Unreliable Narrator Framing as a Cost-effective Mechanism for Satire Explanation of Australian and American Satirical News

Kanay Gupta, David Y. Liu, Aditya Joshi
University of New South Wales, Australia
kanay.gupta@student.unsw.edu.au, david.liu2@unsw.edu.au, aditya.joshi@unsw.edu.au

Abstract

Satirical news is a unique form of misinformation: it is intended to be humorous by the original publication source but can be misinterpreted. Explanation of satirical news articles is challenging because recent and local context needs to be incorporated. This is a pilot study that explores explanation generation for satirical news articles from Australia and compares them alongside those from USA. We show that framing explanations with unreliable narrator theory on open-weight models can be a feasible, cost-effective alternative to frontier models. We find that models identify the satirical tropes more accurately on the American source than the Australian one. This work results in a novel dataset and an evaluation benchmark for satire explanation.

1 Introduction

Satirical news is a popular form of social media content which represents a unique form of misinformation. Unlike fake news, the intent of factually inaccurate satirical news is not deceptive. Satirical news functions as a means of delivering political and social commentary through humour and critique, but it is not always understood (Poulsen et al., 2023). Explaining satirical news and its real world commentary requires explaining what makes it satirical and the real-world phenomena it references. Previous work has shown that vision-language models struggle both to detect satire in images and to explain the source of its satirical effect (Nandy et al., 2024).

Betoota Advocate is an Australian news website that publishes satirical content, which typically references recent and geography-specific context which might be absent from a language model's training data. We collect 1000 articles from Betoota Advocate and The Onion, its American counterpart, and generate explanations of satirical news using four open-weight models and two frontier models.

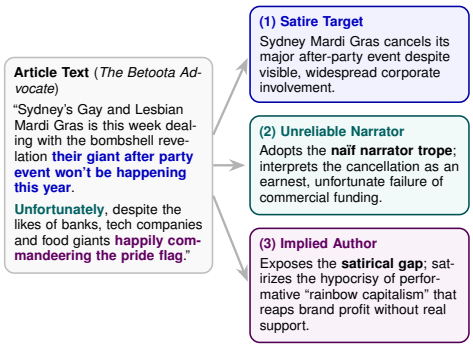

Figure 1: Unreliable narrator framing of an Australian satirical news article. The satire emerges from the unreliable narrator's surface account diverging from the implied author's critique of performative allyship.

We apply Booth (1961)'s unreliable narrator framing (Figure 1) through prompting used alongside internet search tool calling. We observe that the unreliable narrator prompt improves explanation quality in 16 of 17 comparisons with a baseline prompt, with the largest gains on the weakest models and no measurable advantage in the frontier models, suggesting that structured framing compensates for limited capability.

Ours is the first work to evaluate satire explanation on Australian satirical news articles, and to use the literary concept of the unreliable narrator (Booth, 1961) as a prompting framework for satirical news explanation generation (Figure 1). We make three observations. **(1)** Every model identifies the narrator trope more accurately on The Onion than on Betoota, a consistent difference between the two outlets. **(2)** The trope a model assigns reflects the model rather than the article, with each model applying its own preferred trope to both sources regardless of the text. **(3)** Performance degrades as more implicit interpretation is required, from describing the surface account, to inferring the implied author's belief, to explaining the gap between them.

2 Related Work

Work on satire has moved from binary classification toward explanation generation. Across that shift, the recurring difficulty is not recognising that a text is satirical but providing the world knowledge that explains why, and this becomes harder with cultural specificity. Golbeck et al. (2018) introduce a fake news versus satire dataset organised by the theme, but text-only classifiers trained on such data learnt the stylistic shortcuts rather than the content of the joke. Li et al. (2020) showed that combining the headline and image features with ViLBERT outperforms unimodal baselines, though the model still cannot articulate what makes an article satirical. Nandy et al. (2024) reframed the problem as comprehension with YesBut, reporting that vision-language models score below 60% accuracy on explanation generation for the satirical memes. Figurative language has also followed the same shift from labelling to explanation, as seen in FLUTE (Chakrabarty et al., 2022) and sarcasm explanation in multi-party dialogue (Kumar et al., 2022). Abdalla et al. (2024) finds that chain-of-thought prompting recovers much of the performance lost on Arabic satirical news. No prior work evaluates satire explanation on Australian satirical news articles, and no prior work uses unreliable narrator (Booth, 1961) as a prompting framework for explanation generation in the news domain.

3 Methodology

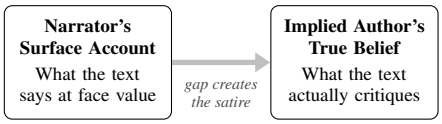

Figure 2: Satirical gap model based on Booth (1961).

Given a satirical news article, the model is tasked with generating: (1) description of the **satire target** which could be a real-world event, person or phenomenon, and (2) an explanation of what is being satirised and how (Figure 1).

We hypothesise that satirical news uses the separation between the narrator of the text from its implied author, which is the set of values the article is actually trying to imply, in order to create the satire in the article (Figure 2). We frame this implicit narrator in satire as an **unreliable narrator** (Booth, 1961) which follows tropes such as the Naif who is naively blind to the obvious wrongness of what they describe and the Clown who reframes conflict or failure as normal or even positive (Brei et al., 2025) (Appendix A). The model is then asked to state (1) which trope applies and why, (2) what the narrator presents on the surface, (3) what the implied author actually means, and (4) how the gap between the narrator and implied author creates the satirical effect. The detailed prompts are in Appendix B.

Australian satire depends on knowledge such as local political figures, recent events or slang, which a model may not hold. To address this issue, the model is given access to perform web searches before answering (Figure 3). Details about the implementation are given in the Appendix C.

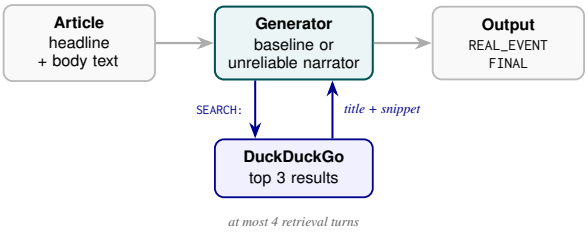

Figure 3: Both prompting conditions share the same retrieval scaffold. The model may emit a `SEARCH:` directive and receive the top three DuckDuckGo results before producing its `REAL_EVENT` and `FINAL` blocks.

Additionally, we test a pre-processing step described in Appendix D but find that it degrades explanation of the satirical target in the majority of cases so we do not use it in the main comparison.

4 Experiment Setup

We use 500 articles from The Betoota Advocate (`www.betootaadvocate.com`) and 500 from The Onion (`www.theonion.com`). Each record in the dataset includes a headline and body text.

We evaluate four open-weight instruction-tuned models: Qwen2-VL-7B-Instruct, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. The Qwen2.5-7B model isolates the effect of removing vision within a single family while Llama and Mistral test whether the structured prompting has an effect across different families. As a comparison we additionally run GPT-5.6 Luna and GPT-5.6 Sol over 100 Betoota articles.

We score the generations with Gemini 3 Flash using a rubric (Pombal et al., 2025) in which the judge first writes feedback for the task and then assigns a score. All conditions are scored on satire target (1-5, accuracy of the target, *i.e.*, the identi-

fied real-world event, person or phenomenon being satirised) and explanation (1-5, accuracy and completeness of the satire mechanism) (Appendix F). UN outputs are additionally scored on four dimensions: trope accuracy (0/1, whether the judge finds the assigned trope well justified by the article), surface account, implied author and mechanism (each 1-5). We compare all judge scores with human ratings in Section 5.1.

5 Results

The unreliable narrator (UN) prompt improves explanation quality for every generator on both datasets (Table 1). Across all the comparisons between the two datasets, four models and both sentence classification settings (Appendix D), 16 out of 17 have a positive change when using the UN framework. Mistral on Betoota with sentence classification is the single negative outcome. Gains are largest where the baseline performance is the weakest.

Satire target identification also shows improvement as 13 of the 17 outcomes favour the UN prompt. Roughly half of the outputs score a 1 on this dimension. Upon closer analysis of individual outputs, we found that the model frequently restates the article's invented characters and events as though they were literal fact rather than identifying the real world phenomenon the article references (see Appendix G). Generations without a well-formed `FINAL` block are excluded. These are almost entirely from Qwen2-VL-7B, and keeping them or scoring them as 1 does not change the direction of any explanation comparison.

Table 2 reports the four unreliable narrator dimensions which apply only to the UN condition. Trope identification is the weakest component which ranges from 0.25 to 0.74 for the open-weight models while the surface account is stronger, ranging from 3.37 to 4.53. We see performance degrade where additional inference is required beyond the literal content from the article, such as from surface account to implied author, then to satirical mechanism.

Trope identification is where the sources diverge the most. Every model identifies the narrator trope more accurately on The Onion than on Betoota, by 0.08 to 0.28 (Table 2). The distributions models assign are consistent within a model but inconsistent across models. Llama-3.1-8B and Qwen2.5-7B label the majority of articles Naif in both sources,

		BETOOTA (AU)		THE ONION (US)	
Model	Prompt	*target*	*expl.*	*target*	*expl.*
Qwen2-VL-7B	BASE	**2.00**	2.00	**2.01**	1.86
	UN	1.93	**2.46**	1.96	**2.38**
Mistral-7B	BASE	1.76	2.75	2.09	2.45
	UN	**1.97**	**3.02**	**2.14**	**3.02**
Qwen2.5-7B	BASE	2.03	2.31	**2.28**	2.19
	UN	**2.14**	**3.06**	**2.28**	**3.09**
Llama-3.1-8B	BASE	1.68	2.83	2.01	2.91
	UN	**1.97**	**3.57**	**2.33**	**3.80**
GPT-5.6 Luna*	BASE	4.08	4.91	–	–
	UN	**4.14**	**5.00**	–	–
GPT-5.6 Sol*	BASE	4.48	4.98	–	–
	UN	**4.54**	**5.00**	–	–

Table 1: Mean judge scores (1–5) under the baseline (BASE) and unreliable-narrator (UN) prompts. Open-weight models were run with $n \leq 500$, generations that did not produce a well-formed "FINAL" block are excluded. *Frontier models were run on Betoota only with $n = 100$.

while Mistral-7B labels the majority Madman in both (Appendix A). Since all three receive identical inputs, this suggests a model-specific preference rather than a feature of the text, which is consistent with the low trope accuracies in Table 2. The higher trope accuracy on The Onion may reflect greater model familiarity with American satire, although publication style and topic mix may also play a role. The gap between the frontier and open-weight models suggests that the bottleneck lies in recovering the real-world reference rather than in the framing or evaluation of the task.

5.1 Human Evaluation

To validate the LLM as a judge, we sampled 50 Llama-3.1-8B responses on Betoota, stratified across both prompts and all five judge score bands. Three annotators scored them against the rubric in Table 9 with the same information available to the judge. Table 3 reports agreement among the annotators and between their aggregated scores and the judge. Agreement is good for satire target ($\text{ICC}(2, k) = 0.78$) and moderate for mechanism and implied author (0.52 and 0.49), but poor for explanation (0.36), trope (0.22) and surface account (-0.11). The aggregated human score correlates significantly with the judge on five of the six dimensions ($\rho = 0.44$ to 0.76), with the exception being surface account. The judge is stricter than the annotators on every dimension which is consistent between both prompts. We therefore rest our main

Model	*trope*	*surface*	*author*	*mech.*
Betoota Advocate (AU)				
Qwen2-VL-7B	0.29	3.37	2.54	2.27
Mistral-7B	0.25	3.94	3.55	2.84
Qwen2.5-7B	0.35	3.80	3.16	2.75
Llama-3.1-8B	**0.46**	**3.88**	**3.87**	**3.25**
GPT-5.6 Luna	0.95	4.96	5.00	5.00
GPT-5.6 Sol	0.98	5.00	4.94	5.00
The Onion (US)				
Qwen2-VL-7B	0.51	3.68	2.49	2.38
Mistral-7B	0.33	4.07	3.45	2.74
Qwen2.5-7B	0.54	4.36	3.06	3.10
Llama-3.1-8B	**0.74**	**4.53**	**3.83**	**3.77**

Table 2: Mean scores on the four unreliable-narrator (UN) dimensions: *trope* accuracy (0–1), *surface* accuracy, implied *author* belief, and satirical *mech*anism (each 1–5). Bold marks the best open-weight model per column. Frontier models were run on Betoota only, n=100.

claims on satire target and explanation and treat surface account and trope as indicative.

Dimension	n	human	judge	ICC$(2,k)$	ρ
satire target	50	2.70	1.98	0.78	0.68*
explanation	48	3.24	3.00	0.36	0.68*
trope	24	0.70	0.44	0.22	0.76*
surface	24	4.01	3.64	-0.11	0.30
implied author	23	3.49	3.36	0.49	0.62*
mechanism	23	3.40	2.84	0.52	0.44*

Table 3: Human evaluation. *human* is the mean of three annotators; ICC$(2,k)$ is the two-way random effects intraclass correlation for that mean; ρ is Spearman correlation between the aggregated human score and the judge, with * marking $p < 0.05$.

6 Discussion

The unreliable narrator prompt gives small models a structure to fill. Under the baseline prompt they tend to state that satire is present and name its target without explaining how the article produces it, whereas the UN prompt explains the satire by naming the trope, the surface account, the implied author's belief and the gap between them (Table 4). A control prompt with the same four steps but no narrator terminology gives similar or larger gains (Appendix E), so this structure is a key ingredient of the framing. Since satirical articles are published continuously and at volume, a small open-weight model with a structured prompt offers a cost-effective way to produce explanations at scale.

Both frontier models more than double the satire target score under identical prompts, scaffolding and judge, showing that the rubric can award high scores and that the low ceiling reflects the capability of open-weight models. The UN framing gives no clear improvement at this scale given the lack of headroom, suggesting that structured framing compensates for part of the limited capability of small models on explanation, **which makes it a cost-effective alternative to frontier models**.

Input Headline: "Girlfriend Takes A Pause From Scrolling To Declare She's Going To Read 20 Books This Year"

Baseline Explanation (2 / 5). The satire is about the superficial nature of New Year's resolutions and the tendency to prioritize social media over personal growth.

UN Explanation (5 / 5). ...The surface account presents Bianca Sutton as a woman who has decided to break away from her TikTok addiction and read 20 books this year. However, the implied author, who is likely a satirist, actually sees this as a continuation of her usual behaviour, where she spends most of her time on social media. The gap between the narrator's presentation of a serious change in behaviour and the implied author's recognition of it as a superficial and ultimately unchanged pattern creates the satirical effect.

Table 4: An example where a baseline prompt fails to produce a response that explains the satire.

7 Conclusion

We evaluated whether formulating satirical news explanation through the framework of an unreliable narrator is beneficial. We compared a structured unreliable narrator prompt against a baseline explanation on 500 Australian and 500 American satirical articles.

The unreliable narrator prompt improved explanation quality in 16 of 17 comparison combinations, where the gains were largest for the weakest models. Trope identification was consistently more accurate on the American source than the Australian one for every model tested, which may reflect greater familiarity with American satire. At the frontier scale, the advantage of the unreliable narrator framing disappeared, but the satire target scores more than doubled under identical conditions showing that the grounding was limited by model capability rather than the evaluation rubric. Future work should target the failures in grounding directly through verification during retrieval and should test whether framing transfers to other culturally specific satirical publications and to multimodal input.

Limitations & Ethical Considerations

What drives the gain. The baseline and UN prompts differ in task decomposition, instruction specificity and terminology, not only in theoretical framing. In follow-up experiments, a structured prompt with the same four steps but without narrator terminology gave larger explanation gains than UN in 7 of 8 cells (Appendix E). The benefit therefore appears to come largely from decomposing the task. The narrator framing remains useful for its interpretable intermediate outputs (trope, surface account, implied author), which show where explanations fail.

Evaluation. All scores come from a single proprietary judge (Gemini 3 Flash), and we do not test whether the results hold under other judges. The judge sees the article and the answer but no verified reference material, so it cannot fully check real-world claims about recent or local events. Structured answers are longer and use the requested vocabulary, which the judge may reward independently of understanding. Human validation covers 50 responses from one model on one source. Agreement is weak for explanation and trope and negative for surface account. Trope accuracy has no gold label and is the judge's own assessment.

Retrieval. Both prompts share the search scaffold, but different prompts can trigger different searches, so the conditions may not see the same evidence. The frontier models mostly answered without searching. A no-retrieval ablation would separate the framing effect from retrieval.

Publications and culture. We use one publication per country, so differences between Betoota and The Onion cannot be separated into cultural effects and publication effects such as style, political leaning, topic mix or difficulty. Several publications per country (e.g., The Babylon Bee alongside The Onion) are needed for cultural claims.

Ethics. All news articles are publicly available. All annotators are authors of the work.

References

Omar W. Abdalla, Aditya Joshi, Rahat Masood, and Salil S. Kanhere. 2024. Comparison of multilingual and bilingual models for satirical news detection of Arabic and English. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 173–178, Canberra, Australia. Association for Computational Linguistics.

Wayne Clayton Booth. 1961. *The rhetoric of fiction*, volume 21983. London.

Anneliese Brei, Katharine Henry, Abhisheik Sharma, Shashank Srivastava, and Snigdha Chaturvedi. 2025. Classifying unreliable narrators with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20766–20791, Vienna, Austria. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jennifer Golbeck, Jennine Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Matthew Mauriello, Elo Lewis, George Mirano, William IV, Sean Mussenden, Tammie Nelson, Sean Mcwillie, Akshat Pant, and Paul Cheakalos. 2018. Fake news vs satire: A dataset and analysis. pages 17–21.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. ***YesBut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895, Miami, Florida, USA. Association for Computational Linguistics.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre Martins. 2025. M-prometheus: A suite of open multilingual LLM judges. In *Second Conference on Language Modeling*.

Shannon H Poulsen, Robert M Bond, and R Kelly Garrett. 2023. Comparing beliefs in falsehoods based on satiric and non-satiric news. *PloS one*, 18(1):e0278639.

William Edward Riggan Jr. 1978. *PICAROS, MADMEN, NAIFS, AND CLOWNS: THE UNRELIABLE FIRST-PERSON NARRATOR.* Indiana University.

Appendix

A Unreliable Narrator Tropes

We take the unreliable narrator tropes described by Brei et al. (2025) and Riggan Jr (1978):

- **Naif**: naively blind to the obvious wrongness of what they describe. The narrator reports a troubling situation in an earnest, matter-of-fact register, and the satire comes from the reader seeing what the narrator does not.
- **Clown**: reframes conflict or failure as normal or even positive. The narrator acknowledges that something has gone wrong but treats it as unremarkable or as a cause for optimism.
- **Picaro**: treats a cunning or self-serving act as completely reasonable. The narrator presents self-interest, evasion or exploitation as ordinary and justified conduct.
- **Madman**: reports with excessive, disproportionate emotional investment. The narrator's intensity is wildly out of scale with the triviality of the subject.

Model	Source	Naif	Clown	Madman	Picaro
Llama-3.1-8B	AU	354	65	47	34
	US	411	39	37	3
Qwen2.5-7B	AU	307	82	77	31
	US	347	56	76	20
Mistral-7B	AU	35	68	295	98
	US	44	92	343	21

Table 5: Narrator tropes assigned under the unreliable narrator prompt, $n = 500$ per source. Each model applies a consistent preference across both sources, but the preferences differ between models.

B Prompts

We compare two methods, both of which share the same basic scaffold and output, described as follows: (A) **Baseline:** The prompt says that the model is an expert at analysing satire in news articles and asks for "a straightforward explanation of what the article is satirising." (B) **Unreliable Narrator (UN):** The prompt says the model is an expert in narratological analysis and supplies a background paragraph. The paragraph defines unreliable narrator and the four tropes.

The two methods compared in this work are matched in that they must output a "REAL_EVENT:" block followed by a "FINAL:" block. The specification for the "REAL_EVENT:" is identical between the baseline and the unreliable narrator prompting and they only differ in the specification for the "FINAL:" block. The full prompts are as follows:

Baseline prompt.

```
You are an expert at analyzing satire in news
articles. You have access to a web search tool
to look up current events, politicians,
celebrities, or slang required to understand
the joke.

STRICT RULES:
1. If you need context, output EXACTLY:
   'SEARCH: [short keyword query]'. Use short
   keywords, NOT full sentences.
2. If you output SEARCH, STOP WRITING
   immediately. Do NOT add anything else. I will
   reply with the results.
3. Once you have enough context, output BOTH of
   the following - in this order, in the SAME
   turn, with no other text:
   REAL_EVENT: [A factual 1-3 sentence
   description of the real-world event, person,
   or phenomenon the article is referencing. Be
   specific - name the actual people, dates, or
   events involved.]
   FINAL: [A straightforward explanation of what
   the article is satirising.]
4. NEVER output SEARCH in the same turn as
   REAL_EVENT or FINAL.
```

Unreliable narrator prompt.

```
You are an expert at analyzing satire using
narratological theory. You have access to a web
search tool to look up current events,
politicians, celebrities, or slang required to
understand the joke.

BACKGROUND: Satirical news sites write as an
'unreliable narrator' - a journalist who
presents absurd or exaggerated events with a
straight face. The narrator's surface account
diverges from the implied author's true intent,
and this gap IS the satire. The narrator
typically fits one of these tropes:
  - Naif: naively blind to the obvious wrongness
    of what they describe
  - Clown: reframes conflict or failure as normal
    or even positive
  - Picaro: treats a cunning or self-serving act
    as completely reasonable
  - Madman: reports with excessive,
    disproportionate emotional investment

STRICT RULES:
1. If you need context, output EXACTLY:
   'SEARCH: [short keyword query]'. Use short
   keywords, NOT full sentences.
2. If you output SEARCH, STOP WRITING
   immediately. Do NOT add anything else. I will
   reply with the results.
3. Once you have enough context, output BOTH of
```

```
   the following - in this order, in the SAME
   turn, with no other text:
   REAL_EVENT: [A factual 1-3 sentence
   description of the real-world event, person,
   or phenomenon the article is referencing. Be
   specific - name the actual people, dates, or
   events involved.]
   FINAL: [Analyse the satire through the
   unreliable narrator lens: (1) identify which
   narrator trope applies and why, (2) state what
   the narrator presents on the surface, (3)
   state what the implied author actually
   believes, and (4) explain how the gap between
   narrator and implied author creates the
   satirical effect.]
4. NEVER output SEARCH in the same turn as
   REAL_EVENT or FINAL.
```

The user turn supplies the article headline and body text. The sentence classification condition additionally prepends a note explaining the `REAL` and `EXAGGERATED` labels applied to each body sentence.

C Internet Search

The model is given access to perform web searches. The model outputs "SEARCH:" along with a keyword, and the system returns the top three DuckDuckGo results with a title and a snippet, and then model continues. The model is able to do a total of 4 search loops before outputting the "REAL_EVENT" and "FINAL" blocks. We refer to the "REAL_EVENT" block as the satire target, as it is not always an event but the block name is unchanged from our earlier runs.

D Preprocessing

Qwen2-VL-7B-Instruct is prompted as the classifier to generate a JSON array of sentence and label pairs which was applied identically to both baseline and unreliable narrator prompting. The tagged body is substituted for the raw body in the generation prompt, along with a note explaining the labels in order to test whether the pre-processing improved the results.

The pre-processing raises the explanation scores under the baseline prompt in 5 of 8 comparisons (mean + 0.13 with the largest being Mistral-7B on Betoota at +0.38) but has essentially no effect under the unreliable narrator prompt (mean -0.02). Satire target scores, however, degrade consistently as they fall in 14 of 16 comparisons (mean -0.235) (Table 8). Labelling a sentence "EXAGGERATED" appears to discourage the model from using the information for the satire target even when a real event is referenced in the exaggerated sentence. Therefore, we do not use the sentence classifier pre-processing in the main comparison.

E Structured Control Prompt

To test whether the UN gain comes from the narrator framing or from the decomposition, we ran a structured control prompt (STRUCT) on all four open-weight models and both sources (no sentence classification). It uses the baseline system prompt and search rules, and replaces only the `FINAL` specification with the UN four-step structure, with the narrator background, tropes and narrator/implied-author vocabulary removed:

```
FINAL: [Analyse the satire step by step:
(1) identify the main satirical technique the
article uses and why, (2) state what the
article presents on the surface, (3) state
what the writer actually believes, and (4)
explain how the gap between what is presented
and what the writer believes creates the
satirical effect.]
```

Table 6 compares it with the baseline and UN prompts. STRUCT improves explanation over the baseline in all 8 cells and scores higher than UN in 7 of 8, the exception being Llama-3.1-8B on The Onion. On satire target the three prompts are close and the differences are mixed. STRUCT answers are the longest of the three, so they carry the same length confound as UN. The result indicates that the narrator terminology is not what drives the explanation gain, but it does not show that STRUCT reflects better understanding.

		target			*expl.*		
	Model	BASE	UN	STR	BASE	UN	STR
Betoota	Qwen2-VL	2.00	1.93	2.07	2.00	2.46	2.75
	Mistral	1.76	1.97	1.79	2.75	3.02	3.60
	Qwen2.5	2.03	2.14	2.19	2.31	3.06	3.50
	Llama	1.68	1.97	1.85	2.83	3.57	3.72
Onion	Qwen2-VL	2.01	1.96	2.13	1.86	2.38	2.48
	Mistral	2.09	2.14	1.98	2.45	3.02	3.32
	Qwen2.5	2.28	2.28	2.32	2.19	3.09	3.27
	Llama	2.01	2.33	2.13	2.91	3.80	3.60

Table 6: Mean judge scores (1–5) for the baseline (BASE), unreliable narrator (UN) and structured control (STR) prompts, without sentence classification. Malformed generations excluded as in Table 1.

F LLM as Judge

Each dimension is scored in a separate call to the LLM, so the judge does not see more than one criterion at a time. The judge is run at temperature 0 and receives only the headline, body text and the explanation to be evaluated. We separately score satire

target and explanation (Table 9), since a model can provide a confident explanation built on a hallucinated target. UN conditions are additionally scored on trope_accuracy (0-1), surface_account, implied_author and mechanism which are each (1-5).

G Qualitative Analysis

The scoring rubric is in Table 9. Tables 10, 11, 7 and 12 show four qualitative examples.

Table 10 shows the pattern where under the baseline prompt the model names the target and stops, while the unreliable narrator prompt produces the trope, the surface account, the implied author's belief and the gap between them from the same knowledge.

Table 11 shows a second comparison in which the unreliable narrator response raises the explanation score from 2 to 5, but its satire target score falls from 3 to 1 as the model constructs a biography for Bianca Sutton, including a follower count and an employment history, for a person the article invented.

Table 7 shows the same issue in a single response. The model produces a good satire explanation and scores 4, but accepts the article's premise that Canada is a minnow nation, meaning an underdog candidate, and reports a real athlete and a real event as though the article were straightforwardly describing them leading to scoring 1 on satire target.

Table 12 shows the limit of the framing as both prompts restate Billy Crenshaw and his ice cream carton as the satire target, missing the phenomenon the article is mocking leading to a score of 1. The unreliable narrator prompt still produces a structured explanation, but it applies that structure to the fiction which shows that the framing organises the explanation without improving the grounding that it relies on. Roughly half of all satire target scores are like this, which shows that the retrieval scaffold should be a target for future work.

Input Headline: "Struggling Winter Olympics Minnow Nation Surprises With First Gold Medal"
Satire target (1 / 5). Mikaël Kingsbury, a Canadian freestyle skier, won the gold medal in men's moguls at the 2026 Winter Olympics in Milano Cortina, Italy.
Explanation (4 / 5). The Betoota Advocate employs the naif narrator trope to present a heartwarming story of a struggling nation, Canada, finally achieving a rare Winter Olympics victory. The implied author, however, recognises the true significance of Kingsbury's achievement, as he is one of the most decorated athletes in the history of moguls skiing. The gap between the narrator's portrayal of a fairy tale underdog story and the implied author's recognition of Kingsbury's true talent and legacy creates a satirical effect.

Table 7: An unreliable narrator output in which the explanation is coherent and scores 4, while the satire target scores 1. The model names a real athlete and a real event, but accepts the article's premise that Canada is a minnow nation rather than recognising it as the joke.

		target		*expl.*	
Model	Prompt	raw	+SC	raw	+SC
Betoota Advocate (AU)					
Qwen2-VL-7B	BASE	**2.00**	1.64	2.00	**2.04**
	UN	**1.93**	1.74	2.46	**2.48**
Mistral-7B	BASE	**1.76**	1.55	2.75	**3.13**
	UN	**1.97**	1.61	**3.02**	2.98
Qwen2.5-7B	BASE	**2.03**	1.93	2.31	2.31
	UN	**2.14**	2.10	3.06	**3.07**
Llama-3.1-8B	BASE	1.68	**1.69**	2.83	**3.04**
	UN	**1.97**	1.85	**3.57**	3.53
The Onion (US)					
Qwen2-VL-7B	BASE	**2.01**	1.64	**1.86**	1.81
	UN	**1.96**	1.79	**2.38**	2.34
Mistral-7B	BASE	**2.09**	1.54	2.45	**2.68**
	UN	**2.14**	1.63	3.02	3.02
Qwen2.5-7B	BASE	**2.28**	1.96	2.19	**2.30**
	UN	**2.28**	2.11	3.09	**3.10**
Llama-3.1-8B	BASE	**2.01**	**2.01**	2.91	**2.99**
	UN	**2.33**	2.10	**3.80**	3.74

Table 8: Effect of the sentence-classification pre-pass (+SC) against raw article text. Bold marks the better of each pair.

Score	Satire Target Score (`satire_target`)	Explanation Score (`satire_explanation`)
1	No real satire target identified, or entirely wrong / hallucinated.	Irrelevant, hallucinated, fails to explain, or merely restates the headline.
2	Correct general topic, but wrong specific people, dates, or details.	Identifies the target, but fails to explain how or why it is satirical.
3	Correct target with minor factual inaccuracies (e.g., wrong date, right person).	Explains the mechanism, but misses important cultural or political nuance.
4	Correct target with all key people and context; trivial differences only.	Explains target, mechanism, and most cultural context with minor gaps.
5	Fully accurate and specific: correct target (including, as applicable, event, people, dates, and AU context.)	Complete explanation: target, satirical device, and cultural critique.

Table 9: Evaluation rubric for satirical article comprehension across satire target grounding and satirical mechanism explanation.

Input Headline: "Corporations That Plaster Themselves In Rainbows Every Year Not Keen To Bail Out Mardi Gras Sadly"

Baseline Prompt **Score: 2 / 5**	**Unreliable Narrator** **Score: 5 / 5**
The satire is satirizing the hypocrisy of large corporations, such as banks, tech companies, and food giants, who use the pride flag for branding purposes but are unwilling to support the Mardi Gras event financially. The article suggests that these corporations are more interested in profit than in genuine support for the LGBTQ+ community.	The Betoota Advocate employs the **naif narrator trope** to present the situation as a **simple failure of corporate sponsorship**, while the **implied author** is likely satirizing the hypocrisy of corporations using rainbow-themed branding for profit while not actually supporting the community. **The gap** between the narrator's surface account and the implied author's critique of performative rainbow capitalism creates the satirical effect.
Names the target, then stops without explaining the satirical device	*Names all four parts: trope, surface, author, gap*

Table 10: Comparison of qualitative outputs and evaluation scores between the baseline prompt and the unreliable narrator framework.

Input Headline: "Girlfriend Takes A Pause From Scrolling To Declare She's Going To Read 20 Books This Year"

Baseline Prompt **Target: 3 / 5 Explanation: 2 / 5**	**Unreliable Narrator** **Target: 1 / 5 Explanation: 5 / 5**
Satire target. The article is a satire about a woman named Bianka Sutton who claims she will read 20 books this year, but instead of actually reading, she continues to scroll through her TikTok feed.	*Satire target.* Bianca Sutton is a social media influencer with over 38K followers on Instagram, where she posts content about her personal life, including her relationship with her boyfriend, Peter Evans. She is also a professional in the recruitment industry, as indicated by her LinkedIn profile.
Explanation. The satire is about the superficial nature of New Year's resolutions and the tendency to prioritize social media over personal growth.	*Explanation.* The Betoota Advocate employs the **naif narrator trope** in this article. The **surface account** presents Bianca Sutton as a woman who has decided to break away from her TikTok addiction and read 20 books this year. However, the **implied author**, who is likely a satirist, actually sees this as a continuation of her usual behaviour, where she spends most of her time on social media. **The gap** between the narrator's presentation of a serious change in behaviour and the implied author's recognition of it as a superficial and ultimately unchanged pattern creates the satirical effect.
Names the target but does not explain the device; summarises the article's plot rather than the phenomenon.	*Names all four parts, but invents a real-world identity for a fictional character.*

Table 11: A second qualitative comparison. The unreliable narrator prompt raises explanation from 2 to 5, but satire target falls from 3 to 1 because the model fabricates biographical detail for a character the article invented.

Input Headline: "Disappointed Man Reaches Bottom Of Ice Cream Carton Right When He Was Hitting His Stride" (The Onion)

Baseline Prompt	**Unreliable Narrator**
Satire target. Billy Crenshaw, a local man in Clemson, SC, is disappointed after reaching the bottom of an ice cream carton just as he was hitting his stride.	*Satire target.* A man, Billy Crenshaw, is disappointed after finishing an ice cream carton because he had just started feeling satisfied with the dessert.
Explanation. The article is satirizing the idea that people can become so engrossed in eating ice cream that they become disappointed when they finish the carton.	*Explanation.* The satire uses the **naif narrator trope**, presenting Billy Crenshaw's disappointment as a serious and relatable experience. The **implied author**, however, is likely to see the situation as humorous and exaggerated. **The gap** between the narrator's surface account of Billy's disappointment and the implied author's recognition of the absurdity of his reaction creates the satirical effect.

Judge feedback on the satire target: "Restates the fictional character and his ice cream carton as if it were literal fact, instead of identifying the real satirical target: sports-metaphor language applied to overeating."

Table 12: Both prompts restate the article's invented character as the satire target and score 1 on that dimension. The unreliable narrator prompt still produces a structured explanation, but the structure is applied to the fiction rather than to the phenomenon the article is mocking.