Drivel-ology: Challenging LLMs with Interpreting Nonsense with Depth

Yang Wang[1], Chenghao Xiao[2], Chia-Yi Hsiao[2], Zi Yan Chang[3], Chi-Li Chen[3], Tyler Loakman[3], Chenghua Lin[1]
[1]The University of Manchester, [2]Durham University, [3]The University of Sheffield
yang.wang-27@postgrad.manchester.ac.uk, chenghua.lin@manchester.ac.uk

Abstract

We introduce Drivelology, a unique linguistic phenomenon characterised as "nonsense with depth" – utterances that are syntactically coherent yet pragmatically paradoxical, emotionally loaded, or rhetorically subversive. While such expressions may resemble surface-level nonsense, they encode implicit meaning requiring contextual inference, moral reasoning, or emotional interpretation. We find that current large language models (LLMs), despite excelling at many natural language processing (NLP) tasks, consistently fail to grasp the layered semantics of Drivelological text. To investigate this, we construct a benchmark dataset of over 1,200+ meticulously curated and diverse examples across English, Mandarin, Spanish, French, Japanese, and Korean. Each example underwent careful expert review to verify its Drivelological characteristics, involving multiple rounds of discussion and adjudication to address disagreements. Using this dataset, we evaluate a range of LLMs on classification, generation, and reasoning tasks. Our results reveal clear limitations of LLMs: models often confuse Drivelology with shallow nonsense, produce incoherent justifications, or miss implied rhetorical functions altogether. These findings highlight a deep representational gap in LLMs' pragmatic understanding and challenge the assumption that statistical fluency implies cognitive comprehension. We release our dataset[1] and code[2] to facilitate further research in modelling linguistic depth beyond surface-level coherence.

1 Introduction

Large language models (LLMs) have achieved impressive success across a wide range of natural language processing tasks, from machine translation and summarisation to commonsense reasoning and dialogue generation (Tang et al., 2023; Achiam et al., 2023; Qwen Team, 2024; Liu et al., 2024a; Guo et al., 2025; Wang et al., 2025; Goldsack et al., 2025). These models exhibit high degrees of fluency, contextual awareness, and even emergent reasoning capabilities. However, whether such performance reflects genuine understanding or merely statistical pattern-matching remains an open and pressing question (Bender et al., 2021; Rayhan et al., 2023).

The continuous evolution of Internet language as a distinct linguistic style offers a novel and insightful avenue for exploring the depth of understanding in LLMs (Ignacio et al., 2024; Mei et al., 2024). Internet language, characterised by its dynamic evolution and cultural embedding, serves as an effective indicator to assess whether models truly grasp deeper semantics or simply rely on superficial pattern recognition. In particular, we introduce the term *Drivelology*, combining *drivel* (i.e., nonsense) with *-ology* (i.e., the study of), which exemplifies this complexity. Drivelology often involves narratives with dual or multiple layers of meaning, employing non-linear structures and ambiguous expressions that challenge LLMs. Unlike purely nonsensical yet grammatically correct sentences such as "*Colourless green ideas sleep furiously*" (Chomsky, 1957), or simplistic tautologies like "*either it is or it isn't*", Drivelology intentionally embeds subtle cultural references, irony, or satire within superficially trivial or absurd narratives. For example, "*I deeply admire Che Guevara's anti-capitalist spirit, so I bought all his merchandise*" illustrates how Drivelology paradoxically critiques performative activism. Thus, it differs from typical internet content, such as inspirational quotes or prose, by demanding deeper interpretative engagement from both human readers and LLMs.

Existing studies have explored the difficulty for LLMs to understand humour, sarcasm, and irony

[1]https://huggingface.co/datasets/extraordinarylab/drivel-hub
[2]https://github.com/ExtraOrdinaryLab/drivelology.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23086–23108
November 4-9, 2025

(Loakman et al., 2023, 2025; Romanowski et al., 2025; Zheng et al., 2025). However, Drivelology differs fundamentally from these phenomena by employing more complex narratives and deeper ambiguities, making it a uniquely challenging benchmark for assessing LLMs' semantic comprehension. Studying LLMs' ability to handle Drivelology offers insights into their social and semantic reasoning, as it encodes subtle emotions and culturally embedded meanings. Understanding such linguistic forms is essential for developing socially intelligent systems (Gandhi et al., 2023; Kosinski, 2024; Mittelstädt et al., 2024). Moreover, enhancing AI's grasp of Drivelology can potentially boost creativity in AI applications, improving user experience in content creation tools (Hu et al., 2024), and even contribute to model safety (Matamoros Fernandez et al., 2023) through better understanding of contextually ambiguous content.

Our contributions are as follows:

- We design a novel taxonomy for Drivelological narratives to aid in categorising the source of meaning embedded in the text.
- We collect and rigorously annotate a novel benchmark dataset called DRIVELHUB for understanding Drivelology from the internet, consisting of 1,200+ Drivelological examples that are considered *nonsense with depth*.
- We use DRIVELHUB as the basis for four novel tasks: (1) Drivelology Detection: A binary classification task to determine whether a given text is Drivelology or non-Drivelology; (2) Drivelology Tagging: A multi-label classification task to assign one or more categories from our taxonomy (§3.1) to Drivelology samples; (3) Implicit Narrative Writing: An implicit narrative explanation task for a given Drivelology sample; and (4) Narrative Selection: A multiple-choice task where the model selects the correct narrative from five options.

These tasks collectively encompass various levels of Drivelology understanding, ranging from literal content comprehension to more sophisticated narrative reasoning, thereby providing a comprehensive assessment of Drivelology understanding capabilities. We conducted extensive experiments using the DRIVELHUB dataset, evaluating both proprietary and open-source LLMs.

2 Related Work

LLMs Evaluations. Recent LLMs have demonstrated remarkable performance in following human instructions and performing various downstream tasks through zero-shot prompting (Naveed et al., 2023; Liang et al., 2024; Chang et al., 2024; Liu et al., 2024c,b). Various benchmarks have been proposed to evaluate their performance, primarily focusing on assessing the fundamental capabilities of LLMs (Zellers et al., 2019; Sakaguchi et al., 2021; Hendrycks et al., 2021; Suzgun et al., 2023; Zheng et al., 2023; Zhou et al., 2023; Jimenez et al., 2024; Yang et al., 2025; Wang and Lin, 2025; Hong et al., 2025). However, the ability of large models to perform in-depth social reasoning and accurately understand human contexts remains underexplored (Hu and Shu, 2023; Feng et al., 2024).

Humour, Irony, and Sarcasm. Humour, irony, and sarcasm are fundamental elements of human interaction, each requiring a deep understanding of language, context, and social cues (Palmer, 2003; Filik et al., 2016; Köder and Falkum, 2021; Mir and Laskurain-Ibarluzea, 2021; Loakman et al., 2023, 2025). While computational approaches have addressed these phenomena (Yang et al., 2015; Jentzsch and Kersting, 2023; Boutsikaris and Polykalas, 2024), they often focus on resolving a central contradiction between a literal statement and its context (Kreuz and Link, 2002; Misra and Arora, 2019). Classic sarcasm, for example, is typically understood through a single cognitive step: inverting a word's meaning based on a negative premise, as in, "*Forgetting assignments and stressing over grades, what a fun semester*".

We argue that Drivelology presents a more profound challenge, distinguished by two key characteristics: **(1) its compositional, multi-layered structure, and (2) its use of pragmatic paradox and ambiguity.** For example, "*I deeply admire Che Guevara's anti-capitalist spirit, so I bought all his merchandise*" is not a simple semantic inversion. Its critique of performative activism requires synthesising cultural knowledge, making irony just one component of its layered meaning. Furthermore, Drivelology uses pragmatic paradoxes like "*I'm good at everything except what I can't do.*" This statement is not explicitly sarcastic nor ironic; its challenge lies in navigating the ambiguity of the speaker's intent. This reliance on compositional meaning and deliberate ambiguity is what sets Drivelology apart.

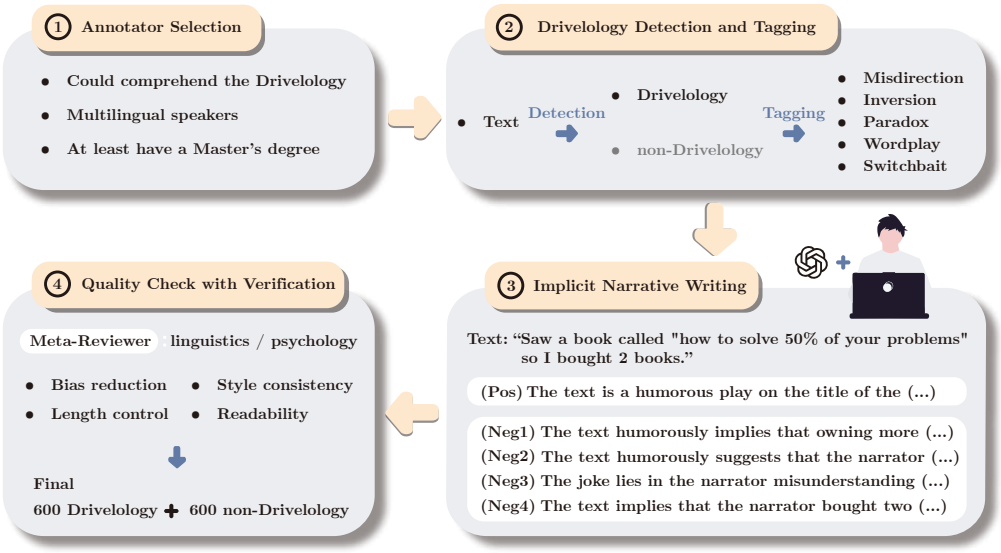

Figure 1: Overview of the multi-stage process for constructing the DRIVELHUB dataset.

Distinguishing Drivelology from Deceptive and Nonsensical Language. To properly situate our work, it is crucial to distinguish Drivelology from related pragmatic concepts. Cappelen and Dever (2019) identify a category termed *deep bullshit*: utterances defined by an indifference to whether the words make any sense at all, resulting in genuine nonsense. For example, a statement like "*Colourless green ideas sleep furiously*" (Chomsky, 1957) qualifies as deep bullshit as it is semantically null. This is distinct from the more widely discussed Frankfurt-style *bullshit*, which is characterised by an indifference to *truth* rather than meaning, often deployed to persuade without regard for fact (Frankfurt, 2005). For instance, a politician might declare they bring a "*fresh perspective, unburdened by the stagnant thinking of Washington insiders*", a statement chosen for its persuasive effect, not its accuracy. Drivelology shares a superficial resemblance with deep bullshit, as both can appear nonsensical. However, the two are fundamentally antithetical in their purpose and construction. Whereas deep bullshit arises from a *disregard for meaning*, Drivelology is meticulously crafted *for the sake of conveying a hidden meaning*. It is, as we define it, "nonsense with depth".

The surface-level absurdity of a Drivelogical text is a deliberate rhetorical framework, designed to guide an audience toward an implicit critique, observation, or emotional payload. Thus, unlike vacuous deep bullshit, Drivelology is rhetorically complex and purposeful. While other forms of bad language, such as lying or misleading, are defined by their deceptive relationship to truth (Cappelen and Dever, 2019), Drivelology's defining feature is its purposeful and creative use of apparent nonsense to generate layered semantics. Clarifying these boundaries is essential for developing AI systems that can appreciate subtle human expression. A truly capable model must differentiate between genuinely meaningless utterances (deep bullshit) and the sophisticated, implicit communication of Drivelology, a task that requires moving beyond surface-level coherence to grasp complex rhetorical intent.

3 The DRIVELHUB Dataset

Our benchmark dataset, DRIVELHUB, is designed to evaluate how well LLMs understand Drivelology. Each entry in the dataset includes: (1) a Drivelology sample, (2) the underlying message that the sample aims to convey, and (3) one or more categories describing the main type of Drivelology contained in the sample. These components form the basis for a variety of tasks that assess different aspects of Drivelology comprehension and reasoning. An overview diagram of the multi-stage process for constructing the DRIVELHUB dataset is presented in Figure 1 and Appendix A.1.

3.1 Taxonomy of Drivelology

Drivelology refers to a unique style of language that blends humour, ambiguity, and rhetorical complex-

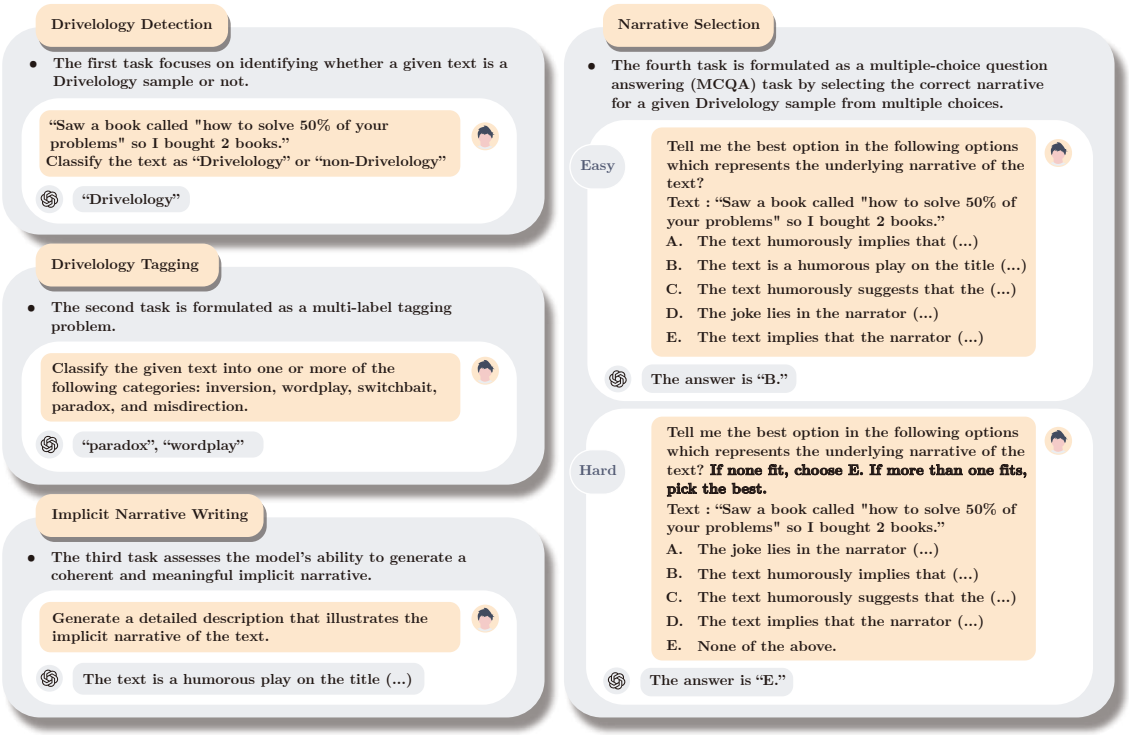

Figure 2: Overview of the Drivelology evaluation framework for LLMs. The figure illustrates four core tasks designed to systematically assess LLMs' ability to understand and reason about Drivelology: Drivelology Detection (binary classification), Drivelology Tagging (multi-label classification), Implicit Narrative Writing (generative reasoning), and Narrative Selection (multiple-choice question answering with both Easy and Hard settings).

ity to create statements that are intentionally puzzling or nonsensical. Unlike ordinary nonsense or straightforward jokes, Drivelology often relies on layered meanings, unexpected twists, and linguistic playfulness to engage readers in deeper interpretation or amusement. The defining characteristics of Drivelology can be broadly categorised into the following taxonomy:

Misdirection. This technique leads the listener down an expected path before a final twist reveals a different, often more literal or absurd, ending. Example: *"Don't give up on your dream so easily! Keep sleeping!"* The expected path is motivational encouragement; the twist is a literal interpretation of "dream."

Paradox. This relies on a statement that appears logically self-contradictory but contains a latent, often humorous or profound truth. The core of the technique is the clash of seemingly incompatible ideas. Example: *"I will not forget this favour until I forget it."* This is a logically circular statement that humorously asserts the certainty of remembering.

Switchbait. This technique hinges on a specific phrase (the "bait") that has a culturally-embedded double meaning. The initial context is then suddenly replaced (the "switch") by a surprising second meaning. The humour is generated by this cynical or culturally-specific reinterpretation of the bait, rather than by derailing a narrative. Example: *"Brit: You've got a gun problem. American: Yeah, at least it's a modern problem."* The bait is the phrase "gun problem." The switch reframes it from a criticism of US gun violence to a dark turn, implying cultural counter-attack on UK knife crime.

Inversion. This technique takes a well-known phrase, cliché, or social script and flips it on its head. The humour arises by reversing a familiar structure to creating a new, often satirical, meaning. Example: *"Other than being good-looking, having a great figure, and having money, I have nothing else."* This inverts the structure of a humble complaint into an arrogant boast.

Wordplay. This is the use of linguistic creativity, often by exploiting the phonetics or polysemy of words. It includes puns, double entendres, and similarities. Example: *"Do you have any raisins? No? How about a date?"* This is a classic homographic pun playing on two meanings of the word "date".

We note that the defining characteristic of Drivelology is not the use of a single technique, but the creative and often simultaneous combination of several within a single utterance to produce its layered, nonsensical effect. This inherent complexity is central to our study, which is why the Drivelology Tagging task is formulated as a multi-label classification problem (see §3.4), allowing a single sample to be annotated with one or more of the following categories.

3.2 Drivelology Collection

To ensure a comprehensive evaluation, we prioritised high diversity in our benchmark dataset by selecting a wide range of topics from different languages. Our data was collected from a variety of popular platforms (e.g., Instagram, Threads, TikTok, Facebook, Line, RedNote, Pinterest, Naver, and YouTube). These platforms were chosen strategically as their largest user demographic falls between 25 to 34 years old,[3] which aligns well with our research focus since Drivelology content predominantly originates from younger generations (Sha, 2024). For non-Drivelology samples, we curated content from sources such as famous quotes, proverbs, and Ruozhiba (a popular online forum). These non-Drivelology samples are also multilingual, covering English, Mandarin, Spanish, French, Japanese, and Korean. We further categorised non-Drivelology samples into two types: normal sentences (such as meaningful quotes or proverbs) and pure nonsense (text that lacks logical structure or meaning). A significant proportion of the pure nonsense samples were collected from Ruozhiba.

3.3 Data Annotation

Labelling Drivelology requires both content comprehension and cultural context understanding. We implemented a rigorous four-step annotation protocol: (1) **Annotator Selection:** We assembled a team of seven multilingual annotators who all held at least a Master's degree and demonstrated proficiency in multiple languages.[4] (2) **Drivelology Detection and Tagging:** Annotators identified texts as either Drivelology or non-Drivelology, and classified Drivelology samples into categories including Misdirection, Paradox, Switchbait, Inversion, and Wordplay. (3) **Implicit Narrative Writing:** We employed a human-in-the-loop process to create the narrative explanations. For each Drivelology sample, human experts drafted and refined the correct narrative explanation. We then utilised GPT-4.5[5] as an assistive tool to generate four plausible but incorrect narrative counterparts, all of which underwent a final stage of manual verification and editing to ensure their quality as effective distractors. (4) **Quality Check:** A meta-reviewer with linguistics and psychology expertise reviewed all annotations. The meta-reviewer also revised the narratives as needed to ensure consistent length, uniform writing style, and improve overall readability. Further details concerning the annotation process can be found in Appendix A.1.

3.4 Task Design

To evaluate an LLM's ability to understand Drivelology, we designed four tasks to assess different facets of social and non-linear reasoning. An overview of these tasks is provided in Figure 2.

Drivelology Detection. A binary classification task where the model must determine if a given text is Drivelology or non-Drivelology.

Drivelology Tagging. A multi-label classification task where the model assigns one or more descriptive categories (see §3.1) to a Drivelology sample to capture its layered rhetorical structure.

Narrative Writing. A generative task where the model explains the implicit narrative and underlying meaning of a Drivelology sample, requiring it to move beyond a surface-level reading.

Narrative Selection. A multiple-choice question answering (MCQA) task where the model is tasked with selecting the correct narrative for a Drivelology sample from several options. The **Easy** version offers one correct answer and four distractor, whilst the **Hard** version adds a "none of the above" option, requiring deeper reasoning, as this option should only be chosen if none of the provided narratives adequately capture the underlying meaning of the Drivelology sample. This additional step significantly increases the task's complexity, as it prevents reliance on simple elimination strategies.

[3] According to Statista's social media demographics data as of September 2025. https://www.statista.com/topics/1164/social-networks

[4] The annotation team consisted of four authors of this paper and three paid annotators recruited for their linguistic expertise.

[5] gpt-4.5-preview-2025-02-27

4 Experiments

4.1 Models and Settings

We evaluate the performance of state-of-the-art LLMs in a zero-shot setting. We utilise both proprietary models including GPT-4 (Achiam et al., 2023) and Claude-3 (Anthropic, 2024), as well as open-sourced models including Qwen3 (Qwen Team, 2025), Qwen2.5 (Qwen Team, 2024), Llama3.1 (Grattafiori et al., 2024), Llama3 (Grattafiori et al., 2024), and DeepSeek V3 (Liu et al., 2024a).

To minimise variance across task prompts, we design three distinct prompts for each task and report the average performance over three runs (one for each prompt). Detailed descriptions of the prompts and additional experimental settings are provided in Appendix B.1.

4.2 Evaluation Metrics

We use accuracy for the Drivelology Detection task, F1 score for the Drivelology Tagging task, and accuracy for the MCQA task. For the generation task that involves writing narrative explanations, we apply reference-based evaluation metrics commonly used in text generation studies (Celikyilmaz et al., 2020). Specifically, we use BERTScore (Zhang et al., 2020) and an LLM-as-a-judge evaluation paradigm (Zheng et al., 2023). Recent work shows that GPT-based evaluation aligns well with human judgments (Chan et al., 2023; Liu et al., 2023; Hu et al., 2024; Gu et al., 2024), and thus we select GPT-4 series for LLM-as-a-judge evaluation. The judge was tasked to rate each generated narrative on a 1-to-5 Likert scale based on its semantic quality. Note that we use different GPT variants for different purposes: gpt-4.5 for data annotation, gpt-4o-mini for zero-shot experiments, and gpt-4.1 for LLM-as-a-judge evaluation in text generation tasks. This helps reduce potential evaluation bias toward GPT-4's own generation (Hu et al., 2024; Liu et al., 2024b).

5 Main Results

The main results in Table 1 show a clear hierarchy in model performance. Deepseek-v3 is the dominant model, achieving the top score in five of the six evaluated metrics. The contrast between the two evaluation metrics in the Narrative Writing task is particularly noteworthy. While BERTScore-recall values are high across all models, suggesting a universal proficiency in generating fluent text, the GPT-4-as-a-judge scores provide a much clearer picture of true narrative quality. On this scale, deepseek-v3 (3.59) and claude-3.5-haiku (3.39) are the only models to score comfortably above three, indicating their outputs were judged as possessing high semantic quality. In stark contrast, other models like llama-3-8b-instruct (2.63) and qwen3-8b-instruct (2.64) fall below this threshold, suggesting their narratives failed to capture the required depth and were deemed qualitatively weaker by the LLM-as-a-judge. The most striking performance gap is present in the MCQA task. The Hard setting causes a steep decline in accuracy for all models, exposing a critical weakness in subtle reasoning. Notably, qwen3-8b-instruct is an outlier here, scoring 26.78%, which far surpasses the next-best model. In the Classification tasks, deepseek-v3 again confirms its superior understanding by leading in both the Detection (81.67%) and Tagging (55.32%) tasks.

5.1 Prompt Language Influence

An analysis of the impact of prompt language on model performance reveals a metric-dependent pattern. As shown in Figure 3, the choice between English and Mandarin prompts is not a neutral choice but a significant factor that influences evaluation outcomes. We identify two distinct and opposing patterns. Firstly, English prompts consistently yield superior performance on tasks that reward lexical precision and complex logical reasoning. In the Narrative Writing task, this is most evident in the BERTScore results, where every model scores higher when prompted in English. This suggests that English instructions may prime the models to generate outputs with greater lexical overlap with the reference translations, a feature to which BERTScore is highly sensitive (Hanna and Bojar, 2021). A similar advantage for English prompts is observed in the MCQA task. The uniform improvement in both Easy and Hard settings implies that English may serve as a more robust internal *language of reasoning*. Conversely, Mandarin prompts produce a consistent, albeit smaller, advantage on tasks that prioritise direct content comprehension. The improved performance under GPT-as-a-judge, which evaluates qualitative coherence, indicates that Mandarin prompts better align the models with the semantic and narrative intent of the source material. The consistent gains in the classification tasks further suggest that for direct comprehension and categorisation, instructions in Mandarin are more effective.

Models	Narrative		MCQA		Classification	
	BERT	GPT	Easy	Hard	Detect	Tag
gpt-4o-mini	85.81	2.90	81.89	4.67	75.00	49.52
claude-3.5-haiku	86.51	3.39	83.17	11.56	71.90	52.03
llama-3-8b-instruct	84.67	2.63	77.39	1.67	57.81	39.90
llama-3.1-8b-instruct	85.60	2.75	77.56	1.89	58.57	36.21
qwen2.5-7b-instruct	85.51	2.78	77.50	3.78	62.66	42.49
qwen3-8b-instruct	85.91	2.64	83.17	**26.78**	65.00	38.04
deepseek-v3	**87.11**	**3.59**	**86.83**	15.50	**81.67**	**55.32**

Table 1: Main results. For the narrative explanation writing task, we report BERTScore-recall (BERT) and GPT-4-as-a-judge (GPT) evaluation scores. For the narrative selection task, we report accuracy. For the Drivelology classification tasks, we report accuracy for the detection task and weighted F1 score for the tagging task. The best scores are **bold** and the second best ones are underlined.

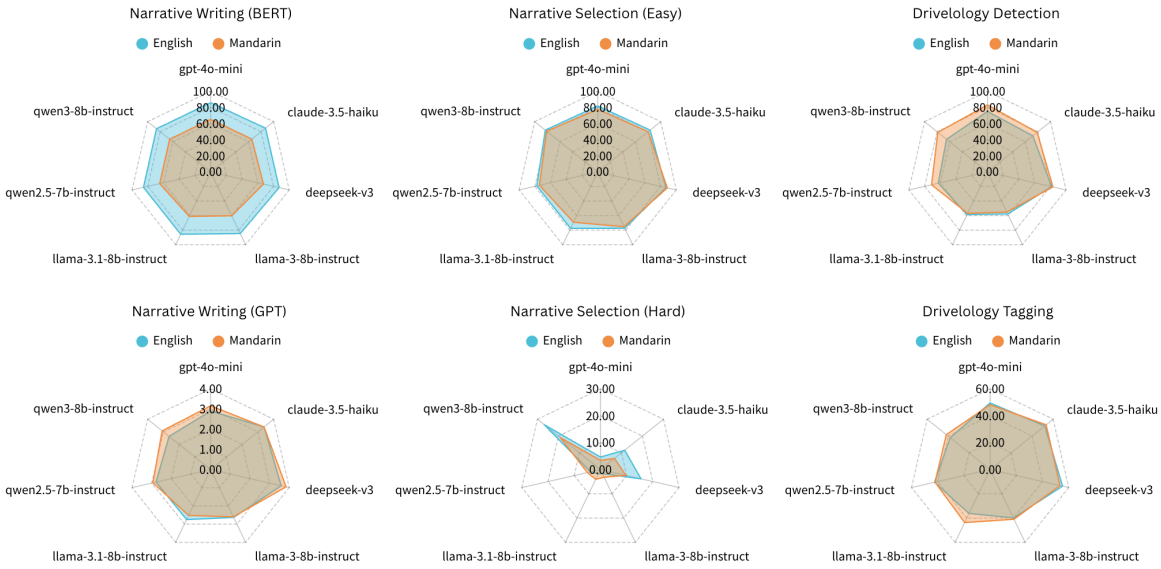

Figure 3: Model performance on the multilingual DRIVELHUB dataset, contrasted by prompt language (English vs. Mandarin). Each reported score is the average performance over three distinct prompts to minimise variance.

5.2 Model Size Scaling in the Qwen3 Series

The results in Table 2 illustrate the impact of model size on performance. We focus on the Qwen3 series across the MCQA and Classification tasks, as these tasks exhibited the widest performance variance in Table 1, making them most suitable for studying scaling effects. For the Easy MCQA task, performance gains are consistent but modest: as model size increases from 4B to 14B, accuracy improves by approximately 3% for English prompts and 6% for Mandarin. The Hard task reveals a spiking scaling effect. With English prompts, accuracy leaps from a mere 6.00% for the 4B model to 45.83% for the 14B model. The trend is even more pronounced with Mandarin prompts, where the score skyrockets from 2.44% to 47.89%. This indicates that the more complex reasoning required by the Hard task is a key differentiator that is unlocked by larger model sizes. Therefore, the ability to handle such complex reasoning appears to be an emergent property in the Qwen3 architecture, strongly correlated with its parameter count. For the Detection task, performance does not consistently improve with size. Notably, when prompted in Mandarin, the 8B model significantly outperforms both its smaller and larger counterparts. The Tagging task reveals yet another pattern: a noticeable dip in performance at the 8B size, which then recovers to achieve the best score at 14B for both languages. These findings indicate that the benefits of model scaling are task-dependent.

Prompt	Size	MCQA		Classification	
		Easy	Hard	Detect	Tag
English	4B	81.00	6.00	66.80	43.21
	8B	83.17	26.78	65.00	38.04
	14B	83.94	45.83	66.22	47.61
Mandarin	4B	77.61	2.44	62.86	46.10
	8B	81.11	19.11	78.81	41.71
	14B	83.50	47.89	71.78	49.13

Table 2: MCQA and Classification results in the Qwen3 series of different sizes. This table shows the performance on both tasks when prompted in English and Mandarin. The Full version containing all tasks can be found in Table 5.

5.3 Role of Language in the MCQA Task

A closer look at the MCQA results from Table 1 reveals that aggregate scores mask significant performance variations across the different languages in DRIVELHUB. As shown in the breakdown in Figure 4, we can analyse the difficulty of each language's content for the models. Deepseek-v3 consistently demonstrates the most robust cross-lingual performance, achieving the highest accuracy across nearly all languages in both the Easy and Hard settings. The analysis also pinpoints which languages pose a greater challenge. Korean and Mandarin consistently result in the lowest accuracy, especially in the Hard task, marking their content as the most difficult for the models to process.

6 Analysis and Discussion

6.1 Analysis of Model Reasoning

In the Narrative Writing task, claude-3.5-haiku and deepseek-v3 achieve the highest GPT-4-as-a-judge scores (3.39 and 3.59) and also perform strongly in the Drivelology Tagging task (52.03% and 55.32%). Thie correlation between their performance in both tasks raises an important question: *Do these models arrive at correct Drivelology classifications through appropriate reasoning that reflects a true understanding of the underlying meaning?* To investigate this question, we analyse their reasoning processes across several representative examples.

For example: "*Meng Po: Those who have forgotten their names, please follow me.*" Deepseek-v3 categorise this as *switchbait*, emphasising the cultural significance of Meng Po, a mythological figure who administers the Soup of Forgetfulness in Chinese folklore. Its reasoning explicitly highlights the importance of cultural knowledge, suggesting they treat Meng Po's mythological role as important context that readers must understand to appreciate the Drivelology. In contrast, claude-3.5-haiku categorises it as a *paradox*, focusing on the logically self-contradictory statement: "how can someone who has forgotten their name respond to such a call?" This divergence in reasoning approaches suggests varying degrees of cultural knowledge *internalisation* among models. Claude-3.5-haiku appears to have so thoroughly internalised the cultural context of Meng Po that it treats it as implicit knowledge, allowing it to focus on the logical structure of the text rather than its cultural elements. This observation raises important questions about how different models process and prioritise cultural knowledge versus logical reasoning in their analysis of culturally-embedded texts, and how such internalisation affects their ability to identify different categories of Drivelology.

6.2 Analysis of Human Reasoning

Drivelology challenges not only LLMs but also human annotators, who often bring diverse perspectives and interpretations to the same text. Because Drivelology is intentionally ambiguous, contradictory, or ironic, it invites multiple plausible readings. Annotators rely on their own linguistic, cultural, and contextual knowledge, which means that the same Drivelology sample can evoke different analytical frameworks depending on who is interpreting it.

Consider the statement: "*I hate two kinds of people: the first kind is those who don't finish their...*" From a *paradox* perspective, the speaker claims to dislike people who speak incompletely, yet the sentence itself is left unfinished, ironically exemplifying the very behaviour it criticises. This self-contradiction highlights the speaker's insincerity and creates a paradoxical effect. Alternatively, from a *misdirection* viewpoint, the statement sets up the expectation of a complete list, but then abruptly stops, leaving the audience anticipating an answer that never comes. The humour and irony arise from this unresolved expectation. Another example is: "*I deeply admire Che Guevara's anti-capitalist spirit, so I bought all his merchandise.*" Here, the *paradox* lies in admiring Guevara's anti-capitalist stance while simultaneously engaging in capitalist consumerism by buying his merchandise. This contradiction turns ideological admiration into commercial participation. The *switchbait* interpre-

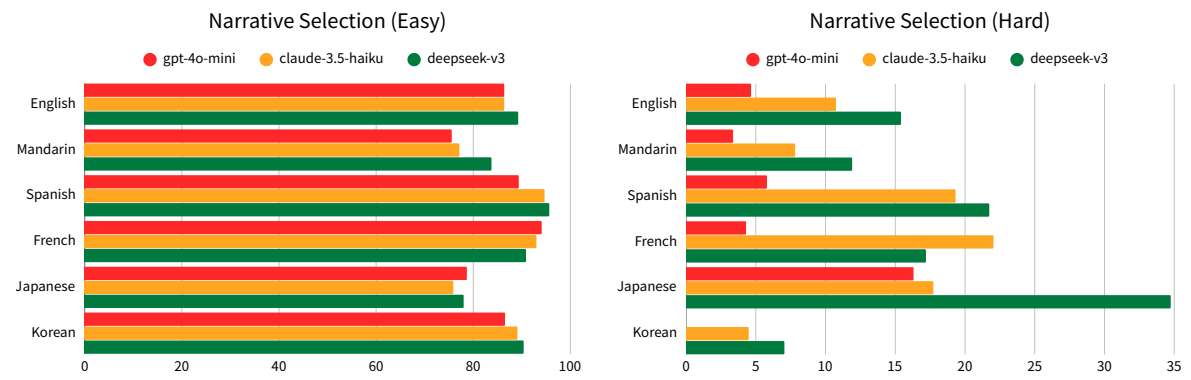

Figure 4: A language-based breakdown of Narrative Selection (MCQA) accuracy from Table 1. The charts disaggregate the overall Easy and Hard accuracy scores based on the original language of the Drivelology sample.

tation depends on cultural knowledge: recognising Che Guevara as a symbol of anti-capitalism is key. Without this context, the contradiction, and the humour, may not be apparent. The text's irony and layered meaning rely on shared cultural and historical understanding, making switchbait also an appropriate label.

7 Conclusion

In this work, we introduced Drivelology, a unique linguistic phenomenon that challenges the semantic and pragmatic understanding of LLMs. We constructed and evaluated the DRIVELHUB dataset across multiple languages and task settings. Our extensive experiments reveal a critical and consistent gap between statistical fluency and genuine comprehension in state-of-the-art LLMs. While models can generate syntactically coherent text, they largely fail to grasp the layered, culturally-embedded meanings central to Drivelology. We found that complex reasoning, particularly on the "Hard" MCQA task, remains a significant bottleneck, though performance scales predictably with model size. Conversely, performance on classification tasks showed non-linear scaling, suggesting that simply increasing parameter count is not a panacea for all reasoning deficits. The failure of LLMs to interpret Drivelology underscores a deep representational gap in their ability to model complex social and cultural contexts. Our work provides a concrete benchmark for the community to address these deeper challenges. Future research should focus not only on scaling models but also on developing novel training paradigms that explicitly target the multi-layered reasoning that defines sophisticated human communication.

Limitations

Language Imbalance. Over one third of the samples in the DRIVELHUB dataset are in Mandarin (see Table 4). This results in a slight language imbalance, which may affect the generalisability of our findings across different linguistic and cultural contexts. We do our best in §5 to control for potential content distribution bias arising from this imbalance. Additionally, as the DRIVELHUB dataset is still being expanded, we will continue to focus on addressing distribution differences by increasing the representation of Drivelology samples in underrepresented languages.

Limited Computation Resources. Due to budget constraints, we were unable to evaluate stronger proprietary LLMs such as GPT-5, Claude-3.7, or DeepSeek R1, as their usage costs are prohibitively high. For open-source models, we restricted our experiments to 14B parameter models because of limited computational resources, and were unable to run larger models within our available infrastructure. We encourage researchers and the broader community to expand on this work by evaluating larger or more advanced LLMs as resources permit.

Focus on Understanding Rather Than Generation. In this paper, we focus on evaluating the understanding and reasoning abilities of LLMs with respect to Drivelology, rather than their capacity to generate fluent and human-like Drivelology text. While generation is an important aspect, it falls outside the main scope of our study. Nevertheless, we include a discussion in Appendix C with sample generations, illustrating that current LLMs often require over 20 attempts to produce Drivelology that achieves comprehensive alignment between topic, rhetorical category, and sentence structure.

Ethics Statement

Copyright and License. All data samples used in this study are collected exclusively from publicly available content on social media platforms. We respect the intellectual property rights of original authors by ensuring that no proprietary or paywalled material is included. The dataset is released solely for research purposes under a license that prohibits commercial use and redistribution of original content.

Content Review and Harm Mitigation. To uphold ethical standards, we carefully review all collected samples and filter out any content that may be offensive, harmful, or violate privacy. Our annotation process is designed to ensure that sensitive information is excluded and that the dataset does not propagate hate speech, harassment, or other forms of harmful language.

Intended Use. The dataset and accompanying resources are intended strictly for academic research and the advancement of natural language processing technologies. Users are advised to adhere to ethical guidelines and local regulations when using the dataset.

Acknowledgments

Tyler Loakman is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation (EP/S023062/1).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Leonidas Boutsikaris and Spyros Polykalas. 2024. A comparative review of deep learning techniques on the classification of irony and sarcasm in text. *IEEE Transactions on Artificial Intelligence*.

Herman Cappelen and Josh Dever. 2019. *Bad language*. Oxford University Press.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from AGI: Are LLMs all we need? *Transactions on Machine Learning Research*. Survey Certification.

Ruth Filik, Alexandra Ţurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Harry G. Frankfurt. 2005. *On bullshit*. Princeton University Press, Princeton, NJ.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2025. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hanhua Hong, Chenghao Xiao, Yang Wang, Yiqi Liu, Wenge Rong, and Chenghua Lin. 2025. Beyond one-size-fits-all: Inversion learning for highly effective nlg evaluation prompts. *arXiv preprint arXiv:2504.21117*.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024. Cracking the code of juxtaposition: Can ai models understand the humorous contradictions. *arXiv preprint arXiv:2405.19088*.

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Marvin John Ignacio, Thanh Tin Nguyen, Hulin Jin, and Yong-guk Kim. 2024. Meme analysis using llm-based contextual information and u-net encapsulated transformer. *IEEE Access*, pages 1–1.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Franziska Köder and Ingrid Lossius Falkum. 2021. Irony and perspective-taking in children: The roles of norm violations and tone of voice. *Frontiers in Psychology*, 12:624604.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Roger J Kreuz and Kristen E Link. 2002. Asymmetries in the use of verbal irony. *Journal of language and social psychology*, 21(2):127–143.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024b. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Comparing apples to oranges: A dataset & analysis of llm humour understanding from traditional puns to topical jokes. *Preprint*, arXiv:2507.13335.

Ariadna Matamoros Fernandez, Louisa Bartolo, and Luke Troynar. 2023. Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, 12(1).

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. Slang: New concept comprehension of large language models. *arXiv preprint arXiv:2401.12585*.

Montserrat Mir and Patxi Laskurain-Ibarluzea. 2021. Spanish and english verbal humour: A comparative study of late-night talk show monologues. *Contrastive Pragmatics*, 3(2):278–312.

Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Justin M Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14(1):27449.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Jerry Palmer. 2003. *Taking humour seriously*. Routledge.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abu Rayhan, Rajan Rayhan, and Swajan Rayhan. 2023. Artificial general intelligence: Roadmap to achieving human-level capabilities. *DOI*, 10:13140.

Adrianna Romanowski, Pedro HV Valois, and Kazuhiro Fukui. 2025. From punchlines to predictions: A metric to assess llm performance in identifying humor in stand-up comedy. *arXiv preprint arXiv:2504.09049*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Rula Sha. 2024. Literary effect of "nonsense literature": A perspective of rhetorical style. *Journal of Eastern Liaoning University (Social Sciences)*, 26(1).

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Enhancing dialogue generation via dynamic graph knowledge aggregation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616, Toronto, Canada.

Yang Wang and Chenghua Lin. 2025. Tougher text, smarter models: Raising the bar for adversarial defence benchmarks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6475–6491, Abu Dhabi, UAE. Association for Computational Linguistics.

Yang Wang, Chenghao Xiao, Yizhi Li, Stuart E Middleton, Noura Al Moubayed, and Chenghua Lin. 2025. Adversarial defence without adversarial defence: Enhancing language model robustness via instance-level principal component removal. *arXiv preprint arXiv:2507.21750*.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2367–2376.

John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. 2025. SWE-bench multimodal: Do ai systems generalize to visual software domains? In *The Thirteenth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yilun Zheng, Sha Li, Fangkun Wu, Yang Ziyi, Lin Hongchao, Zhichao Hu, Cai Xinjun, Ziming Wang, Jinxuan Chen, Sitao Luan, and 1 others. 2025. Fanchuan: A multilingual and graph-structured benchmark for parody detection and analysis. *arXiv preprint arXiv:2502.16503*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Dataset Details

A.1 Overview of the Annotation Process

Labelling Drivelology presents significant challenges, not only because it demands a deep familiarity with both content and cultural context, but also due to the potential for divergent interpretations among annotators from varied backgrounds. For each Drivelology sample, we annotate the underlying narrative and the category of the Drivelology. To ensure high-quality and precise annotations, we designed a multi-step annotation protocol as follows: (1) **Annotator Selection.** We recruited multilingual annotators, and ensured that they could comprehend the Drivelology. Eight human judges[6] participated in the annotation process, all of whom are proficient Mandarin and English speakers (some speak more than three languages) and have at least a Master's degree. (2) **Drivelology Detection and Tagging.** Each annotator was tasked with determining whether a given text is non-Drivelology or Drivelology. Non-Drivelology includes both normal, meaningful sentences and pure nonsense that lacks rhetorical or semantic structure. If the text is identified as Drivelology, the annotators then perform a multi-label classification task, assigning one or more of the following categories to the sample: Misdirection, Paradox, Switchbait, Inversion, and Wordplay. (3) **Implicit Narrative Writing.** Given a Drivelology sample, we first prompt GPT-4 to generate narrative descriptions, illustrating the Drivelology's narrative and explaining the underlying meaning. Human annotators then double-check and modify the contents through dialogue interactions with the GPT-4 model to obtain a correct narrative. Additionally, we prompt GPT-4 to generate four hard negative counterparts to form a multiple-choice question answering task for our experiment. As narrative writing is inherently open-ended and involve subjectivity, we additionally frame this as selection tasks, and ensure that the correct option is clearly and objectively superior than the negative options to mitigate subjectivity. Following Achiam et al. (2023), We primarily rely on human annotators to obtain gold-standard annotations, while allowing the annotators to collaborate with GPT-4. (4) **Quality Check with Verification.** To further minimise annotation errors, an experienced meta-reviewer with a background in linguistics and psychology systematically reviewed all annotated samples. The meta-reviewer excludes the samples with ambiguous or controversial narratives as some of them may introduce bias. This process ensures the quality of the annotated components for benchmark dataset construction.

[6]The original annotation was performed by seven annotators, and a psychology/linguistics expert made the final decision.

A.2 Dataset Distribution

Table 4 presents the language distribution of samples in the DRIVELHUB dataset. As shown, the dataset is skewed toward Mandarin, which accounts for 277 out of the total Drivelology samples. In contrast, other languages such as Japanese and Korean are present only in limited quantities. To characterise the distribution and overlap of annotation categories in our dataset, we present an UpSet plot (Lex et al., 2014) in Figure 5, summarising intersections among the five Drivelology categories.

B Experimental Details

B.1 Experiment Prompts

To ensure reproducibility and transparency, we provide the exact prompts used in each of our experimental tasks. These prompts were carefully designed to probe different aspects of Drivelology comprehension and generation across various LLMs. Below, we detail the prompts for each task: Drivelology Detection (Figure 7), Drivelology Tagging (Figure 8), Narrative Writing (Figure 9 for generation and Figure 10 for evaluation), Narrative Selection (Figure 11 for Easy and Figure 12 for Hard).

C Drivelology Generation

To explore the generative capabilities of LLMs in Drivelology, we conducted a case study using GPT-4. Our goal was to assess whether the model can produce contextually natural and pragmatically rich Drivelology examples, focusing on both surface form and deeper pragmatic alignment.

We designed two experimental settings: (1) Minimal Guidance and (2) Guidance With Category Definitions. In the Minimal Guidance setting, the model received only a brief introduction to Drivelology, without any example texts or category definitions. In the Category Definitions setting, the model was provided with detailed definitions of the five Drivelology categories (see §3.1). For each stage, we tested three prompting strategies: zero-shot, one-shot, and five-shot.

Text	Translated Text	Taggings
夜店這種地方還是少去，耳朵會聾掉。我陪朋友去過一次，後來男友叫我不要去，我都聽不見。	Nightclubs are the kind of place you should go to less, your ears will go deaf. I went once to accompany a friend, and later my boyfriend told me not to go, but I couldn't hear him.	wordplay
愛一個人是藏不住的，但愛兩個一定要藏住。	Loving someone cannot be hidden, but loving two people must be hidden.	switchbait
母親節已經想好要送什麼了。給自己買件新衣服，送媽媽一個漂亮的女兒。	Mother's Day gift is already decided. Buy myself a new dress and give my mom a beautiful daughter.	misdirection
同學：你都怎麼作弊？明天段考。我：偷偷的把課本的內容都記在腦袋裡，老師根本抓不到。	Classmate: How do you usually cheat? The midterm is tomorrow. Me: I secretly memorize all the contents of the textbook in my head, the teacher can't catch me at all.	inversion
只要夫妻两个人互相信任，四个人就能相安无事。	As long as the husband and wife trust each other, four people can get along in peace.	inversion, wordplay
以前我老婆對我眞的超兇的，後來我就讓他去學空手道跟劍道。至少現在他打我之前，會先跟我鞠躬。	In the past, my wife was really super mean to me. Later, I let her go learn karate and kendo. At least now, before she hits me, she will bow to me first.	inversion, switchbait
高速公路旁的警语写着：开车请看前方。	The warning sign by the highway reads: Please keep your eyes on the road while driving.	inversion, paradox
女孩从不会在意你开什么颜色的法拉利。	A girl will never care what color Ferrari you drive.	misdirection, inversion, wordplay
學生：老師，我媽要我問一下我的成績出來了嗎？老師：你等一下。學生：好的。老師：09 55 34 20 47。學生：打不通。老師：這是成績。	Student: Teacher, my mom asked me to check if my grades are out yet? Teacher: Just a moment. Student: Okay. Teacher: 09 55 34 20 47. Student: Can't get through. Teacher: That's your grade.	misdirection
我：今年過年我要帶女朋友回去喔。老媽：幾歲，哪裡人？我：到時候你們自己問他，他很溫柔可愛體貼，沒什麼缺點。老媽：你就是他最大的缺點。	Me: This year during Lunar New Year, I'm bringing my girlfriend home. Mom: How old is she? Where is she from? Me: You can ask her yourself then. She's gentle, cute, thoughtful—she doesn't have many flaws. Mom: You're her biggest flaw.	misdirection
私の長所は素直に間違いを認めることです。短所は、決して間違いを改めないことです。	My strength is that I can honestly admit my mistakes. My weakness is that I never correct my mistakes.	paradox
お客様のおかげで忍耐力がアップしてきました。	Thanks to the customer, my patience has improved.	inversion, wordplay
제가못하는것빼고는다잘해요	I'm good at everything except what I can't do.	paradox
A: 돌잔치결혼장례식등등한달전에얘기하셈연차올려야하니B: 한달전은빡세네장례식한달전예고면살인아님?	A: Let me know a month in advance for events like funerals. B: A month's notice for a funeral? That's premeditated murder!	paradox
여자친구:나살찐거같아?남자친구:넌살이문제가아니야 °	Girlfriend: Do you think I gained weight? Boyfriend: Your problem isn't your weight.	inversion
집에불이났다. 온가족이당황해서소리친다. 아버지: 야, 119가몇번이야? 119가몇번이냐고!!!아들: 아버지, 이럴때일수록침착하셔야돼요. 제가114에전화해서물어볼게요	The house caught fire. The whole family was panicking and shouting. Father: Hey, what's the number for 119? What's the number for 119!!! Son: Dad, you need to stay calm in situations like this. I'll call 114 and ask.	misdirection, switchbait
Quel est le coquillage le plus léger? La palourde.	What is the lightest shell? The clam.	wordplay
Une vague amoureuse du vent lui demande : Est-ce que tu peux me faire une petite bise aujourd'hui?	A wave, in love with the wind, asks: "Can you give me a little kiss today?"	wordplay
Que horrible cuando tu mamá te da instrucciones y tú estás medio dormida, entonces no te acuerdas si tenías que lavar la basura, colgar al perro o sacar a pasear la ropa.	How awful when your mom gives you instructions while you're half asleep, so you can't remember whether you were supposed to wash the trash, walk the dog, or take the clothes out for a walk.	misdirection
C'est l'histoire de deux pommes de terre. Une d'elles se fait écraser et l'autre s'écrie: Oh purée !	It's the story of two potatoes. One of them gets crushed, and the other exclaims: Oh mashed potatoes!	wordplay
Pourquoi est-ce que Hulk a un beau jardin? Parce qu'il a la main verte.	Why does Hulk have a beautiful garden? Because he has a green thumb.	switchbait, wordplay
Que fait un employé de chez Sephora à sa pause clope ? Il parfumer.	What does a Sephora employee do during their cigarette break? They perfume.	switchbait, wordplay
Qu'est ce qu'une lampe moche ? Un LEDron.	What do you call an ugly lamp? A LED-boring.	wordplay
Pourquoi est ce que Potter est triste ? Parce que personne Harry à sa blague.	Why is Potter sad? Because no one Harry gets his joke.	switchbait, wordplay

Table 3: Representative examples of Drivelology.

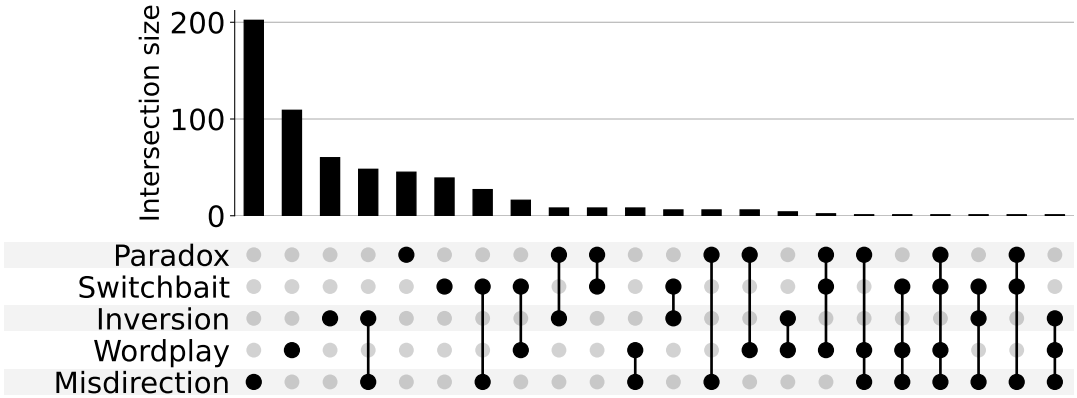

Figure 5: UpSet plot (Lex et al., 2014) illustrating the overlap and intersection sizes among Drivelology categories. Each vertical bar represents the number of samples belonging to a specific combination of categories, as indicated by the connected black dots below. Categories include Misdirection, Paradox, Switchbait, Inversion, and Wordplay.

Language	Drivelology	Non-Drivelology	Total
Mandarin	277	194	471
English	93	75	168
Spanish	69	68	137
French	62	80	142
Korean	52	92	144
Japanese	47	91	138
Total	600	600	1200

Table 4: Language distribution of Drivelology and non-Drivelology samples in the DRIVELHUB. Mandarin includes Simplified Chinese and Traditional Chinese.

C.1 Findings

Our investigation into GPT-4's generative capabilities reveals a significant gap between mimicking linguistic forms and achieving genuine pragmatic depth.

Minimal Guidance. When prompted with only a brief description of Drivelology, GPT-4 relied heavily on surface-level cues such as *paradoxical*, *unexpected twist*, or *nonsensical*. The resulting outputs were typically simple, declarative statements containing superficial contradictions (e.g., "He's an honest liar" or "I bought a one-way ticket with unlimited uses"). These examples mimicked the form of Drivelology but lacked the semantic depth, layered meaning, and interpretive tension that define the genre.

With Category Definitions. Providing explicit category definitions led to more complex outputs, including richer character interactions, emotional cues, and linguistic characteristics like personification. For example: "He says he's vegetarian, but only eats plants that scream – like carrots that wail when pulled from the ground." While this sentence demonstrates greater creativity and engagement with paradox and wordplay, it still falls short of Drivelology's essential qualities. The supposed contradiction is not inherently ironic, and the interpretive tension remains weak. Additionally, increasing the number of examples (five-shot) did not improve output quality. Instead, it often exposed deeper structural and semantic issues. Some outputs suffered from syntactic misalignment (e.g., "It's not that you can't love others, it's that love can't you," which is ungrammatical and uninterpretable), while others exhibited shallow or logically incompatible contradictions (e.g., "It's not that I don't want to work hard, it's that I've worked so hard it looks like I'm not trying").

Across all settings, GPT-4 struggled to internalise the subtle requirements of Drivelology. Out of 20 generations prompted with examples, only one output achieved comprehensive alignment between topic, rhetorical category, and sentence structure. For example: "这本书太深奥了，我花了一整晚没看懂封面。(This book is too profound, I spent the whole night and still couldn't understand the cover)." Although providing more examples led to slightly more complex narratives, the outputs consistently lacked Drivelology's hallmark features: contextual misdirection, interpretive layering, and rhetorical paradox. These shortcomings were especially pronounced in scenarios requiring cultural knowledge, emotional nuance, and inferential reasoning. Overall, our findings highlight the persistent challenges LLMs face in generating text that aligns with the deeper pragmatic and rhetorical demands of Drivelology.

Prompt	Model Size	Narrative		MCQA		Classification	
		BERT	GPT	Easy	Hard	Detect	Tag
English	4B	77.45	2.64	81.00	6.00	66.80	43.21
	8B	85.91	2.64	83.17	26.78	65.00	38.04
	14B	86.00	2.67	83.96	46.66	73.57	45.19
Mandarin	4B	67.79	2.96	77.61	2.44	62.86	46.10
	8B	65.07	3.08	81.11	19.11	78.81	41.71
	14B	64.23	3.19	82.56	51.69	77.62	49.35

Table 5: Performance of Qwen3 models of varying sizes (4B, 8B, 14B) across different tasks.

D Future Work

While this work successfully introduces the DRIVELHUB dataset and benchmarks the limitations of current LLMs, the rich structure of our data opens up significant avenues for future research. We outline two key directions: advancing model training methodologies and developing a robust framework for evaluating Drivelology generation.

D.1 Advancing Model Training with the MCQA Task

We have identified that the Narrative Selection (MCQA) task within our dataset is a perfect fit for GRPO (Shao et al., 2024). GRPO is an advanced preference optimisation technique that allows a model to learn from the relative preferences within a group of candidate responses, rather than relying on simple pairwise (Rafailov et al., 2023) or scalar rewards (Schulman et al., 2017). By learning from a full ranking of multiple candidates, the model receives a much richer training signal. The design of our MCQA task naturally lends itself to this paradigm. For each Drivelology sample, we provide one correct narrative and several carefully crafted distractors. This setup creates an inherent group-wise ranking (i.e., the correct option is preferred over all incorrect options), which can be directly leveraged by GRPO. Future work should explore fine-tuning LLMs using GRPO on the DRIVELHUB MCQA data. We hypothesise that this approach could substantially improve a model's ability to discern subtle semantic and pragmatic distinctions, thereby enhancing its capacity for the deep, non-linear reasoning required to truly comprehend Drivelology. This would represent a significant step toward closing the gap between statistical fluency and genuine cognitive understanding that our current work highlights.

D.2 Developing Metrics for Drivelology Generation

Our current study focuses primarily on the understanding and reasoning abilities of LLMs rather than their capacity for generation. A significant area for future work is to establish a comprehensive framework for evaluating generated Drivelology. A key limitation to address is the absence of metrics capable of quantifying the qualities that make a text Drivelological. Simply measuring fluency or grammatical correctness is insufficient. A robust evaluation would require developing novel metrics to assess specific aspects of a generated Drivelology text, such as: (1) **Entertainability**: How humorous, witty, or engaging is the output? (2) **Cohesion and Paradoxical Depth**: How well does the output maintain surface-level coherence while simultaneously embedding a meaningful, non-obvious contradiction or twist? (3) **Originality**: How surprising or non-formulaic is the output? Does it avoid simply rehashing common Drivelological text or existing templates? (4) **Cultural Resonance**: How well does the output tap into shared cultural knowledge, social scripts, or contemporary memes to create its meaning?

Furthermore, a robust framework must test for *controllable generation* – the ability to create Drivelology that meets specific constraints, like producing an "inversion" about "technology." Success here would be a strong signal of true comprehension. While developing this framework is challenging, it is essential for two reasons: it would allow for more rigorous model assessment and provide clearer targets for training. This creates a powerful feedback loop where better evaluation drives better generation, which in turn deepens the model's core reasoning, ultimately leading to LLMs that can truly master "nonsense with depth."

Human Guidelines:

Annotation Guidelines for Drivelology Dataset

Introduction

These guidelines are designed to assist annotators in accurately labelling samples for the Drivelology dataset. Annotators should familiarise themselves with the definitions and characteristics of Drivelology and non-Drivelology texts before proceeding.

Definitions

- Drivelology:
  - Description: Statements that appear logically coherent but contain deeper, often paradoxical meanings. These challenge conventional interpretation by blending surface-level nonsense with underlying depth, often incorporating elements of humour, irony, or sarcasm. Understanding Drivelology requires contextual insight and emotional interpretation.
  - Examples:
    * "I bought a book on how to solve 50% of my problems, so I bought two books."
    * "Loving someone cannot be hidden, but loving two people must be hidden."
- non-Drivelology:
  - Description: This includes pure nonsense (grammatically correct but semantically meaningless statements) and normal sentences, including quotes or proverbs, that convey clear or straightforward information without the layered complexity characteristic of Drivelology.
  - Examples:
    * "The cat sat on the mat." (normal sentence)
    * "Colourless green ideas sleep furiously." (pure nonsense)

Annotation Tasks

- Drivelology Tagging
  - Task: Classify Drivelology samples into one or more categories only if the sample is Drivelology:
    * Misdirection: A rhetorical technique where the focus shifts but connects back to the original topic through indirect hints.
    * Paradox: A statement that combines ideas that do not logically fit together but conveys a deeper meaning.
    * Switchbait: A language trick that changes meaning based on cultural knowledge or idioms.
    * Inversion: Rearranging the usual order of words or ideas to create a surprising effect.
    * Wordplay: Creative use of language through puns or double meanings.
  - Instructions:
    * Identify the primary characteristics (i.e., the first strong impression) of the text.
    * Assign one or more categories based on the definitions above.
- Implicit Narrative Writing
  - Task: Generate a detailed description illustrating the implicit narrative of the Drivelology text.
  - Instructions:
    * Analyse the text to uncover underlying themes, messages, or emotional undertones.
    * Write a narrative that reflects the deeper significance of the text, going beyond a surface-level summary.
    * Generate four contextualised, plausible, but ultimately incorrect narrative, wrong understanding of the given Drivelology text, each within three sentences as distractors. Keep the length and style the same as the correct narrative, and keep these negative narratives difficult to tell from the positive narrative.

Quality Assurance

Each annotation will undergo a review process where a meta-reviewer will assess the annotations for consistency and accuracy. Annotators should mark down any samples that exhibit ambiguities or uncertainties during the annotation process. The meta-reviewer will review these marked samples and finalise the answer based on a thorough evaluation.

Figure 6: Guidelines for human annotators.

Prompt1:

Instruction:
Classify whether the given text is a Drivelology sample or not.

Definition:
- Drivelology: Statements that appear logically coherent but contain deeper, often paradoxical meanings. These challenge conventional interpretation by blending surface-level nonsense with underlying depth, often incorporating elements of humor, irony, or sarcasm, and requiring contextual understanding and emotional insight to unravel their true significance.
- non-Drivelology: This includes pure nonsense (grammatically correct but semantically meaningless statements, such as "boys will be boys") and normal sentences, including quotes or proverbs, that convey clear or straightforward information without the layered complexity characteristic of Drivelology.

Input Text:
{text}

Output Format:
Please provide the output in JSON format with the following keys:
- answer: Specify whether the text is "Drivelology" or "non-Drivelology."
- reason: Provide a clear explanation of why the text is classified as Drivelology or not.

Prompt2:

Instruction:
Classify whether the given text is a Drivelology sample or not.

Definitions:
- Drivelology: Statements that appear logically coherent but contain deeper, often paradoxical meanings. These challenge conventional interpretation by blending surface-level nonsense with underlying depth, often incorporating elements of humor, irony, or sarcasm, and requiring contextual understanding and emotional insight to unravel their true significance.
- non-Drivelology: This includes pure nonsense (grammatically correct but semantically meaningless statements) and normal sentences, including quotes or proverbs, that convey clear or straightforward information without the layered complexity characteristic of Drivelology.

Input Text:
{text}

Instructions for Reasoning:
Analyse the input text by comparing it to the definitions above. Identify whether it contains logical coherence, paradox, layered meaning, or requires emotional/contextual insight. If uncertain, select the category that best fits and explain your reasoning.

Output Format:
Please provide the output in JSON format with the following keys:
- answer: Specify "Drivelology" or "non-Drivelology."
- reason: Clearly explain why the text was classified as such, referencing specific features from the definitions.

Prompt3:

Classify the text as "Drivelology" or "non-Drivelology."

Definitions:
- Drivelology: Logically coherent statements with paradox, layered or hidden meaning, often using humor, irony, or sarcasm. These require emotional or contextual insight to interpret.
- non-Drivelology: Pure nonsense or straightforward statements without hidden complexity.

Text:
{text}

Reasoning:
Decide based on the definitions above. If uncertain, choose the closest fit and briefly explain.
Output (JSON only):

```
{
    "answer": "Drivelology",
    "reason": "The text contains underlying meaning, fitting the Drivelology definition."
}
```

Figure 7: Prompts for Drivelology Detection task.

Prompt1:

Instruction:
Classify the given text into one or more of the following categories: inversion, wordplay, switchbait, paradox, and misdirection.

Definitions:
- inversion: INVERSION DEFINITION.
- wordplay: WORDPLAY DEFINITION.
- switchbait: WITCHBAIT DEFINITION.
- paradox: PARADOX DEFINITION.
- misdirection: MISDIRECTION DEFINITION.

Input Text:
{text}

Output Format:
Please provide the output in JSON format with the following keys:
- answer: List the applicable comma-separated categories for the text (e.g., "category1, category2").
- reason: Provide a clear explanation for why the text is classified into each category.

Prompt2:

Instruction:
Analyse the input text and classify it into one or more of the following categories: inversion, wordplay, switchbait, paradox, and misdirection. Use the definitions below to guide your classification.

Definitions:
- inversion: INVERSION DEFINITION.
- wordplay: WORDPLAY DEFINITION.
- switchbait: WITCHBAIT DEFINITION.
- paradox: PARADOX DEFINITION.
- misdirection: MISDIRECTION DEFINITION.

Input Text:
{text}

Output Format (JSON):
```
{
    "answer": "category1, category2, ...",
    "reason": "Explain why the text fits each chosen category based on the definitions."
}
```

Prompt3:

Instruction:
Examine the input text and determine which of the following categories it belongs to: inversion, wordplay, switchbait, paradox, and misdirection. Base your classification strictly on the definitions provided below.

Definitions:
- inversion: INVERSION DEFINITION.
- wordplay: WORDPLAY DEFINITION.
- switchbait: WITCHBAIT DEFINITION.
- paradox: PARADOX DEFINITION.
- misdirection: MISDIRECTION DEFINITION.

Input Text:
{text}

Please provide the output in JSON format:
```
{
    "answer": "category1, category2, ...",
    "reason": "Briefly explain how the text fits each selected category, using the definitions as a basis."
}
```

Figure 8: Prompts for Drivelology Tagging task.

Prompt1:

You need to first read and understand the text given. Generate a detailed description to illustrate the implicit narrative of the text.

Text: {text}

Output format should be JSON with the following keys:
- narrative: The narrative of the text in English.

Prompt2:

Read and understand the provided text carefully.

Task: Generate a detailed description that illustrates the implicit narrative of the text.

Input Text: {text}

Output Format:
{

"narrative": "The narrative of the text."

}

Please ensure the output is in JSON format and contains the key "narrative" with the developed description of the implicit narrative derived from the input text.

Prompt3:

Read and understand the provided text carefully.

Task:
Generate a detailed description that illustrates the implicit narrative of the text. By "implicit narrative," we mean the underlying message, theme, perspective, or emotional undertone that is not directly stated but can be inferred from the text. Your description should go beyond surface-level summary and provide insights into the text's underlying themes, perspectives, or intentions.

Input Text:
text

Output Format:
Output only the JSON object, with no extra commentary or explanation.
{

"narrative": "A detailed description of the implicit narrative, including the underlying theme, emotional tone, and implied perspective."

}

Figure 9: Prompts for Narrative Writing task.

Prompt1:

Task:

Evaluate how accurately the candidate narrative matches the given reference narrative.

Use a scale from 1 to 5, where 1 indicates the least accuracy and 5 indicates the highest accuracy.

- Candidate Narrative: {candidate}

- Reference Narrative: {reference}

Output Format:

Please provide the output in JSON format with the following key:

- score: The score indicating the level of matching, ranging from 1 to 5.

Figure 10: Prompts for evaluating Narrative Writing task.

Prompt1:

Tell me the best option in the following options which represents the underlying narrative of the text?

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. {narrative_5}

Output format should be JSON with the following keys:
- answer: The option the text belongs to, and it should be uppercase among A, B, C, D, E.

Prompt2:

Tell me the best option from the list below that represents the underlying narrative of the text. By "underlying narrative," we mean the main theme, implicit message, or perspective the text conveys.

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. {narrative_5}

If more than one option seems plausible, pick the one that best represents the main narrative.

Output format should be JSON with the following keys:
- answer: The option should be a single uppercase letter among A, B, C, D, or E.

Output only the JSON object, with no extra commentary.

Prompt3:

Tell me which option best represents the underlying narrative (main theme, message, or perspective) of the text.

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. {narrative_5}

If more than one fits, pick the best.

Output only JSON:
- answer: One uppercase letter: A, B, C, D, or E.

Example:

```
{
   "answer": "B"
}
```

Figure 11: Prompts for Easy Narrative Selection task.

Prompt1:

Tell me the best option in the following options which represents the underlying narrative of the text?

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. None of the above.

Output format should be JSON with the following keys:
- answer: The option the text belongs to, and it should be uppercase among A, B, C, D, E.

Prompt2:

Tell me the best option from the list below that represents the underlying narrative of the text. By "underlying narrative," we mean the main theme, implicit message, or perspective the text conveys.

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. None of the above.

If none of the options fully fit, select "E. None of the above." If more than one option seems plausible, pick the one that best represents the main narrative.
Output format should be JSON with the following keys:
- answer: The option the text belongs to, and it should be a single uppercase letter among A, B, C, D, or E.

Output only the JSON object, with no extra commentary.

Example output:
{

"answer": "B"

}

Prompt3:

Tell me which option best represents the underlying narrative (main theme, message, or perspective) of the text.

Text: {text}

A. {narrative_1}
B. {narrative_2}
C. {narrative_3}
D. {narrative_4}
E. None of the above.

If none fit, choose E. If more than one fits, pick the best.

Output only JSON: - answer: One uppercase letter: A, B, C, D, or E.

Example:
{

"answer": "B"

}

Figure 12: Prompts for Hard Narrative Selection task.