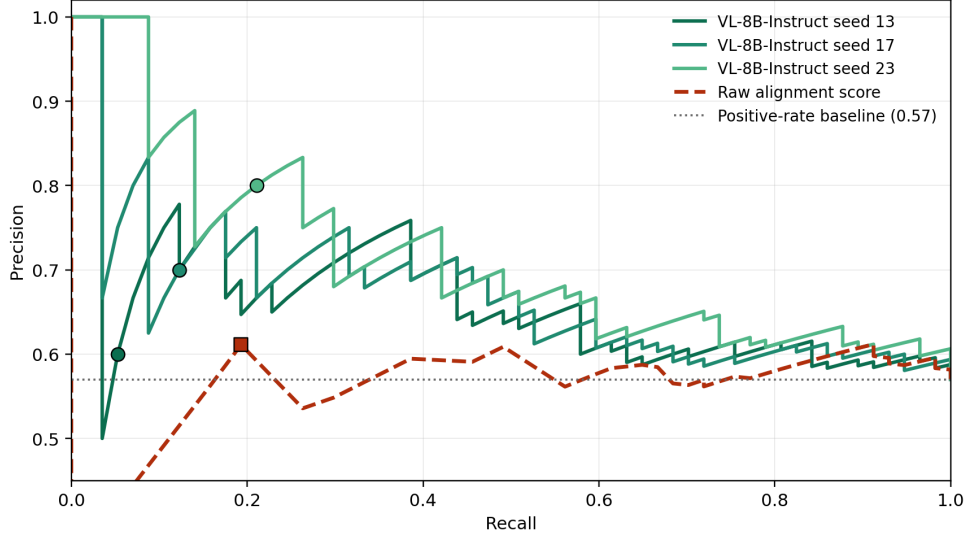


Figure 1: Study 16 precision-recall curves for the best learned finalist (PlusAlignScore + Qwen3-VL-8B-Instruct + dropout 0.1) and the raw alignment-score baseline. Circles mark transferred validation thresholds; the square marks the alignment baseline at threshold 0.89.

transfer poorly. Figure 1 shows the practical problem directly: the frozen validation thresholds land in a very high-precision, low-recall corner of otherwise better available precision-recall curves.

6 Analysis

Alignment-score reasoning helped more than the raw score alone.

In the broad matched-seed comparison, PlusAlignScoreAndReasoning achieved the highest mean development-test F_1 (0.5553), beating both CaptionAndImageOnly and PlusAlignScore. This suggests that the scalar alignment score alone is too compressed: the reasoning field gives the model more usable context about *why* a caption does or does not fit the image, and that extra explanation transfers better across model families than the score alone. Appendix D includes a full example of one of these image-caption alignment-score reasoning outputs so the reader can see the structure of the added context directly.

The richer suites also inject an external model prior.

Both suites expose frozen GPT-generated auxiliary judgments: GPT-4o mini supplies the image description, and GPT-5.4 supplies the alignment score and structured reasoning from the description-caption pair. The ablation therefore asks whether a downstream humor judge benefits from a stronger external side signal rather than whether the base Qwen backbone can infer the same structure on its own. Thresholding the raw GPT-5.4 alignment score already reaches development-test $F_1 = 0.556$, so the teacher feature is strong, yet the best fine-tuned precision-oriented finalist still improves precision by about 0.10 while using lower coverage.

The holdout exposed calibration fragility, not just seed noise.

The study 16 holdout results are much worse than the earlier development-test table: the strongest learned holdout model reaches only 0.700/0.129/0.213 precision/recall/ F_1 , far below the roughly

Case	Caption and short image cue	Votes	Text p_{funny}	VL p_{funny}	Reading
Text TP, VL FN	“classroom heater still on like we haven’t left february and i’m sweating”; contemplative Patrick reaction	42–28	0.349 (T)	0.407 (N)	The text model clears its lower threshold on a relatable campus complaint, while the VL model stays conservative despite a compatible reaction image.
VL TP, text FN	“daylight saving stole an hour during finals week and i carry on”; crying frog behind a smiling mask	41–25	0.269 (N)	0.500 (T)	The actual image makes the “holding it together” joke legible enough for VL to pass threshold.
Shared FN, novel image	“me watching low steps refill overnight at first real sun”; triumphant sunset silhouette	50–20	0.095 (N)	0.182 (N)	On a positive novel-image case, both models stay far below threshold, illustrating that the holdout failure is still dominated by under-calling positives.

Table 6: Illustrative study 16 cases for a strong text finalist and the strongest learned holdout finalist. T = predicted funny, N = predicted not funny.

0.59–0.60 development-test F_1 values of the leading finalists, with coverage near 0.10. Part of that drop is likely data-source shift rather than threshold transfer alone: the weekly studies were not content-matched, and study 16 mixed in more time-sensitive Sidechat jokes alongside a different balance of generic, in-group, and pop-culture captions. The raw alignment baseline sharpens that interpretation: it beats every learned finalist on holdout F_1 (0.293) while losing on AUPRC and AUROC. That is exactly the pattern I would expect if the learned models still rank examples somewhat better, but their transferred validation thresholds are too conservative for the holdout vote distribution. Figure 1 shows this directly: the learned curves sit above the raw alignment baseline, but the chosen operating points live in a low-recall corner.

The development-test winner did not survive unchanged on the holdout.

Before the holdout, the strongest confirmed balanced- F_1 model was the VL configuration PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking at 0.599 on studies 11/15. On study 16, that same configuration drops to mean $F_1 = 0.132$ with recall 0.076. Within this holdout comparison, the strongest learned finalist is instead the simpler VL configuration PlusAlignScore + Qwen3-VL-8B-Instruct at mean $F_1 = 0.213$, with the text-only PlusAlignScoreAndReasoning + Qwen3.5-4B model close behind at 0.206. Table 6 makes this concrete: the text model wins when the auxiliary reasoning field states the joke mechanism in a way that survives thresholding, whereas VL wins when the image itself carries the joke more directly.

Threshold tuning still mattered, but the tuned thresholds did not transfer.

All 15 final confirmation runs satisfied the validation recall floor, but only 2 of the 15 seed-specific holdout runs still satisfy recall ≥ 0.15 on study 16. With only 36 validation positives, the floor was already weak; the holdout results show that it also calibrates poorly under distribution shift. Threshold robustness across study rounds is therefore the clearest remaining weakness, especially because pooled performance is much better on overlap rows (0.265 F_1) than on novel-image rows (0.080 F_1).

Limitations.

Studies 11/15 were inspected repeatedly during iteration, so they remain development evidence rather than a pristine blind evaluation. Study 16 is cleaner but still shares 14 image IDs with the earlier development test, covering 70/100 holdout rows. The strongest PlusAlignScore/PlusAlignScoreAndReasoning results also depend on frozen GPT-generated metadata, and the shift from validation to study 16 shows that the current threshold policy is not yet robust across study rounds. The dataset was also designed week by week for experimentation rather than strict distributional consistency, so the mix of generic memes, Columbia-specific jokes, pop-culture references, intentionally weak prompts, and timely Sidechat jokes varies across studies.

7 Conclusion

This project shows that a humor filter can learn useful signal from captions, images, and frozen GPT-generated image-caption alignment fields. On the development split, the strongest systems used the alignment-score reasoning field and beat both simpler input setups and prompt-only frontier baselines. On the reserved holdout, the main problem was not that the accepted captions stopped being funny; it was that the filter became too conservative and rejected too many good jokes. The current system therefore looks most promising as a strict but usually reliable comedy gate, with better cross-study calibration as the clearest next step.

References

- [1] Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, and Robert Nowak. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning, 2024.
- [2] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [3] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.
- [4] Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 09–15 Jun 2019.
- [5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.
- [6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.
- [8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020.
- [9] Qwen Team. Qwen3.5-4b. <https://huggingface.co/Qwen/Qwen3.5-4B>, 2025. Model card, accessed 2026-03-11.
- [10] Qwen Team. Qwen3-vl-4b-thinking. <https://huggingface.co/Qwen/Qwen3-VL-4B-Thinking>, 2025. Model card, accessed 2026-03-11.
- [11] OpenAI. GPT-4o mini: Advancing cost-efficient intelligence, 2024. Official product announcement, accessed 2026-03-11.
- [12] OpenAI. Introducing GPT-5 for developers, 2025. Official developer announcement, accessed 2026-03-11.
- [13] Google. Models | gemini api, 2026. Official Gemini model overview, accessed 2026-03-11.
- [14] Anthropic. Models overview | claude api docs, 2026. Official Claude model overview, accessed 2026-03-11.
- [15] The Humor Project (Columbia University). The humor project. <https://thehumorproject.org>, 2026. Accessed 2026-03-11.
- [16] Humor Study. Humor study. <https://humorstudy.org>, 2026. Caption-voting platform, accessed 2026-03-12.

A Search Details

The learned-model search process had four stages that are omitted from the main paper to keep the body focused on the scientific comparisons rather than the development diary. I first ran the 39 base suite/model combinations, then reran broad matched-seed comparisons after changing checkpoint selection from validation loss to validation F_1 , then ran a 49-job finalist hyperparameter sweep, and finally confirmed the top 5 finalist configurations on matched seeds 13/17/23 under the final precision-first threshold policy. Across those stages, the learned-model portion of the study comprised 103 single-adapter LoRA training jobs on Modal, each using one NVIDIA L40S-backed container.

B Results Directory Contents

Full run-level results are available in the project `results` directory in Gradescope. That directory contains the complete saved outputs for the learned-model sweeps, frozen-finalist holdout evaluations, and prompt-only frontier baselines used in this project. This appendix summarizes the saved run artifacts and retains only the two most report-relevant run tables: the frozen-finalist study 16 holdout runs and the prompt-only frontier baselines.

B.1 CaptionAndImageOnly Development-Test Runs

The directory `results/test1` contains 129 saved markdown summaries for CaptionAndImageOnly development-test runs. Each file records the run name, checkpoint family, data sources, prompt fields, training hyperparameters, runtime, validation metrics, test metrics, and confusion-matrix counts.

B.2 PlusAlignScore Development-Test Runs

The directory `results/test8` contains 112 saved markdown summaries for PlusAlignScore development-test runs. These artifacts record the same training and evaluation fields as CaptionAndImageOnly, but with the frozen image-caption alignment score included in the prompt inputs.

B.3 PlusAlignScoreAndReasoning Development-Test Runs

The directory `results/test9` contains 136 saved markdown summaries for PlusAlignScoreAndReasoning development-test runs. These are the most information-rich learned-model artifacts and include the frozen image-caption alignment score reasoning field in addition to the score itself.

B.4 Frozen-Finalist Study 16 Holdout Runs

Table 7 lists all 15 seed-specific frozen-finalist study 16 evaluations.

Table 7: All frozen-finalist study 16 holdout runs.

Result artifact	Model	Study 16 F_1
test8-qwen3-8b-epochs4-seed-13	Qwen/Qwen3-8B	0.1270
test8-qwen3-8b-epochs4-seed-17	Qwen/Qwen3-8B	0.2319
test8-qwen3-8b-epochs4-seed-23	Qwen/Qwen3-8B	0.0938
test8-qwen3-vl-8b-instruct-dropout10-seed-13	Qwen/Qwen3-VL-8B-Instruct	0.0968
test8-qwen3-vl-8b-instruct-dropout10-seed-17	Qwen/Qwen3-VL-8B-Instruct	0.2090
test8-qwen3-vl-8b-instruct-dropout10-seed-23	Qwen/Qwen3-VL-8B-Instruct	0.3333
test9-qwen3-5-4b-dropout10-seed-13	Qwen/Qwen3.5-4B	0.2535
test9-qwen3-5-4b-dropout10-seed-17	Qwen/Qwen3.5-4B	0.2090
test9-qwen3-5-4b-dropout10-seed-23	Qwen/Qwen3.5-4B	0.1562
test9-qwen3-vl-4b-thinking-baseline-seed-13	Qwen/Qwen3-VL-4B-Thinking	0.0952
test9-qwen3-vl-4b-thinking-baseline-seed-17	Qwen/Qwen3-VL-4B-Thinking	0.1739
test9-qwen3-vl-4b-thinking-baseline-seed-23	Qwen/Qwen3-VL-4B-Thinking	0.1270
test9-qwen3-vl-8b-thinking-dropout0-seed-13	Qwen/Qwen3-VL-8B-Thinking	0.1515
test9-qwen3-vl-8b-thinking-dropout0-seed-17	Qwen/Qwen3-VL-8B-Thinking	0.2388
test9-qwen3-vl-8b-thinking-dropout0-seed-23	Qwen/Qwen3-VL-8B-Thinking	0.0984

B.5 Frontier Baseline Runs

Table 8 lists every prompt-only frontier setting evaluated on the development test. The tuned column uses the same validation-selected precision-first threshold policy as the learned models.

Table 8: All frontier baseline development-test runs.

Provider	Mode	Test F_1 @ 0.5	Tuned test F_1
Claude Sonnet 4.6	zero-shot	0.551	0.544
Claude Sonnet 4.6	4-shot	0.503	0.484
GPT-5.4	zero-shot	0.488	0.521
GPT-5.4	4-shot	0.498	0.525
Gemini 3 Flash	zero-shot	0.464	0.494
Gemini 3 Flash	4-shot	0.484	0.387

C Frozen-Finalist Seed Details

This appendix section records the exact validation-selected thresholds and seed-specific validation/holdout outcomes for the 15 frozen-finalist study 16 evaluations.

Table 9: Seed-specific frozen-finalist thresholds and metrics. Thresholds are selected on validation and then transferred unchanged to study 16.

Finalist seed	τ	Val P	Val R	Val F_1	Holdout P	Holdout R	Holdout F_1	Holdout cov.
PlusAlignScore + Qwen3-8B (epochs 4), s13	0.731	0.625	0.139	0.227	0.667	0.070	0.127	0.060
PlusAlignScore + Qwen3-8B (epochs 4), s17	0.500	0.385	0.139	0.204	0.667	0.140	0.232	0.120
PlusAlignScore + Qwen3-8B (epochs 4), s23	0.500	0.429	0.083	0.140	0.429	0.053	0.094	0.070
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s13	0.500	0.556	0.139	0.222	0.600	0.053	0.097	0.050
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s17	0.562	0.429	0.167	0.240	0.700	0.123	0.209	0.100
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s23	0.469	0.381	0.222	0.281	0.800	0.211	0.333	0.150
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s13	0.321	0.400	0.333	0.364	0.643	0.158	0.254	0.140
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s17	0.755	0.400	0.222	0.286	0.700	0.123	0.209	0.100
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s23	0.562	0.375	0.167	0.231	0.714	0.088	0.156	0.070
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s13	0.651	0.462	0.167	0.245	0.500	0.053	0.095	0.060
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s17	0.531	0.320	0.222	0.262	0.500	0.105	0.174	0.120
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s23	0.531	0.333	0.139	0.196	0.667	0.070	0.127	0.060
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s13	0.562	0.438	0.194	0.269	0.556	0.088	0.152	0.090
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s17	0.562	0.538	0.194	0.286	0.800	0.140	0.239	0.100
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s23	0.562	0.600	0.167	0.261	0.750	0.053	0.098	0.040

Table 10: Seed-specific study 16 confusion counts for the frozen finalists.

Finalist seed	TP	FP	FN	TN
PlusAlignScore + Qwen3-8B (epochs 4), s13	4	2	53	41
PlusAlignScore + Qwen3-8B (epochs 4), s17	8	4	49	39
PlusAlignScore + Qwen3-8B (epochs 4), s23	3	4	54	39
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s13	3	2	54	41
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s17	7	3	50	40
PlusAlignScore + Qwen3-VL-8B-Instruct (dropout 0.1), s23	12	3	45	40
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s13	9	5	48	38
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s17	7	3	50	40
PlusAlignScoreAndReasoning + Qwen3.5-4B (dropout 0.1), s23	5	2	52	41
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s13	3	3	54	40
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s17	6	6	51	37
PlusAlignScoreAndReasoning + Qwen3-VL-4B-Thinking (baseline), s23	4	2	53	41
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s13	5	4	52	39
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s17	8	2	49	41
PlusAlignScoreAndReasoning + Qwen3-VL-8B-Thinking (dropout 0), s23	3	1	54	42

D Prompt Templates

GPT-4o mini image descriptions.

The exact prompt used to create `simple_image_description` was:

```
Describe to me this picture in detail with the subject’s demeanor and the image itself.

In addition, try to see how the background, if one exists, relates to the general context of the image. For
example, if an image is from a movie scene and the background is relevant, you should include the
context, not just describe it and by all means try all your best to understand how the background and
the foreground relate, and especially if the background is a prominent building or anything worth
mentioning, specifically mention what it is.

The image will likely be a picture of a person or people.

Also if you recognize that the image is from a specific show or set or a specific cultural event make sure
to make note of that.

Specifically use the following scene description for context, which may include names of celebrities,
famous figures, television shows, movies, and/or popular memes:

*****
${celebrityRecognitionOutput}
*****

If the subject is not human or does not exist at all, anthropomorphize it.
Find humorous elements about the picture. Keep your response to 150 words.
In the case that this image is a place, or action described in depth to an extent that is relevant to the
context.for example if the image is of an iconic building, make sure to specifically mention it.
```

GPT-4o mini produced these descriptions once per image before downstream fine-tuning or split-specific evaluation, and those frozen outputs were reused for every later experiment.

Example GPT-4o mini image description output.

One representative output looked like this:

The image features Leonardo DiCaprio as Jay Gatsby from *The Great Gatsby*, exuding charm and confidence as he raises a glass in a celebratory toast. His demeanor is inviting and enigmatic, with a slight smile that suggests he's in the midst of a lavish party.

In the background, vibrant fireworks and blurred figures create a lively atmosphere, emphasizing the film's themes of opulence and the exuberance of the Jazz Age. The overall scene captures a moment of indulgence and celebration, reflecting Gatsby's extravagant lifestyle.

The juxtaposition of Gatsby's suave appearance against the festive backdrop underscores his allure and the fleeting nature of glamour. It's as if he's inviting the viewer to join in the revelry, while also hinting at the deeper complexities of his character. The image encapsulates the spirit of the era, where wealth and excitement were ever-present, yet often ephemeral.

GPT-5 alignment-score prompt.

The exact prompt used to generate `image_alignment_score` and `image_alignment_score_reasoning` was:

```
You are an expert meme analyst. Analyze how well a caption matches an image based on its description. Consider:
- Does the caption's emotional tone match the image's emotional affordance?
- Does the caption exploit the visual reaction shown in the image?
- Is there a mismatch between what the image shows and what the caption says?
- How well does the humor mechanism align with the image's reaction type?

Return ONLY valid JSON:
{"alignment_score": 0.0-1.0, "emotional_match": "high|medium|low",
"reaction_exploitation": "strong|moderate|weak", "mismatch_type": "description or none",
"humor_mechanism": "brief description", "improvement_suggestions": ["suggestion1", "suggestion2"]}

Image: {images.image_description}
Caption: {captions.content}

Analyze alignment (JSON only).
```

Example GPT-5 alignment-score output.

One representative JSON output looked like this:

```
{
  "alignment_score": 0.91,
  "emotional_match": "high",
  "reaction_exploitation": "strong",
  "mismatch_type": "none",
  "humor_mechanism": "personified shock and dismayed realization; the dog's upright, wide-eyed posture effectively conveys a sudden 'wait, what?' reaction that fits losing an hour before an exam",
  "improvement_suggestions": [
    "Make the caption slightly more conversational to match the dog's human-like stance, e.g. 'me realizing daylight saving just robbed me of an hour before my midterm'",
    "If desired, heighten the panic for even tighter alignment with the surprised expression, e.g. 'me realizing my midterm is after daylight saving stole my last hour of studying'"
  ]
}
```

Text-only Qwen prompt template.

The learned text-only models scored the two labels using this exact chat template, with suite-specific prompt fields filled from Table 1:

```
System:
You are a humor judge. Decide whether the caption is funny for the described image. Use only the provided fields. Return exactly one label: funny or not funny.

User:
Image description: {simple_image_description}
Caption: {caption}
[optional] Image alignment score: {image_alignment_score}
[optional] Image alignment score reasoning: {image_alignment_score_reasoning}
```

Vision-language Qwen prompt template.

The learned VL models received the downloaded image plus this exact text template:

```
System:
You are a humor judge. Determine whether the caption is funny for the image. Use only the provided image
and text fields. Return exactly one label: funny or not funny. Do not explain your answer.

User:
[image bytes]
Caption: {caption}
[optional] Image alignment score: {image_alignment_score}
[optional] Image alignment score reasoning: {image_alignment_score_reasoning}
Is this caption funny for the image? Return exactly one label.
```

Frontier-baseline prompt and settings.

The zero-shot and few-shot frontier baselines used this exact instruction string. This prompt reflects the available Columbia supervision used for those baselines and should not be read as the intended deployment scope of the overall humor judge:

```
You are evaluating whether a caption would be funny to a specific classroom audience. The audience consists
of Columbia students voting on AI-generated captions for user-uploaded images. Judge audience-
specific humor, not just whether the caption matches the image literally. The score should represent
the probability that this audience would give the caption a strict majority of upvotes. Return only a
JSON object matching the requested schema. Set predicted_label to 'funny' when funny_score >= 0.5 and '
not funny' otherwise. Keep short_rationale to one short sentence, at most 18 words, and avoid
mentioning hidden reasoning.
```

The required JSON schema was {"funny_score", "predicted_label", "short_rationale"}. Few-shot mode used 4 image-disjoint train exemplars chosen with seed 17, with 2 positive and 2 negative examples and no shared image IDs with validation or development test.

E Additional Qualitative Cases

Case	Caption	Votes	Alignment	Text / VL	Why it matters
Teacher TP, learned FN	"spilled coffee on backpack and now everything is over"	37–30	0.94	0.294 N / 0.321 N	The raw alignment baseline would pass this true positive at 0.89, but both learned filters abstain. This is a clean example of threshold conservatism rather than missing alignment.
Teacher FP, learned TN	"cafeteria mystery meat labeled vegan and my soul filed a complaint"	22–44	0.92	0.276 N / 0.438 N	The high teacher score is misleading on this audience. Here the learned filters correctly avoid a false positive that the raw alignment threshold would surface.
Text TP, VL FN	"me realizing daylight saving stole an hour before my midterm"	35–29	0.91	0.378 T / 0.469 N	The reasoning field's overconfidence cue is enough for the text model to pass threshold, while the VL model remains just below its operating point.
Borderline negative rejected by all	"abc7 helicopters show up while we're tanning and i'm trying not to laugh"	24–42	0.88	0.269 N / 0.202 N	This example sits just below the raw alignment threshold and is also rejected by both learned filters, showing how hard borderline cases can flip under the study 16 votes.

Table 11: Additional study 16 cases from seed 13. T = predicted funny, N = predicted not funny. These examples complement Table 6 by showing when the raw alignment baseline is better or worse than the learned filters.

F Sample Rated Image-Caption Pairs

Figures 2 and 3 show four screenshots in the exact presentation format used in the weekly studies. These are representative examples of the image-caption pairs that students rated. All four examples received a strict majority of positive votes and therefore count as gold-positive examples under the report's labeling rule.

suddenly remembering I left
my laundry in the dryer three
days ago



Example 1: 88 likes, 40 dislikes (gold funny).

Wait, that was on the syllabus?



Example 2: 101 likes, 35 dislikes (gold funny).

Figure 2: Representative rated image-caption pairs from the weekly Columbia humor studies.

me waiting for my registration
time slot to open like



Example 3: 87 likes, 42 dislikes (gold funny).

this is what getting called on
feels like



Example 4: 70 likes, 49 dislikes (gold funny).

Figure 3: Additional representative rated pairs showing the same screenshot format presented to participants.