MORABLES: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables

Matteo Marcuzzo ♠ **Alessandro Zangari** ♠ **Andrea Albarelli** ♠
Jose Camacho-Collados ◊ **Mohammad Taher Pilehvar** ◊
♠Dept of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice
{name.surname}@unive.it, albarelli@unive.it
◊School of Computer Science and Informatics, Cardiff University
camachocolladosj@cardiff.ac.uk, pilehvarmt@cardiff.ac.uk

Abstract

As LLMs excel on standard reading comprehension benchmarks, attention is shifting toward evaluating their capacity for complex abstract reasoning and inference. Literature-based benchmarks, with their rich narrative and moral depth, provide a compelling framework for evaluating such deeper comprehension skills. Here, we present MORABLES, a human-verified benchmark built from fables and short stories drawn from historical literature. The main task is structured as multiple-choice questions targeting moral inference, with carefully crafted distractors that challenge models to go beyond shallow, extractive question answering. To further stress-test model robustness, we introduce adversarial variants designed to surface LLM vulnerabilities and shortcuts due to issues such as data contamination. Our findings show that, while larger models outperform smaller ones, they remain susceptible to adversarial manipulation and often rely on superficial patterns rather than true moral reasoning. This brittleness results in significant self-contradiction, with the best models refuting their own answers in roughly 20% of cases depending on the framing of the moral choice. Interestingly, reasoning-enhanced models fail to bridge this gap, suggesting that scale – not reasoning ability – is the primary driver of performance.

1 Introduction

The evaluation of Large Language Models (LLMs) for natural language understanding remains a key challenge, driving the creation of benchmarks that reflect the depth and complexity of their evolving capabilities (Chang et al., 2024; Hodak et al., 2024; Dong et al., 2024; White et al., 2025). Traditional benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) have reached a point of performance saturation, and some of their tasks have been criticized for structural limitations. A prime example is Natural Language Inference (NLI), a core reading comprehension task featured heavily in these benchmarks. Despite its central role, studies have revealed significant shortcomings in how NLI datasets function as evaluation tools (Naik et al., 2018; Gururangan et al., 2018; Nie et al., 2020). For instance, McCoy et al. (2019) demonstrated that models trained and evaluated on MNLI (Williams et al., 2018) often rely on superficial heuristics rather than genuine understanding. While adversarial datasets have been introduced to mitigate the influence of spurious cues, they often lead to artificial examples that may not reflect real-world language use (Bihani and Rayz, 2025).

Recognizing these limitations, recent research has begun to explore alternative benchmarking approaches that better capture the richness of human language understanding (Ghosh and Srivastava, 2022; Sravanthi et al., 2024; Li et al., 2024). One promising direction involves literature-based benchmarks, which are designed to probe higher-order reasoning and comprehension using authentic narratives. Unlike many traditional tasks, benchmarks derived from literary texts present models with more natural linguistic structures and challenge them to interpret implicit themes that require deeper inferential skills (Kočiský et al., 2018).

Building on this line of work, we propose the use of moral fables drawn from literary tradition as a novel tool to assess moral inference in modern LLMs. We introduce MORABLES, a curated benchmark of 709 short stories and fables primarily drawn from Western literary tradition. Each entry includes a high-quality transcription or translation of an original fable, paired with a moral attributed to the original author or translator. Our proposal consists of a human-verified Multiple Choice Question Answering (MCQA) task where the original moral is presented alongside a diverse set of four alternative options (Figure 1 provides an example). We create multiple variants of our benchmark based on story and choice modifications to test for poten-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27727–27751
November 4-9, 2025 ©2025 Association for Computational Linguistics

The Wolf and the Crane (Perry 156)

A wolf swallowed a bone which got stuck in his throat. The pain was excruciating, so the wolf started looking for someone who could be induced to remove the accursed thing in exchange for a reward. The wolf asked each of the animals if they would help him and finally the crane was convinced by the wolf's solemn promises. Trusting her long beak to the wolf's gaping maw, the crane carried out the dangerous cure. Yet when the crane demanded the promised reward, the wolf simply said, "You ungrateful creature! You extracted your head unharmed from my mouth and still you ask for a reward?".

(A) There is nothing more valuable than liberty.
(B) **Expect no reward for serving the wicked.**
(C) Gratitude can turn pain into promise.
(D) Unity is mankind's greatest good, while ungrateful dissension is a brave and slavish thing.
(E) Desperation can turn foes into allies.

Figure 1: A sample entry from the MORABLES core MCQA dataset, with the correct option (B) in bold.

tial biases, including data contamination and shallow shortcut answering. Lastly, alongside MCQA evaluation, we assess the quality of LLM-generated free-text morals via human annotation.

Our findings show that, although larger models perform well, certain adversarial modifications and alternative evaluation procedures suggest that these models may rely more on thematic matching with story events than genuine moral reasoning. For instance, this weakness is illustrated by our discovery that even the best models contradict their own judgments in roughly 20% of cases when the moral inference task is reframed. Furthermore, our analysis of reasoning-augmented models indicates that explicit reasoning contributes less to performance than overall model size. Lastly, our experiments with human evaluation of free-text morals indicate that this task is highly challenging and potentially open ended, with semantic similarity to the original moral showing only a weak correspondence. This highlights moral inference as a complex and multi-faceted task.

2 Related Work

2.1 Fables, Stories and Morals

Moral Fables in NLP The only previous work explicitly addressing fables and morals in NLP is by Guan et al. (2022), who introduced a bilingual dataset of crowd-sourced stories and morals. However, their broader definition of a story – "a series of coherent events involving several interrelated characters, implying support or opposition of some behavior" – results in many entries that deviate from classical fables, including Wikipedia excerpts and forum anecdotes. Our contribution differs substantially, focusing solely on historically sourced fables which are manually verified.

Narrative and story datasets Several datasets have been proposed to investigate machine reasoning within the context of stories, specifically targeting the tasks of selecting and generating suitable story endings (Mostafazadeh et al., 2016; Guan et al., 2019). Notable examples include ROCStories (Mostafazadeh et al., 2016), WritingPrompts (Fan et al., 2018), roleplayerguild (Louis and Sutton, 2018), and PG-19 (Rae et al., 2020), but none are explicitly centered on moral narratives.

Morality and ethics datasets Previous work explored morality and ethics, though in ways fundamentally different from fable-based morals, primarily aiming to evaluate machine alignment with human ethical reasoning. Notable datasets include Moral Stories (Emelin et al., 2021), ETHICS (Hendrycks et al., 2021a), and Scruples (Lourie et al., 2021), each focusing on scenarios requiring morally appropriate actions or ethical judgments.

2.2 Machine Reading Comprehension

Most existing machine reading comprehension tasks are structured as question answering, with broad categorizations based on the types of questions and answers (Qiu et al., 2019; Rogers and Rumshisky, 2020; Rogers et al., 2023). *Span-based* tasks involve selecting a continuous span of text from a given context as the answer. Notable datasets in this category include SQuAD/SQuAD2.0 (Rajpurkar et al., 2016, 2018), TriviaQA (Joshi et al., 2017), MS Marco (Nguyen et al., 2017), NewsQA (Trischler et al., 2017), and HotpotQA (Yang et al., 2018), covering a wide range of domains and settings. *Cloze-style* tasks, on the other hand, involve filling in blanks within a sentence or passage, as seen in datasets like CNN/Daily Mail (Hermann et al., 2015) and WikiReading (Hewlett et al., 2016).

The most adopted format is the *MCQA* task, which offers a pre-defined set of options to choose from as answers to a passage. These choices are typically created by domain experts and are often designed to sway the respondent's decision. Prominent datasets include RACE (Lai et al., 2017), MMLU (Hendrycks et al., 2021b), ARC (Clark et al., 2018), and MMLU-pro (Wang et al., 2024), all of which are widely used in LLM evaluation. However, it is worth noting that this approach is vulnerable to annotation artifacts and shallow cues, leading models to perform well without true understanding (Rogers and Rumshisky, 2020).

Lastly, *free-form answer* tasks have also been explored in datasets like NarrativeQA (Kočiský et al., 2018) and DuReader (He et al., 2018). While metrics such as BLEU (Papineni et al., 2002) and, more recently, BERTScore (Zhang et al., 2020) have been used for evaluation, they still show limited alignment with human judgments of similarity (Leung et al., 2022; Herbold, 2024).

3 The MORABLES Benchmark

In this section, we describe the MORABLES benchmark, which is publicly available.[1]

3.1 Data Collection

MORABLES is sourced from a variety of open sources, detailed in Appendix A. Aesop, the most prominent Western fabulist, serves as the primary source and has influenced many later authors, such as La Fontaine, to compose fables in a similar style.

Our definition of fable follows that of Paul E. Jose and Krieg (2005), which defines it as characterized by three main aspects: *(i)* they are short, *(ii)* they feature talking animals with a metaphorical meaning (though some fables contain historical figures), and *(iii)* they involve morally significant actions and outcomes.

Collection procedure The data collection was carried out in the following steps. First, we identified suitable sources, defined as those that *(i)* contain both a fable and an associated moral, *(ii)* are derived from historical literature and official translations, and *(iii)* are in the open domain. Then, the stories and their morals were extracted from the relevant websites or books. Since multiple sources may report the same story, we conducted an in-depth duplicate removal process. This involves the utilization of similarity scores such as word Intersection over Union (IoU) and BERTScore (Zhang et al., 2020). For Aesop's fables, its Perry Index (Perry, 1952) is also used to check for duplicates.

At the end of this process, we obtained 709 pairs of fables and their corresponding morals, whose statistics are reported in Table 1. The fables are primarily short texts, typically composed of a few long sentences that detail conversations between characters. In contrast, the morals are expressed as single, concise statements. A more detailed analysis of the textual content, including common themes, characters, and readability, can be found in Appendix A.

Text Statistic	Value
Number of fable/moral pairs	709 pairs
Avg length: fables	133.4 words
Avg length: morals	11.6 words
Unique words: fables + morals	7,278 words
Avg. number of sentences: fables	5.6 sentences
Avg. sentence length: fables	25.0 words

Table 1: Statistics for the original fables and morals in MORABLES. Word and sentence counts are calculated using the `word_tokenize` and `sent_tokenize` functions from `NLTK` (Bird, 2006).

3.2 Core Dataset Construction

We develop a MCQA-based language understanding task based on the retrieved fables and their morals. The **MORABLES** dataset features carefully constructed answer choices, each designed to test the respondent's decision-making process.

MCQA distractors The MORABLES core dataset requires models to select the correct moral among five candidates. To create a challenging benchmark, we focus on two key objectives: *(i)* developing plausible negative alternatives, thereby reducing the effectiveness of superficial cues and shortcut learning, and *(ii)* incorporating human validation to ensure both dataset quality and that distractors are not overly plausible or misleading.

In the following paragraphs, we outline the procedure for generating challenging negative choices. We implement a systematic information extraction process to identify similar characters and entities, salient features and alternatives, and to incorporate distractor adjectives and newly generated morals. All extraction processes were performed using GPT-4o (prompts detailed in Appendix E), with outputs verified by human annotators for accuracy and appropriateness.

(1) Similar-character moral Our first distractor choice consists of a moral from another story in the dataset, featuring similar characters but with a different development. This approach aims to reveal potential biases or shortcut strategies based solely on the entities in the story. However, not

[1] `https://huggingface.co/datasets/cardiffnlp/Morables`

all fables contain characters that appear in other stories. To address this, we expand our search by generating plausible alternatives for each character (*e.g.*, frogs → toads, fox → jackal) and matching stories based on these substituted entities.

(2) Trait-injected moral Our second distractor is also derived from the moral of a different fable, but it is modified to further assess the models' reliance on superficial cues. Specifically, we extract prominent features or traits of the fable's characters and insert them into an incorrect moral. As with the previous option, we prioritize stories with similar characters to make the choice more compelling, while ensuring that the moral selected is distinctly different from the first option.

(3) Feature-based moral Our third distractor is an LLM-generated moral, where the model is given only the traits and characteristics of the fable's characters and must base the moral solely on those attributes. Although this moral may be structurally similar to authentic ones, it is necessarily detached from the narrative's events or overarching message, as it lacks access to the story's context.

(4) Partial-story moral The final distractor is also LLM-generated, but it is based on an excerpt from the story (specifically, the first 10% of sentences). This option is particularly challenging to craft, as the moral or the source story may sometimes be inferred from the excerpt, especially in shorter fables. To mitigate this risk, we structure the generation prompt to encourage creativity and a unique moral, providing a hand-crafted example as guidance.

Proof-checking The most important factor when creating negative answer choices is that they remain plausible but fundamentally incorrect. While some choices may be partially correct or overly generic, the original moral must always be the correct answer. To ensure this, we conduct both a semi-automatic similarity check and a human-driven validation of our benchmark. Each distractor is assessed for its similarity to the original one using standard similarity measures (word IoU and BERTScore). Morals that exhibit high similarity (> 0.5 IoU, > 0.4 BERTScore F1) are then manually evaluated – roughly 15% of feature-based morals and 25% of partial-story morals. The main guideline for evaluation is to accept a generated moral if it is coherent but fails to correctly capture the underlying meaning of a fable. We further conduct an in-depth human evaluation of both the dataset and the generated options, detailed in the next Section. When a suitable alternative could not be AI-generated, it was curated manually.

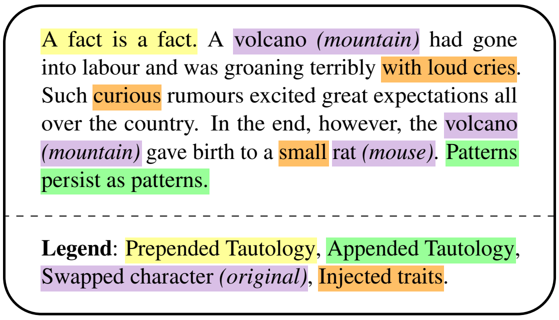
Finally, we examined additional options for the dataset, including the incorporation of opposite morals, as detailed in Appendix B. The appendix presents related experiments and our rationale for excluding these options from the final benchmark.

3.3 Human Annotation and Validation

It is essential to acknowledge that human oversight or annotator bias in dataset creation can result in overly challenging questions, which may lead to distorted metrics (Rubin and Donkin, 2024). In our study, this may occur due to: *(i)* alternative options appearing more suitable for the fable, *(ii)* options with meanings that are too similar, or *(iii)* the original moral reflecting outdated ethical values, making it less appealing. To address these issues, we implement a comprehensive human validation procedure with multiple annotators from diverse backgrounds. Annotation is conducted using Label Studio (Tkachenko et al., 2025), with a screenshot of the GUI provided in Appendix C.

Human validation procedure Five graduate-level English native speakers from diverse academic backgrounds (Computer Science, Neuroscience, Theology, and Literature) were selected as annotators after a screening where they annotated 20 fables of varying difficulty, from straightforward to intentionally ambiguous. Cohen's Kappa scores (McHugh, 2012) indicated substantial agreement (0.6–0.75). However, some difficult fables proved ambiguous even for the annotators, underscoring the need for human-driven benchmark refinement.

The dataset was divided into batches, with each fable receiving two independent annotations. Annotators were presented with each fable and five candidate morals, and were asked to select the most appropriate. To identify ambiguity we allowed annotators to choose multiple answers, instructing them to do so only if they found the options equally valid. Fables with two incorrect annotator answers were flagged as ambiguous and manually reviewed, generating alternative answers when appropriate. In total, 152 fables contained at least a moral or distractor option that needed to be replaced. Morals reflecting outdated values were replaced with updated versions coherent with the story, with new distractors created as needed.



A fact is a fact. A volcano *(mountain)* had gone into labour and was groaning terribly with loud cries. Such curious rumours excited great expectations all over the country. In the end, however, the volcano *(mountain)* gave birth to a small rat *(mouse)*. Patterns persist as patterns.

Legend: Prepended Tautology, Appended Tautology, Swapped character *(original)*, Injected traits.

Figure 2: An example with multiple adversarial modifications; colors indicate alteration types. Original characters (in parentheses) are not present in the actual text.

Variant	# of entries	Choice answers
Core	709	GT + distractors
NOTO	709	NOTO + distractors
TF	3,545	GT or distractor
ADV	709	GT + (adversarial) distractors

Table 2: Scale and answer choice composition for the four MORABLES dataset variants.

3.4 Dataset Variants: TF and NOTO

Following the work of Salido et al. (2025) and Wang et al. (2025), we also explore two alternative evaluation procedures: *(i)* substituting all the correct answers with “None of the other options” (NOTO version); and *(ii)* transforming the task into a binary one, where each choice is framed as a True or False question responding to the prompt “*True or False: The moral is ...* ” (TF version). We also conduct a short investigation into the impact of answer ID naming and token bias (Wei et al., 2024), detailed in Appendix D. Briefly, answer IDs had little impact on results, but token bias appeared in some open-weight models.

3.5 Adversarial MCQA

LLMs are trained on vast amounts of data, which likely includes classical fables such as those found in this dataset. Although some lesser-known fables may not have appeared during pre-training, contemporary research must assume that models have seen the majority of internet-available content. To address this, we introduce a set of adversarial modifications to the MORABLES dataset, designated as the ADV variant. Here, LLMs are challenged through subtle alterations to the fables or multiple-choice options, detailed below. Figure 2 illustrates an example of all possible fable modifications.

(1) Character swap The first modification swaps characters and entities within the fable with plausible alternatives. This approach tests whether models are simply associating specific characters with morals. The modification is implemented using basic string matching and replacement.

(2) Trait injection The injection of major traits and characteristics in fables follows the same process as trait-injected morals described in Section 3.2. We first select a different fable-moral pair and, when appropriate, inject character traits from the candidate fable into the story being modified. Not all adjectives are used, as some may cause incoherence. As before, we prioritize fables with similar characters, excluding those already used. The candidate pair’s moral is then included among negative answer choices.

(3) Tautology injection Our final modification prepends or appends short, self-contained sentences – specifically, tautologies (*e.g.*, “It is what it is”) – designed to add no additional meaning to the story. We begin with a manually compiled list of tautology-moral pairs, used as a seed to automatically generate a larger set (with some detailed in Appendix A). Each tautology is paired with a moral matching its message (*e.g.*, “Accept things as they are”). The list is then manually filtered to exclude sentences that could alter the story’s meaning.

Proof-checking Since these modifications introduce new morals, we verify that they are not semantically similar to the correct choice using the same procedure described earlier. We test both adding and substituting the new moral among the answer choices (limiting the total to five), finding that the number of choices does not significantly affect model performance (see Appendix D).

3.6 Summary of Dataset Variants

Table 2 summarizes the composition and scale of each variant of the MORABLES benchmark. The TF variant reframes the dataset into 3,545 distinct True/False statements (709 fables × 5 choices). Whether and how the distractors are modified in the ADV variant depends on the combination of modifications being tested (see Section 4.1.2).

4 Evaluations

We follow standard LLM evaluation for MCQA tasks (Zheng et al., 2024), prompting models to predict the answer ID token (*e.g.*, A, B, C). The first generated token, stripped of spaces and newlines, is used as the answer. For models with an accessible temperature parameter, we set it to 0 for deterministic results. While output probabilities for option tokens could be used in open models, we confirm this yields equivalent results (see Appendix D). Choice options are shuffled to give each token roughly equal chance of being correct.

LLMs evaluated We report results for several recent LLMs, grouped for clarity as: *(i)* small-open, *(ii)* large-open, *(iii)* large-closed, and *(iv)* reasoning. Among open models, we test Llama 3.3 (70B) (Grattafiori et al., 2024), Mixtral 8x22B (Mistral-AI, 2024), and DeepSeek V3 (2024-12-26) (DeepSeek-AI, 2024) as large-model representatives. To assess smaller models, we include Llama 3.1 (8B), Mistral 7B (Jiang et al., 2023), and Qwen 2.5 (7B) (Yang et al., 2025). For closed-source models, we select GPT-4o (2024-08-06) (OpenAI et al., 2024), Gemini 2.0 Flash (2025-02-05) (Anil et al., 2025) and Claude 3.5 Sonnet (2024-10-22) (Anthropic, 2024), owing to their strong performance and stability. We also evaluate two reasoning models: the closed-source GPT-o3-mini (2025-01-31) (OpenAI, 2025) and the open-weight DeepSeek R1 (2025-01-20) (DeepSeek-AI et al., 2025). While more advanced models such as OpenAI's o1 exist, their cost is prohibitive even for a benchmark of this size.

We present results using the best prompt identified in our preliminary tests, which provides simple guidelines and a single example (one-shot). Initial experiments showed that one example significantly improved performance, while additional examples had minimal effect. These results may thus be considered an upper bound, as zero-shot settings yield inferior performance (4–6% drop in accuracy). A comparison of zero- and one-shot results, along with all prompts used, is provided in Appendix E.

4.1 Results

Results for the MORABLES core dataset, shown in Table 3, reveal a pronounced disparity in performance between large and small LLMs. Larger open models – Llama 3.3 70B and DeepSeek V3 – consistently outperform their smaller counterparts, with a performance gap of over 40 percentage points between the worst small model, Mistral 7B (28.4%), and the best large model, Llama 3.3 70B (73.6%). Qwen 2.5 7B substantially outperforms other similarly-sized models, though a significant gap remains compared to the larger models.

	Model	**Accuracy %**
Small-open	Mistral 7b	28.4 ± 0.8
	Llama 3.1 8b	34.0 ± 1.5
	Qwen 2.5 7b	46.8 ± 0.3
Large-open	Mixtral-8x22b	56.9 ± 1.4
	Llama 3.3 70B	73.6 ± 0.6
	DeepSeek V3	70.6 ± 0.8
Large-closed	Gemini 2.0 Flash	80.7 ± 0.3
	GPT-4o	84.0 ± 0.5
	Claude 3.5 Sonnet	**84.8** ± **0.4**
Reasoning	GPT-o3-mini	66.3
	DeepSeek R1	77.0

Table 3: Average accuracy on MORABLES over 3 runs (± std). Reasoning models are tested only once due to their cost.

Among closed models, Gemini 2.0 Flash, Claude 3.5 Sonnet, and GPT-4o achieve the highest accuracies, demonstrating a clear advantage over the best open LLMs. Reasoning-oriented models show mixed results: DeepSeek R1 outperforms its non-reasoning counterpart, while GPT-o3-mini lags behind, likely due to its smaller scale.

Error Analysis Table 4 presents the choice distribution for our experiments on the MORABLES core dataset, showing the average selection percentage for each choice. The "partial-story" moral is the most frequent incorrect choice across nearly all models, indicating a primary failure mode where they over-rely on initial narrative cues rather than holistically comprehending the entire plot. For large models, the "feature-based" moral is the second most common error, revealing a notable performance gap between open and closed-source systems. This distractor accounted for 8.8% of total responses from Llama 3.3 70B, 11.5% from Mixtral-8x22b, and 10.0% from DeepSeek V3, considerably higher than those from closed-source counterparts like GPT-4o (4.9%) and Claude 3.5 (3.7%). Smaller models follow the same pattern but exhibit higher error rates across all distractor categories. For example, Llama 3.1 8b is uniquely vulnerable to the "similar-character" (12.9%) and "trait-injected" (13.2%) distractors, suggesting a broader difficulty discriminating among plausible-sounding but incorrect morals. Overall, this analysis reveals distinct failure modes: larger models are prone to nuanced thematic misinterpretations, while smaller models are misled by simpler, superficial cues.

Model	GT	Similar-char.	Trait-injected	Feature-based	Partial-story	Invalid
Mistral 7b	28.4 ± 0.8	4.0 ± 0.4	8.0 ± 0.3	20.4 ± 0.6	29.3 ± 0.7	9.8 ± 0.9
Llama 3.1 8b	34.0 ± 1.5	12.9 ± 0.8	13.2 ± 1.2	18.1 ± 0.5	21.7 ± 0.5	0.0
Qwen 2.5 7b	46.8 ± 0.3	2.6 ± 0.2	6.3 ± 0.2	16.8 ± 0.6	27.5 ± 1.0	0.0
Mixtral-8x22b	56.9 ± 1.4	3.3 ± 0.5	4.0 ± 0.4	11.5 ± 0.4	16.2 ± 0.6	8.0 ± 1.1
Llama 3.3 70B	73.6 ± 0.6	2.1 ± 0.3	1.5 ± 0.1	8.8 ± 0.4	13.1 ± 0.9	1.0 ± 0.1
DeepSeek V3	70.6 ± 0.8	2.5 ± 0.2	2.1 ± 0.2	10.0 ± 0.4	14.6 ± 0.7	0.0
Gemini 2.0 F.	80.7 ± 0.3	1.5 ± 0.1	2.8 ± 0.2	6.8 ± 0.1	8.2 ± 0.0	0.0
GPT-4o	84.0 ± 0.5	1.6 ± 0.1	1.3 ± 0.1	4.9 ± 0.1	8.1 ± 0.5	0.0
Claude 3.5 S.	84.8 ± 0.4	2.6 ± 0.3	2.2 ± 0.2	3.7 ± 0.2	6.8 ± 0.4	0.0
GPT-o3-mini	66.3	2.7	3.7	10.4	16.8	0.0
DeepSeek R1	77.0	2.3	2.1	5.6	10.4	0.0

Table 4: Average distribution of answer choices over 3 runs (%, ± std) for the core Morables dataset. Each column corresponds to a specific type of answer choice: The *GT* (ground truth) reflects accuracy and *Invalid* specifies cases in which the model's output was not in a valid format (after normalization).

Model	Accuracy %		Cons.
	TF	NOTO	
Mistral 7b	31.7 ± 0.4	6.5 ± 0.8	99.3
Llama 3.1 8b	63.4 ± 0.1	14.7 ± 1.8	55.3
Qwen 2.5 7b	74.1 ± 0.3	26.1 ± 1.3	73.4
Mixtral-8x22b	63.2 ± 0.4	3.1 ± 0.1	85.7
Llama 3.3 70B	77.9 ± 0.4	12.5 ± 0.4	70.1
DeepSeek V3	81.3 ± 0.1	18.4 ± 1.5	50.8
Gemini 2.0 F.	76.0 ± 0.1	29.3 ± 0.2	83.7
GPT-4o	81.3 ± 0.8	**30.1** **± 0.3**	78.6
Claude 3.5 S.	82.2 ± 0.2	17.3 ± 1.3	62.3
GPT-o3-mini	**83.3**	19.0	55.7
DeepSeek R1	76.1	25.5	81.9

Table 5: Average accuracy (± std) over 3 runs for the Morables TF and NOTO variants. The rightmost column (*Consistency*) measures internal consistency, *i.e.*, the percentage of incorrect NOTO choices that models previously labeled as True in the TF task.

4.1.1 TF/NOTO

Results for the TF and NOTO evaluation variants (detailed in Section 3.4) are presented in Table 5, with a granular breakdown of TF metrics in Table 6. The accuracy scores for TF (Table 5) are generally higher than or comparable to those achieved in the standard evaluation. However, the TF framing causes a significant class imbalance (20% positives, 80% negatives). Indeed, the results of Table 6 indicate models tend to over-predict the "*True*" class, thereby frequently misclassifying distractors. This pattern is reflected in consistently high recall (models identify the correct moral), but low precision (they mistakenly accept distractors).

Model	Prec %	Rec %	F1 %
Mistral 7b	22.3 ± 0.1	**96.9** **± 0.1**	36.2 ± 0.1
Llama 3.1 8b	32.1 ± 0.2	74.6 ± 1.1	44.9 ± 0.4
Qwen 2.5 7b	41.1 ± 0.4	68.2 ± 0.3	51.3 ± 0.4
Mixtral-8x22b	34.4 ± 0.2	92.5 ± 0.1	50.1 ± 0.3
Llama 3.3 70B	47.2 ± 0.4	89.7 ± 0.1	61.8 ± 0.4
DeepSeek V3	52.6 ± 0.4	67.8 ± 0.5	59.2 ± 0.1
Gemini 2.0 Flash	45.1 ± 0.1	92.5 ± 0.1	60.6 ± 0.1
GPT-4o	51.8 ± 1.3	91.6 ± 0.8	66.2 ± 0.8
Claude 3.5 Sonnet	53.4 ± 0.1	92.1 ± 0.4	**67.6** **± 0.1**
GPT-o3-mini	**55.8**	79.1	65.5
DeepSeek R1	45.1	89.8	60.1

Table 6: Average classification metrics for the binary Morables TF variant over 3 runs (± std).

The NOTO variant reveals that models are highly reluctant to select the "*None of the others*" option. We hypothesize two reasons for this: *(i)* the distractor options are sensible morals, even if incorrect for the given fable, and *(ii)* the inherent sycophancy of LLMs (*i.e.*, their tendency to agree with user suggestions), an artifact of RLHF finetuning (Sharma et al., 2024). To diagnose this behavior, we introduce a *Consistency* metric (rightmost column, Table 5). When a model incorrectly selects a moral instead of NOTO, we compare this choice against its own judgment from the TF setting. A choice is *Consistent* if the model previously labeled that moral as *True*, indicating internal consistency in selecting an answer it genuinely believes is valid. Conversely, a choice is *Inconsistent* if previously labeled *False*. This critical finding suggests the model does not believe its own answer is correct, but selects it anyway, revealing a strong aversion to the NOTO option. Therefore, high consistency suggests errors stem from a stable but flawed thematic interpretation, while low consistency supports the sycophancy hypothesis, as models would rather select a wrong answer than none.

The high consistency of Mistral 7b and Mixtral-8x22b reflects their poor performance, evidenced by the large precision-recall gap in Table 6. Llama 3.1 8b's low consistency is initially surprising; however, further analysis shows that it is considerably more likely than Mistral 7b to predict *False* (53% false predictions compared to only 12% for Mistral 7b). Notably, Mixtral-8x22b also shows a high rate of False predictions (46.3%), suggesting Llama 3.1 8b's inconsistency stems from a broader difficulty with moral inference, likely due to its small size.

Despite their larger sizes, DeepSeek V3 and, to a lesser extent, Llama 3.3 70B exhibit notable inconsistency. DeepSeek V3 often rejects a moral in the TF setting but then selects it around half the time in NOTO, while Llama 3.3 70B does so about 30% of the time. Remarkably, Claude 3.5 also demonstrates significant inconsistency, changing its stance in roughly 38% of cases. Surprisingly, GPT-o3-mini is also highly inconsistent, despite strong TF metrics. GPT-4o, DeepSeek R1, and Gemini 2.0 Flash show relatively stable performance across both settings, though a 20% rate of self-refuted answers remains a non-negligible gap.

4.1.2 Adversarial MCQA

Table 7 shows the performance of GPT-4o and Llama 3.3 70B on the MORABLES ADV variants across different modification combinations. In general, the modifications lead to a noticeable decline in performance, with their effects worsening when combined. Tautologies prove to be the most impactful modification, particularly when added at the end of the story.

Altering the characters within fables without further changes has minimal impact, with adjective injections yielding similarly modest effects (though slightly more pronounced for GPT-4o). Inserting tautologies is more impactful, causing a noticeable performance drop as models frequently select them as the answer. The two models also exhibit different positional biases: GPT-4o favors initial tokens, while Llama 3.3 favors final ones. As expected, combining multiple modifications leads to the sharpest decline, with accuracy dropping by up to 12.3% for GPT-4o and 10.5% for Llama 3.3.

Char.	Adj.	Pre.	App.	Acc %
GPT-4o (baseline accuracy: 84.0 ± 0.5)				
✓	-	-	-	82.2 ± 0.9
-	1.8%	-	-	81.0 ± 0.3
-	-	2.8%	-	79.8 ± 0.4
-	-	-	4.9%	80.0 ± 0.3
-	-	3.2%	4.6%	75.7 ± 0.1
✓	0.9%	-	-	78.6 ± 0.2
✓	-	4.7%	5.6%	72.9 ± 0.4
-	1.7%	3.8%	3.7%	74.7 ± 0.3
✓	1.1%	5.2%	4.9%	71.7 ± 0.1
Llama 3.3 70B (baseline accuracy: 73.6 ± 0.6)				
✓	-	-	-	72.1 ± 0.4
-	1.3%	-	-	72.5 ± 0.6
-	-	2.2%	-	73.6 ± 0.4
-	-	-	9.5%	68.9 ± 0.2
-	-	4.2%	8.0%	65.0 ± 1.6
✓	1.8%	-	-	72.3 ± 0.2
✓	-	3.6%	8.5%	63.9 ± 0.4
-	0.9%	4.1%	7.4%	65.7 ± 0.4
✓	0.6%	3.3%	8.4%	63.1 ± 0.8

Table 7: Results for adversarial (ADV) variants of the MORABLES dataset, showing the average selection rate of the adversarial moral and overall task accuracy (± std over 3 runs). Modification types are abbreviated as: *Char.* (character swapping), *Adj.* (adjective injection), and *Pre./App.* (tautology at start/end). Selection rate for *Char.* is omitted, as this modification does not affect distractor morals.

4.2 Free-Text Moral Evaluation

To further contribute to the evaluation of free-text generation, we assess moral generation from three different LLMs (GPT-4o, Claude 3.5, and Llama 3.3 70B). The same annotators involved in our dataset validation separately rated three generated morals per fable on a 1–5 Likert scale, according to the following alignment rubric:

(1) *No alignment*: The moral does not relate to the story's core message;

(2) *Minimal alignment*: The moral is related to the story but misses the main point;

(3) *Partial alignment*: The moral captures a part of the story's message;

(4) *Strong alignment*: The moral is a good fit but doesn't capture the full essence;

(5) *Perfect alignment*: The moral perfectly encapsulates the story's essence.

Table 8 summarizes our results, including the BERTScore (Zhang et al., 2020) F1 similarity between each generated moral and the reference for an approximate measure of semantic similarity.

Model	Rating	BERTScore
GPT-4o	3.83 ± 1.29	0.29 ± 0.24
Claude 3.5 S.	3.76 ± 1.29	0.24 ± 0.21
Llama 3.3 70B	3.66 ± 1.28	0.22 ± 0.19

Table 8: Average metrics for free-text morals generated by three different LLMs. Annotators rated each moral on a 1–5 scale based on its alignment with the corresponding fable. BERTScore (F1) is also reported.

All three models produce morals that, according to annotators, generally align with their respective fables. However, the high standard deviation suggests considerable variability, likely reflecting the inherent subjectivity of interpreting morals and the challenges of evaluating open-ended free-text outputs. Meanwhile, BERTScore F1 values remain fairly low for all models, indicating that the wording of generated morals often diverges substantially from the original references. A qualitative analysis of the divergence between BERTScore and human ratings is provided in Appendix C.

To explore the link between automated similarity measures and human judgment, we calculated the Pearson correlation coefficient between BERTScore and average annotator ratings. The correlation coefficients are positive albeit relatively low – 0.35 for Claude 3.5, 0.28 for GPT-4o, and 0.33 for Llama 3.3 – indicating only a weak positive relationship between semantic similarity and human perception of alignment (see the plot in Appendix C). This suggests that automated semantic similarity scores are unable to fully capture the complexity of the task, a finding that aligns with previous work on the limited agreement between such metrics and human judgment (Leung et al., 2022; Herbold, 2024)

4.3 Discussion

Our findings reveal several noteworthy patterns. First, smaller models struggle with the core moral inference task, emphasizing its non-trivial nature. In contrast, large open models such as Llama 3.3 70B perform well, although closed-source models still maintain a significant advantage.

Analysis of the `TF` and `NOTO` variants reveals that models frequently regard multiple options as plausible, often selecting the least incorrect option, echoing observations by Wang et al. (2025). This is reflected in the models' high recall but low precision in the `TF` setting (Table 6) and diminished accuracy in the `NOTO` scenario. Notably, models exhibit inconsistent behavior, sometimes preferring to choose a moral they previously rejected in `TF` over a "*None of the others*" option. This may reflect the propensity of LLMs to prioritize internal probability rankings over strict correctness, making "*None of the others*" rarely the preferred selection when juxtaposed with plausible alternatives (Wang et al., 2025). As noted by Salido et al. (2025), this may also signal strong memory-based associations.

Reasoning models yield mixed results: DeepSeek R1 almost matches closed-source systems, while GPT-o3-mini performs notably worse – both in terms of accuracy on the core dataset as well as in consistency between `TF` and `NOTO` – illustrating pronounced limitations. This suggests that model scale remains more important than reasoning capabilities for this task.

The impact of adversarial modifications may also indicate memorization, as evidenced by the significant performance drop in GPT-4o and Llama 3.3. Both models show a tendency to focus on the initial or final tokens of narratives, which heavily influences their decisions. This aligns with previous findings that neural language models often exhibit "attention sinks" at the beginnings and ends of documents (Xiao et al., 2024; Zhang et al., 2024).

5 Conclusion & Future Work

We introduce MORABLES, a high-quality, manually curated benchmark of fables paired with their morals. By building several challenging tasks, we probe the nuanced reasoning and narrative understanding capabilities of LLMs. Our results reveal a substantial performance gap between small and large models, with only the largest models achieving strong accuracy. However, even large models display inconsistent behavior under specialized evaluation, such as the `TF` and `NOTO` variants, and remain vulnerable to adversarial modifications that expose reliance on memorization and positional biases rather than genuine moral inference. These findings highlight important challenges for future research. Future work will expand the dataset to include fables from diverse cultures, facilitating the study of cultural bias and differing ethical perspectives. Moreover, we plan to leverage the holistic nature of moral lessons – which often arise from the broader narrative rather than specific textual elements – as a testbed for explainability research.

Limitations

The primary limitation of MORABLES lies in the high likelihood that most modern LLM have encountered not only the stories but also their associated morals during their pre-training phase. This raises significant concerns about memorization, which could compromise our ability to draw definitive conclusions regarding their moral inference capabilities, as the models may simply be recalling information they have already seen. While our adversarial modifications aim to partially address this issue, a more effective (but resource-intensive) solution would be to create entirely new, unseen short stories to test whether LLMs can still infer a moral in those cases.

Another limitation is that the dataset is predominantly sourced from Western stories. This is largely due to the greater availability of English fables that come with verified morals. Consequently, this focus restricts our ability to study potential cultural biases. Likewise, the dataset and our evaluation was performed for English only. Lastly, the rapid release of new models makes it challenging to publish results for models developed concurrently with ongoing experimentation. However, while future work should benchmark newer models, significant effort will be required to determine whether any observed performance improvements stem from enhanced reasoning capabilities or increased memorization tendencies.

Finally, we acknowledge a methodological consideration regarding GPT-4o's evaluation. Since the model's outputs were used to generate the distractor choices for our dataset, there is a risk of "self-preferential bias" in its results on the MCQA benchmark. However, we found no clear evidence of such an effect: GPT-4o's performance was not an outlier compared to its peers in a way that would indicate an unfair advantage. Furthermore, the risk was inherently limited because the model was not exposed to the original fable's text during the choice-generation process. Therefore, while the potential for bias is noted, its inclusion remains valuable for comparative analysis.

Ethics Statements and Broader Impact

MORABLES was collected from public web resources under licenses that allow use and redistribution for research purposes. The text and the moral of the fables has been extracted and formatted by the authors of this work. Human annotators were hired to verify the quality of the dataset and of AI-generated morals. Annotators were asked to read the fables and determine the correct moral from five multiple choices, as well as to rate automatically generated morals on a Likert scale from 1 to 5. Though all fables go through an initial screening process for offensive content, annotators were allowed to skip a fable if it contained content they felt uncomfortable with and encouraged to leave an explanatory comment. Annotators were paid around £20 per hour, well above the minimum hourly wage in the UK, where annotation took place. Annotators are recruited from diverse backgrounds, including Computer Science, Neuroscience, Theology, Literature.

As discussed in the previous section, our research focuses exclusively on fables from the Western tradition that have been translated into English. This choice is largely influenced by the prevalence of fable-moral pairs in English literature. However, fables are a universal phenomenon found in cultures around the world, and nearly all human societies utilize them. As a result, our dataset is unfortunately more than likely biased toward moral lessons that reflect Western values. This limitation may pose two significant issues: first, it fails to adequately represent values that differ across cultures, and second, models trained on this dataset may inadvertently reinforce these biases. Future work should aim to broaden this scope to include fables from diverse cultures globally. It would be particularly interesting to explore how models respond to fables with messages that diverge from conventional Western themes.

Acknowledgments

We thank our team of annotators – Felix Chadwick-Smith, Fayeja Farhana, Dhruvil Raval, Kayleigh Shier, and Hannah Trotman – for their careful and thoughtful work in validating our dataset and rating the model-generated morals. Our sincerest gratitude also goes to the Cardiff NLP group for their support throughout the project and for their detailed review of an early manuscript.

Jose Camacho-Collados was supported by a UKRI Future Leaders Fellowship.

References

Aesop. 2000. *Aesop's Fables*. Project Gutenberg.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, and Anja Hauth et al. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Anthropic. 2024. Claude 3.5 sonnet. Anthropic.

Geetanjali Bihani and Julia Rayz. 2025. *Learning Shortcuts: On the Misleading Promise of NLU in Language Models*, pages 147–158. Springer Nature Switzerland, Cham.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, and Qihao Zhu et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Laura Gibbs. 2017. Aesop's books. `https://aesopsbooks.blogspot.com/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *International Conference on Learning Representations*.

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

Steffen Herbold. 2024. Semantic similarity prediction is better than other semantic similarity measures. *Transactions on Machine Learning Research*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Miro Hodak, David Ellison, Chris Van Buren, Xiaotong Jiang, and Ajay Dholakia. 2024. Benchmarking large language models: Opportunities and challenges. In *Performance Evaluation and Benchmarking*, pages 77–89, Cham. Springer Nature Switzerland.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 56, Institute for Simulation and Training.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024. When LLMs meet cunning texts: A fallacy understanding benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Mistral-AI. 2024. Mixtral 8x22b: Cheaper, better, faster, stronger. Mistral AI.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2017. MS MARCO: A human-generated MAchine reading COmprehension dataset.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2025. Openai o3-mini. OpenAI.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and AJ Ostrow et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenRouter. 2025. OpenRouter - The API for all LLMs. `https://openrouter.ai`. Accessed: 2025-09-15.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Catherine A. D'Anna Paul E. Jose and Dana Balsink Krieg. 2005. Development of the comprehension and appreciation of fables. *Genetic, Social, and General Psychology Monographs*, 131(1):5–37. PMID: 16482782.

Ben Edwin Perry. 1952. *Aesopica: A Series of Texts Relating to Aesop or Ascribed to Him or Closely Connected with the Literary Tradition That Bears His Name*. University of Illinois Press.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A survey on neural machine reading comprehension. *Preprint*, arXiv:1906.03824.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Anna Rogers and Anna Rumshisky. 2020. A guide to the dataset explosion in QA, NLI, and commonsense reasoning. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 27–32, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Mark Rubin and Chris Donkin. 2024. Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology*, 37(8):2019–2047.

Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. 2025. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks. *Preprint*, arXiv:2502.12896.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages

353–355, Brussels, Belgium. Association for Computational Linguistics.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. LLMs may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862, Abu Dhabi, UAE. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chengyuan Li et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. 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*.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Dataset Sources and Feature Analysis

A.1 Data sources

Table 9 lists websites crawled for the dataset collection. All stories are free to use and redistribute for research purposes. Aesop's fables originated from an oral tradition, making it difficult to determine whether some fables attributed to other authors are direct transcriptions or new works inspired by him. Notable sources of such fables in this study include the Roman fabulist Phaedrus and the Italian writer Laurentius Abstemius. Whenever possible, we utilize English translation by G. F. Townsend (Aesop, 2000) and L. Gibbs (Gibbs, 2017)[2].

All morals used in our benchmark are directly sourced from either the original authors or reputable later authors/translators who provided interpretations of these fables:

- *Original Authors/Translators*: Some morals were explicitly added by the original authors of the fables (*e.g.*, Aesop) and subsequently translated.
- *Later Authors/Fabulists*: In many cases, morals were appended by later authors or fabulists who re-interpreted and translated the fables. For example, for a fable like "The Wolf and the Raven" (Perry Index 190), Aesop did not originally provide a moral. However, Sir Roger L'Estrange later added an "epimythium" (a moral appended at the end of the story) in his 1692 translation.

A.2 Analysis of Dataset Features

In addition to standard metrics reported in Section 3, we measure the Flesch-Kincaid readability score (Kincaid et al., 1975). The Fables average a score of 8.30 (with a median of 7.90), indicating that the text is suitable for someone at the 8th-grade reading level. This means that a typical 13- to 14-year-old should be able to read and understand the text with relative ease. A subset of our dataset, attributable to fables by Phaedrus, have a slightly more complex language, with a mean score of 11.59.

The most common characters in the fables are animals, with the fox (70 occurrences), lion (62), dog (61), wolf (52), donkey (36), and eagle (29) being the most frequent. Men are prominently portrayed, often interacting with animals, appearing at least 183 times as "man" and many other times in forms such as farmer (26) and shepherd (22). Major deities from Greek and Roman traditions, such as Zeus and Jupiter, also appear prominently, each featured in 20 fables.

We perform an exploratory topic analysis of the dataset with BERTopic (Grootendorst, 2022), specifically focusing on the content of the morals. An unguided clustering reveals numerous categories which can be loosely categorized as discussions of values and ethics (trust and deception, justice, gratitude, kindness), mindfulness and awareness (life choices and their impact), appreciation (contentment and value, true worth), protection (caution against enemies, suffering and neglect), and facing the human experience with resilience (fate and consequences, courage in adversity).

A.3 Tautologies

Table 10 presents examples of tautologies utilized for the ADV variant of MORABLES. The full set of tautologies is provided in the linked repository.

B Opposite morals and binary task

We considered other distractors for the core dataset, such as semantically opposite generated morals with high word overlap and random sentences from the fable to check for selection biases. However, these options were almost never chosen by the models in preliminary tests, indicating they were too easy and leading to their exclusion. Moreover, including a semantically opposite negative option might have inadvertently provided a superficial cue to the correct answer by allowing models (or even human evaluators) to simply identify the contrasting pair, thereby reducing the choice to just those two options. For completeness, we also implement a binary task where models select between the original and the opposite moral, which we detail below.

Preliminary binary task Following the work of Guan et al. (2022), we devise a preliminary task consisting of a binary classification between two moral choices. The task is to differentiate between the original and its semantic opposite. While Guan et al. (2022) create the incorrect candidate by substituting a random token in the original moral with its antonym, we follow a different approach. Leveraging the strength of LLMs, we prompt GPT-4o to provide "anti-morals" for each provided moral, stressing the fact that the provided answer should be semantically opposite in a binary sense and share as many words as possible as the original moral. The morals are collected in a 3-shot sce-

[2] The author kindly granted permission to use the content published under a Creative Commons license.

Source name	Author / Translator	Source link
The Aesop for Children	Aesop	Project Gutenberg
Aesop Fables	Aesop / G. F. Townsend	Project Gutenberg
The fables of Aesop	Aesop / Croxall, La Fontaine, L'Estrange	Hathi trust
The Comedies of Terence [...]	Terence, Phaedrus / H. T. Riley	Project Gutenberg
Aesop's Books	Aesop / L. Gibbs	Author Blog*

Table 9: List of source web pages used for constructing MORABLES. * Permission to utilize the translations and a parsable XML of the website has been kindly granted by the author.

Tautology	**Moral**
It is what it is.	Accept things as they are.
What is, is.	Embrace the present reality.
The outcome is the outcome.	Accept what results from your endeavors.
Things are what they are.	Accept the finality of reality without resistance.

Table 10: Examples of tautologies and paired morals utilized for the ADV version of MORABLES.

Model	**Acc %**	**Acc$_\emptyset$ %**	**Acc$_{1S}$ %**
Llama 3.3 70B	94.8	90.0	87.7
Gemini 2.0 Flash	96.9	94.2	92.8
GPT-4o	97.6	89.8	95.3
GPT-o3-mini	95.5	-	-
DeepSeek V3	97.6	91.7	90.8
DeepSeek R1	92.7	91.3	89.8

Table 11: Accuracy for the MORABLES Binary preliminary task. *Acc* stands for accuracy in the standard setting, while $Acc_\emptyset$ and Acc_{1S} stand for accuracy when the model is provided no fable or just the first sentence of the fable, respectively.

nario, where examples are hand-crafted. The resulting opposite morals are semantically and linguistically consistent and coherent. Nevertheless, the expectation remains for this binary task to be easy, as morals usually encapsulate sensible and just ethical values. Moreover, morals in themselves are often popular proverbs.

Results of preliminary tests Results for the binary task are presented in Table 11. As expected, opposite morals can be easily distinguished, notably even without providing the fables to the models. This is demonstrated by $Acc_\emptyset$, which measures model accuracy when determining the correct choice without any story text. We also report Acc_{1S}, where only the first sentence of the fable is provided. Interestingly, performance slightly decreases in this case, suggesting that models place considerable emphasis on the task's initial tokens and not just the choice answers.

C User study

C.1 Discrepancies between BERTScore and Human Ratings

Figure 4 shows the weak positive link between semantic similarity (BERTScore) and average human ratings, indicating that greater semantic overlap does not strongly predict human preference in moral evaluation.

A low BERTScore with a high human rating often occurs when the original moral is abstract and the model presents instead a more explicit lesson. Occasionally, however, this may result in the generated moral becoming overly focused on story details. For instance, the moral "*Straws show how the wind blows*" (Perry Index 95) is an abstract lesson on how small details can reveal important clues about future events or the character of an individual. In this case, the generated moral "*The way you are regarded by strangers will reflect how those closer to you perceive you*" fits the narrative of the story, but fails to encapsulate its broader meaning e.

Conversely, a high BERTScore does not always guarantee a high human rating. For instance, Figure 3 illustrates a case where all model-generated morals are semantically similar and aligned with the original moral, thus being correct. However, users favored more detailed, discursive versions like Claude's. We hypothesize this discrepancy signals an annotation artifact, where morals are evaluated comparatively rather than independently, despite instructions to assess each separately.

The Two Roosters And The Eagle (Perry 281)

Two roosters were fighting with one another. The loser hid himself away in a corner, while the rooster who had won the battle flew up on top of the house and flapped his wings, crowing about his victory. An eagle then swooped down and carried the rooster away.

- **True Moral**: Pride goes before destruction.
- **GPT**: Pride comes before a fall.
- **Claude**: The proud are often brought low.
- **Llama**: Pride goes before a fall.

Figure 3: Example of a Fable with its original moral and three LLM-generated morals. Despite all morals being semantically similar and correct, users often preferred more discursive and elaborated morals, such as the one generated by Claude.

C.2 Annotation Graphical User Interface

Figure 5 showcases the two components of the annotation user interface used in this study: *(a)* the MCQA component, in which annotators select the appropriate moral(s) for a given fable, and *(b)* the moral evaluation component, where annotators rate the alignment of each moral with the fable. The referenced annotator guidelines are provided in the linked repository.

D Empirical results on output tokens

Below, we present empirical results on several relevant phenomena identified in the literature that may skew LLM MCQA evaluation outcomes. While not exhaustive, these experiments provide insights and highlight areas for future investigation.

D.1 Output selection

We examine how model performance changes when using the first generated token as the answer (Acc$_{NT}$) compared to using output probabilities for option ID tokens (Acc$_{LP}$), as shown in Table 12. The latter method is applicable to open, local models and can be partially implemented via OpenAI's API for GPT-4o, which provides log probabilities for the 20 most likely tokens; in practice, answer labels are consistently among these top 20. However, this functionality is not supported by other models such as Gemini Flash 2.0, Claude 3.5, and reasoning models (GPT-o3-mini and DeepSeek R1).

We conducted three runs on a reduced sample of 100 data points to assess whether the choice of evaluation procedure significantly affects the results. This analysis is relevant only when the model temperature is set above 0, as otherwise models yield identical outputs regardless of the evaluation method. Accordingly, we set $T = 1$ for our evaluation. For each run, we calculate accuracy by comparing the next generated token as the answer to the top answer based on log probabilities within the same run.

Model	**Acc$_{NT}$ %**	**Acc$_{LP}$ %**
Llama 3.3 70B	70.0 $\pm$ 0.6	71.0 $\pm$ 1.1
GPT-4o	83.0 $\pm$ 1.7	81.5 $\pm$ 2.1

Table 12: Accuracy on a reduced sample (100 data points) of the MORABLES dataset. Acc$_{NT}$ stands for accuracy utilizing the standard next token prediction evaluation, while Acc$_{LP}$ indicates the accuracy for the same run evaluated using log probabilities ($\pm$ std).

We evaluate one open model (Llama 3.3) and one closed model (GPT-4o). As shown in Table 12, the confidence intervals for the two approaches overlap, indicating that both are valid alternatives without a statistically significant advantage for either model.

D.2 Token bias

We investigate the presence of token bias in LLMs for our benchmark in a selection of models (Zheng et al., 2024; Wei et al., 2024). Results are summarized in Table 13. Overall, the token bias phenomenon manifests with varying magnitudes across different models, with advanced models like GPT-4o showing little to no evidence of token bias, while models such as Llama 3.3 and DeepSeek V3 (and, to a lesser extent, Gemini 2.0 Flash) exhibit potential token bias in their selection of certain choices over others. Specifically, Llama 3.3 appears to favor the second option, whereas DeepSeek V3 tends to select the last option less frequently. This behavior occurs regardless of the answer ID naming, which we discuss next.

D.3 Token naming

We also evaluate whether models have different performance based on whether the answer choices are named with letters ("A", "B", *etc.*) or numbers ("1", "2", *etc.*). Results are summarized in Table 15. Similarly to the token bias phenomenon, the results depend on the model, with GPT-4o showcasing basically no difference, while models like Llama 3.3 and even Gemini 2.0 Flash display a small difference in performance. Interestingly, Gemini performs better on letter-named choices, while Llama performs worse. In general, there appears to be no best approach to choice naming in our case.

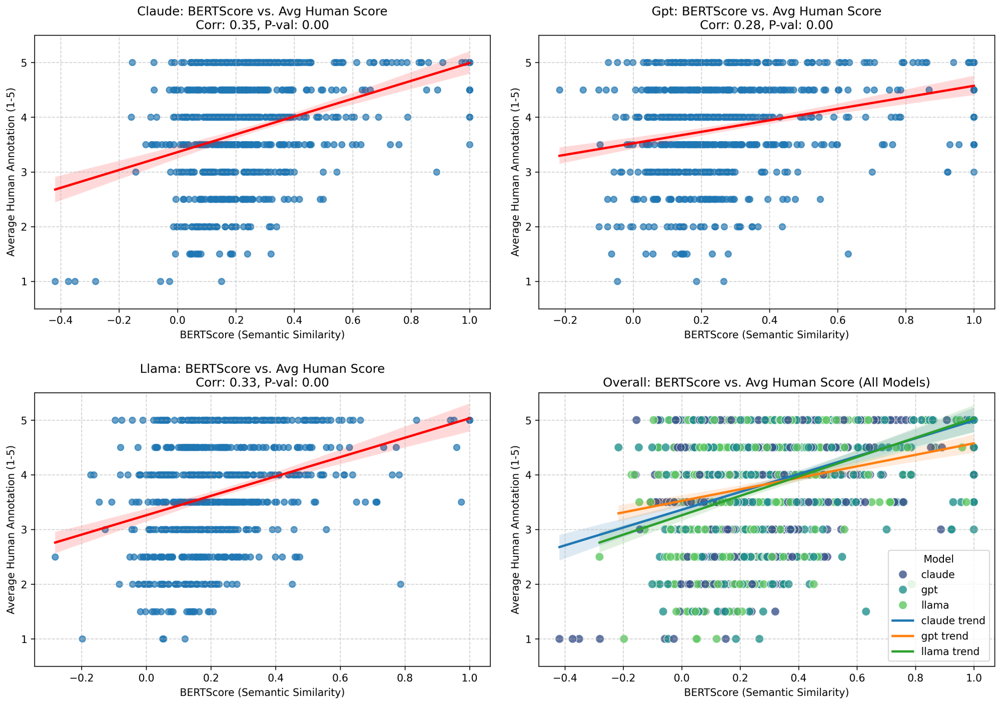


Figure 4: Average Human Rating vs. BERTScore across three LLMs. All models exhibit only a weak positive correlation, highlighting both the difficulty of evaluating morals using semantic similarity metrics and the inherent subjectivity in human moral assessment.

Model	O_1	O_2	O_3	O_4	O_5
Core dataset GT distribution	20.17%	21.16%	19.46%	19.04%	20.17%
Llama 3.3 70B	**16.92%**	**25.53%**	22.43%	17.21%	17.91%
GPT-4o	20.59%	19.46%	20.59%	19.18%	20.17%
Gemini 2.0 Flash	20.29%	18.84%	22.90%	19.13%	18.84%
DeepSeek V3	22.28%	22.14%	21.02%	20.31%	**14.25%**

Table 13: Answer selection percentages on a sample (50%) of the MORABLES dataset. Ground Truth distributions are shown in the first row. Percentages that vary significantly are highlighted in bold.

D.4 Number of choices

Table 14 compares model performance when presented with more than five answer choices (specifically, eight, due to the inclusion of adversarial options) versus when limited to five choices, in which less effective distractors were removed. The reduction to five choices was achieved by eliminating distractors that the model had not previously selected as correct answers. The results indicate that performance remains largely consistent across both settings, with a slight improvement in the five-choice condition, particularly for GPT-4o. This suggests that while the number of choices may slightly impact performance, the nature and quality of the distractors play a more important role.

Model	**Acc$_8$ %**	**Acc$_5$ %**
Llama 3.3 70B	63.1 $\pm$ 0.8	63.9 $\pm$ 0.5
GPT-4o	71.7 $\pm$ 0.1	73.1 $\pm$ 0.7

Table 14: Accuracy on the most challenging ADV variant of the MORABLES dataset, which incorporates all modifications. The table compares performance on more than five choices (left) versus five (right), where the latter set excludes the least plausible distractors.

E Experimental Setup and Prompts

E.1 Implementation Details

To ensure consistent and reproducible access, all models were queried via the OpenRouter API (OpenRouter, 2025).

Model	Acc$_{123}$ %	Acc$_{ABC}$ %
Llama 3.3 70B	73.0 ± 0.6	70.0 ± 1.1
GPT-4o	82.5 ± 2.1	81.5 ± 0.7
Gemini 2.0 Flash	74.0 ± 1.1	77.0 ± 0.6

Table 15: Accuracy comparison on a sample (100 data points) of the MORABLES core dataset, showing performance when token IDs are represented as numbers versus as letters (± std).

E.2 Zero-shot vs Few-shot

Table 16 displays the comparison between 0- and 1-shot setting performance for the MORABLES dataset. Smaller models are excluded from this analysis, as they do not adhere to the format instructions without providing examples.

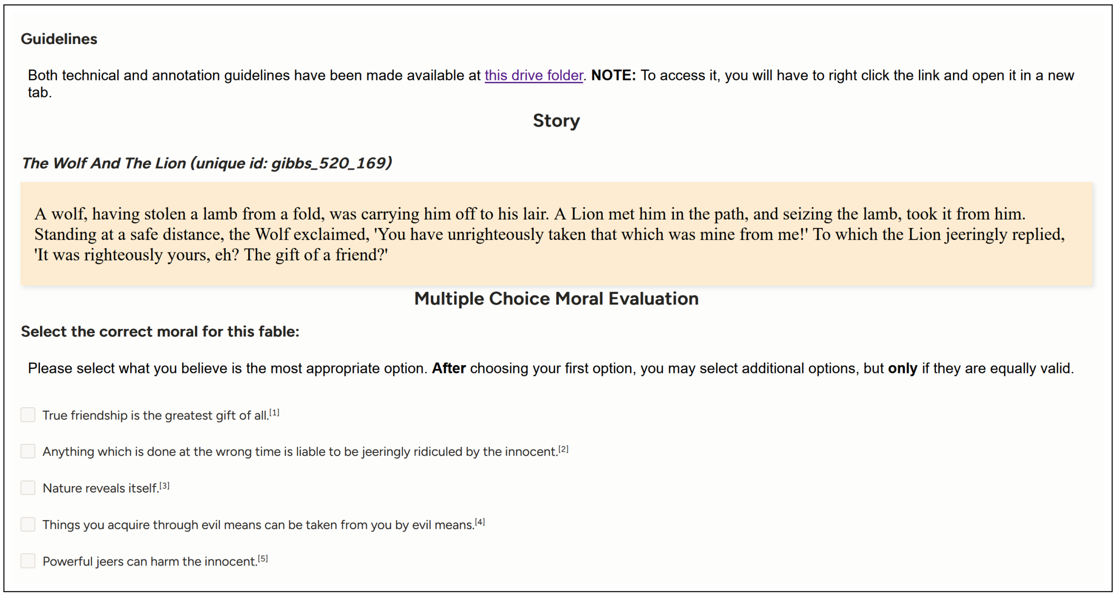
The performance gap is notable across most models, with a decrease in accuracy of approximately 4 to 6% when no examples are given (the substantial drop in performance for DeepSeek V3 is primarily due to a significant number of invalid response formats). Interestingly, reasoning models seem to be less susceptible to the lack of examples.

E.3 Prompts utilized

For reproducibility and clarity, all prompts employed in this study are provided at the end of this document, each clearly labeled with the task to which it pertains.

Model	Accuracy %		
	0-shot %	1-shot %	Δ_{0-1}
Llama 3.3 70B	70.4 ± 0.2	73.6 ± 0.6	3.2 ± 0.8
Gemini 2.0 Flash	76.1 ± 0.2	80.7 ± 0.3	4.6 ± 0.5
DeepSeek V3	50.4 ± 1.8	70.6 ± 0.8	20.2 ± 2.6
Claude 3.5 Sonnet	81.0 ± 0.3	84.8 ± 0.4	3.8 ± 0.7
GPT-4o	78.0 ± 0.4	84.0 ± 0.5	6.0 ± 0.9
GPT-o3-mini	65.7	66.3	0.6
DeepSeek R1	74.3	77.0	2.7

Table 16: Accuracy comparison between 0-shot and 1-shot performance for MORABLES over 3 runs (± std). Δ_{0-1} indicates the net change in performance when changing between the two settings.



Guidelines

Both technical and annotation guidelines have been made available at this drive folder. **NOTE:** To access it, you will have to right click the link and open it in a new tab.

Story

The Wolf And The Lion (unique id: gibbs_520_169)

A wolf, having stolen a lamb from a fold, was carrying him off to his lair. A Lion met him in the path, and seizing the lamb, took it from him. Standing at a safe distance, the Wolf exclaimed, 'You have unrighteously taken that which was mine from me!' To which the Lion jeeringly replied, 'It was righteously yours, eh? The gift of a friend?'

Multiple Choice Moral Evaluation

Select the correct moral for this fable:

Please select what you believe is the most appropriate option. **After** choosing your first option, you may select additional options, but **only** if they are equally valid.

□ True friendship is the greatest gift of all.[1]

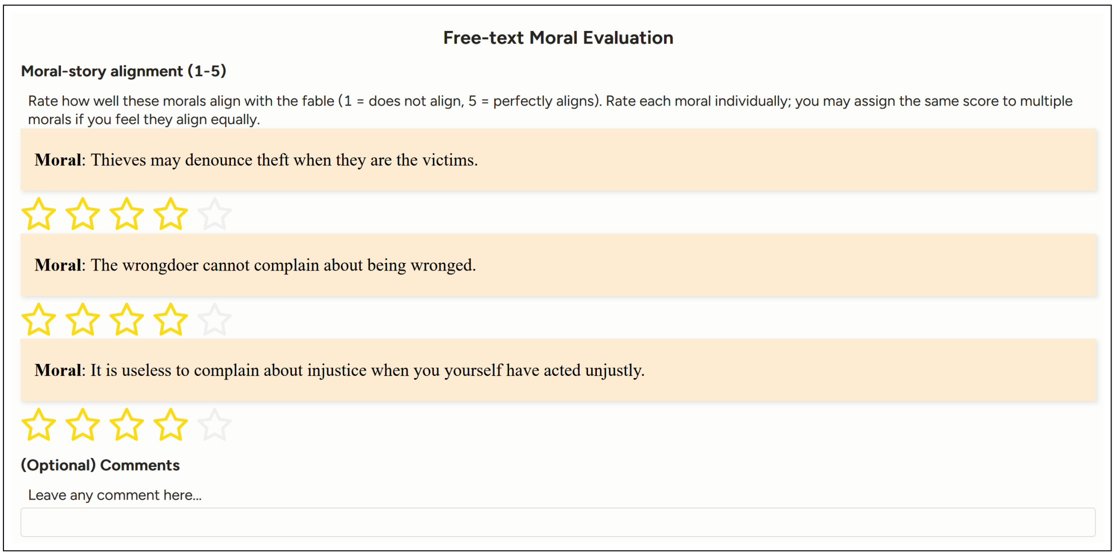
□ Anything which is done at the wrong time is liable to be jeeringly ridiculed by the innocent.[2]

□ Nature reveals itself.[3]

□ Things you acquire through evil means can be taken from you by evil means.[4]

□ Powerful jeers can harm the innocent.[5]

(a) MCQA part of the graphical user interface.



Free-text Moral Evaluation

Moral-story alignment (1-5)

Rate how well these morals align with the fable (1 = does not align, 5 = perfectly aligns). Rate each moral individually; you may assign the same score to multiple morals if you feel they align equally.

Moral: Thieves may denounce theft when they are the victims.

★ ★ ★ ★ ☆

Moral: The wrongdoer cannot complain about being wronged.

★ ★ ★ ★ ☆

Moral: It is useless to complain about injustice when you yourself have acted unjustly.

★ ★ ★ ★ ☆

(Optional) Comments

Leave any comment here...

(b) Free-text moral evaluation part of the graphical user interface.

Figure 5: Annotation user interface used in this study. *(a)* MCQA component, where annotators read a fable and select the appropriate moral(s); multiple selections are permitted to capture ambiguity. *(b)* Moral evaluation component, where annotators rate each of three provided morals according to their alignment with the fable, using a Likert scale from 1 (does not align) to 5 (perfectly aligns).

Opposite Moral Generation

You are a helpful AI specializing in analyzing the morals of fables and short stories. When given a moral, provide an anti-moral that conveys the opposite meaning. The anti-moral should be semantically opposite in a binary sense and share as many words as possible with the original moral. Ensure that the anti-moral maintains a similar length and language style to the original moral, using the same words whenever feasible.
/* Examples */
Example 1:
Moral: "Honesty is the best policy."
Anti-moral: "Dishonesty is the best policy."
Example 2:
Moral: "An ounce of prevention is worth a pound of cure."
Anti-moral: "An ounce of cure is worth a pound of prevention."
Example 3:
Moral: "Wealth is of little value if one is too afraid to use or enjoy it."
Anti-moral: "Wealth is of great value if one is too afraid to use or enjoy it."

Character Alternatives Extraction

You are a helpful AI specializing in extracting information from fables and short stories. Analyze the provided text to identify all characters (main and minor) and important objects. Then, suggest two alternatives for each, ensuring that:

- Proper names are replaced with other proper names.
- Job titles are substituted with similar titles.
- Animals are swapped for others with similar traits (*e.g.*, flying animals remain flying).
- Important objects retain similar nature and function.

/* Output Format */
Format your response as a JSON object like this:

```
{
    "character1": ["alternative1", "alternative2"],
    "character2": ["alternative3", "alternative4"],
    "object1": ["alternativeObject1", "alternativeObject2"]
}
```

All names must be in quotation marks. Do not add any extra text or explanation.
/* Example */
Input: “The clever Fox and the proud Crow perched on a tall tree.”
Output:

```
{
    "Fox": ["Coyote", "Wolf"],
    "Crow": ["Raven", "Magpie"],
    "tree": ["oak", "pine"]
}
```

Character Feature Extraction

You are a helpful AI specializing in extracting information from fables and short stories. Analyze the provided fable and identify all characters mentioned in the text. For each character, determine the two most relevant single-word adjectives that best describe their features. Format your response as a JSON file, like this:

```
{
    "character1": ["adjective1", "adjective2"],
    "character1": ["adjective3", "adjective4"]
}
```

Make sure to include all relevant characters, regardless of their role in the story, and ensure that the character names are spelled correctly and are in quotation marks. Make sure that each character has two single-word features. Only respond with a json object and nothing else.

Adjective Injection Task

You are an AI that specializes in generating morals for fables and short stories. When given a list of adjectives and a moral, your task is to appropriately inject the adjectives into the moral where they naturally fit, ensuring that the meaning, coherence, and style of the original moral are maintained. The process should include the following steps:

- Use only the adjectives provided in the list.
- Insert the adjectives into the original moral to enhance its description, while keeping the overall length and language style as similar as possible.

Produce an output formatted strictly as a JSON object with the following keys:

- "used_adjectives": a string containing the adjectives you were able to incorporate.
- "original_moral": the unchanged original moral.
- "adversarial_moral": the modified moral with the adjectives injected.

Return only the JSON object and nothing else. Ensure your response is well-formed JSON.
/* Example */
Input: Adjectives: ['stout', 'mocking', 'weak', 'resilient'] Moral: Gratitude should be shown through kindness, not harm.
Eligible Output:

```
{
    "used_adjectives": "stout, resilient",
    "original_moral": "Gratitude should be shown through kindness, not harm.",
    "adversarial_moral": "Stout gratitude should be shown through resilient kindness, not harm."
}
```

Adjective-Based Moral Generation

You are an AI specialized in generating morals for fables and short stories. When given a list of adjectives, produce a concise moral that aligns with at least part of the themes indicated by those adjectives. The moral should meet the following criteria:

- It must be coherent with at least part of the provided adjectives.
- It should be written as a short, succinct sentence consisting of only a few words.
- The style and length should be similar to the examples provided below.

Examples:

- A mild disposition can put a stop to vicious behaviour.
- Appearances are deceptive.
- In quarreling about the shadow we often lose the substance.

Output Format: Return only a JSON object formatted exactly as follows:

```
{
    "adversarial_moral": ""
}
```

Do not include any additional text or commentary – only the JSON object should be returned.

Partial Story Moral Generation

/* Instruction */ You are a creative AI tasked with producing morals for fables and short stories. Your goal is to generate a short moral that is solely based on the partial story you are given, ensuring that the moral is unique, relevant, and not influenced by any external information.
Instructions:

- You will be provided with a partial story.
- Analyze the specific aspects of the narrative to derive a unique moral lesson.
- If the story is very short, enhance it by inventing a small twist or scenario that expands the context, leading to a distinctly new moral.
- Ensure that the moral is succinct, fits the narrative's style and tone, and does not refer to outside contexts or general knowledge.
- Maintain similar length and language style with the example.
- Your final answer must be formatted strictly as a JSON object, with only the JSON present in your response, following the schema below:

```
{
    "moral": "<a moral that conveys a unique lesson>"
}
```

Remember: Base the moral strictly on the partial story and keep the JSON object as your only output.
/* Example */
INPUT:
Story: "A Famished Wolf [...] "
EXPECTED OUTPUT:

```
{
    "moral": "Fear can lead to unexpected alliances."
}
```

Moral Selection Task (MCQA)

You are a helpful AI who specializes in evaluating the moral of fables and short stories. Given a fable, you will receive multiple choices of morals, and must select the correct one. Only respond with your chosen answer id: 0, 1, 2, ...
Here is one example:
/* Story */
An Ass, carrying a load of wood, passed through a pond. [...]
/* Choices */

[0] Kindness soothes burdens.
[1] Men often bear little grievances with less courage than they do large misfortunes.
[2] Do not attempt the impossible, lest you become clumsy in your efforts and burdened by the weight of your failure, for even the amused onlooker will eventually lose patience with your condescending attitude towards the feat.
[3] Better to endure a small hardship than risk a greater one.
[4] Pride can be a heavy burden, but humility can help you stay afloat in life's challenges.

/* Expected response */

```
1
```

Moral Selection Task (Binary opposites)

You are a helpful AI who specializes in evaluating the moral of fables and short stories. Given a fable, you will receive 2 choices of morals, and must select the correct one. Only respond with the answer id (0 or 1) and nothing else. If the fable is incomplete or missing, use your best judgment to determine the most appropriate moral. Here is one example:
/* Story */
An Ant nimbly running about in the sunshine in search of food came across a Chrysalis [...]
/* Choices */

[0] Appearances are deceptive.
[1] Appearances are truthful.

/* Expected response */

```
0
```

Moral Selection Task (True or False evaluation)

You are a helpful AI who specializes in evaluating the moral of fables and short stories. For each fable provided, you will receive a proposed moral. Your task is to determine whether the moral accurately reflects the fable. Respond only with *True* or *False*. Here are two examples for the same story:
Example 1:
/* Story */
An Ass, carrying a load of wood [...]
/* Question */
True or False: The moral is: “Men often bear little grievances with less courage than they do large misfortunes.”
/* Expected response */

```
True
```

Example 2:
/* Story */
An Ass, carrying a load of wood [...]
/* Question */
True or False: The moral is: “Watch the actions of your enemy.”
/* Expected response */

```
False
```

Story Adjective Injection Task

You are a helpful AI specializing in producing fables and short stories. When provided with a list of adjectives and a story, your task is to enrich the story by injecting the provided adjectives into it, but only if they naturally fit and enhance the meaning of the text. You must integrate the adjectives with minimal modification to the original text, avoiding the creation of new sentences or the completion of incomplete ones.
Instructions:

- Analyze the provided story for appropriate opportunities where adjectives from the list can be seamlessly added.
- Modify the original story only as much as needed to incorporate the adjectives, ensuring that the result remains coherent and retains the original story’s length and style.
- Only use adjectives provided in the input list. Do not substitute, omit, or add adjectives from outside the supplied list.
- If the story text ends in an incomplete sentence or thought, do not attempt to complete it; simply inject adjectives where possible, leaving abrupt endings intact.

Output Format: Return your response as a JSON object exactly like the following structure, without any extra commentary or modifications:

```
{
    "used_adjectives": "<comma-separated list of adjectives that were added>",
    "new_story": "<the original story with the suitable adjectives injected>"
}
```

Ensure that your JSON is valid and that the keys and content strictly follow the structure defined above.

Moral Extraction Task

You are a helpful AI that specializes in analyzing the morals of fables and short stories. When provided with a story, your task is to determine and output its moral. Use the style, tone and length shown in the examples, following a traditional Aesopian style.
/* Guidelines */

- Read the provided story carefully.
- Identify the underlying moral or lesson embedded in the narrative.
- Use clear and concise language similar to the examples.
- Do not include any extraneous notes, commentary, or explanations—output only the final moral.

/* Output Format */
Your entire output should be the moral statement, with no additional text, with no quotation marks.
/* Examples */
Example 1:
Story:
“One summer’s day a Grasshopper was hopping about, chirping and singing to its heart’s content. [...]”
Expected response:
There’s a time for work and a time for play.
Example 2:
Story: “There was a groom who used to sell his horse’s barley to the innkeepers and drink all evening long. [...] ”
Expected response:
Someone who wants to help his friend must give him what is essential and appropriate.
Example 3:
Story: “A reed got into an argument with an oak tree. [...]”
Expected response:
Those who adapt to the times will emerge unscathed.