ChatGPT

Folklore-Inspired LLM Morality Evaluation Framework

Overview and Motivation

Folktales, folklore, and folk songs have long been used to impart moral lessons across generations. Stories are not only meant to entertain – they **encode lessons reflecting the author's beliefs about the world** [1]. In fact, communicating moral lessons is one of the oldest functions of storytelling, dating back millennia [2]. This suggests that the rich trove of global folklore can serve as a source of *ethical scenarios* for evaluating AI behavior. The proposed framework leverages these traditional tales of morality as a benchmark to test whether a large language model (LLM) can discern right from wrong in context, and whether its choices align with the **intended moral of each story**. The idea is to pose moral questions or dilemmas derived from folktales to an LLM and measure the percentage of "correct" answers – i.e. answers consistent with the folktale's moral lessons – as an indicator of the model's ethical alignment.

By using folklore as an evaluation tool, we ground the assessment in culturally resonant narratives rather than abstract rules. This approach would complement existing safety tests with **narrative-driven moral challenges**. In the following sections, we outline how to design such a benchmark, address potential criticisms (e.g. training data overlap and cultural bias), and compare this folklore-inspired method with Anthropic's current use of universal principles (like UN guidelines) for AI guardrails.

Global Folklore Moral Benchmark (with Local Variants)

Global Moral Benchmark: We can begin by compiling a broad corpus of folktales and fables from around the world, each known for conveying a particular moral or ethical lesson. Researchers have already created international folktale corpora to study values and morals across cultures [3]. By clustering the morals extracted from a diverse folktale dataset, it's possible to reveal both **common moral themes and culture-specific values** [4]. A global benchmark would include scenarios covering universal virtues such as honesty, kindness, generosity, humility (e.g. *"The Boy Who Cried Wolf"* teaches honesty; *"The Tortoise and the Hare"* rewards perseverance), as well as cautionary tales about vices like greed or pride. The LLM would be evaluated on **how often its answers align with these traditional moral outcomes**, giving an aggregate score of moral alignment. Such a global test ensures coverage of widely shared ethical principles found across humanity's storytelling heritage.

Local Variants for Cultural Adaptation: In addition to the global suite, the framework can feature **localized moral tests** derived from the folklore of specific regions or cultures. This acknowledges that moral norms can be culturally relative and that AI systems might need **regional guardrails** to be truly accepted. For example, one might curate a set of African folktales to serve as a sub-benchmark for AI used in African contexts, or a collection of East Asian folklore for models deployed in those societies. Local folk narratives embed the nuances of regional values and social norms [5] [6]. Adapting the evaluation to include these means the LLM is not only tested against generic "global" ethics, but also for sensitivity to **local cultural values**. Such localization could improve user trust: surveys indicate people trust AI more

when it reflects their local norms [7]. However, it's important to balance this with universal ethical standards – local guardrails should *augment*, not violate, fundamental human rights or global ethics [6]. In practice, we could maintain a core global morality benchmark (analogous to a base safety standard) and then layered regional tests that introduce additional criteria or scenario variations drawn from local folklore and folk songs.

Deriving Direct Ethical Dilemmas from Folktales

A key step is translating folktale narratives into **evaluation prompts** that present direct ethical dilemmas or questions to the LLM. Many folktales center on a protagonist's critical decision or a test of character, which we can harness for Q&A style evaluation. For each story in the benchmark, we would distill it down to a scenario and a moral question. Some possible formats include:

- *Multiple-choice dilemmas:* e.g. present the setup of a tale and ask, *"What should the character do in this situation?"* with one answer reflecting the folktale's morally correct action and others reflecting wrong or immoral choices. The model's task is to choose the option that aligns with the tale's lesson (for example, **not** cheating or lying if the story's moral is honesty). A large set of such questions allows us to calculate a percentage of correct moral choices by the LLM.
- *Direct questions about right vs. wrong:* e.g. *"In the story, was it right for X to do Y?"* expecting a justification that matches the story's verdict. For instance, asking if it was right that a character forgave their enemy, when the folktale frames forgiveness as virtuous.
- *Moral-of-the-story identification:* e.g. provide a brief story synopsis and ask the model *"What is the moral or lesson of this story?"* – essentially testing if the LLM can articulate the folk wisdom. This checks whether the model internalized the narrative's ethical takeaway.

The evaluation framework should include **clear ground-truth answers** for each item. In many cases, folktales explicitly state the moral (as in Aesop's fables) or have an ending that implies the correct resolution (the virtuous character is rewarded, the selfish one punished, etc.). These can serve as the "key" to score the model's answer. Because folktale dilemmas are often straightforward and allegorical, they make for intuitive test cases. This approach also ensures we're testing the model on **contextual morality** – the AI must consider the scenario and choose an action that is *morally acceptable within that narrative*.

By using **direct ethical dilemmas drawn from stories**, we ground abstract principles in concrete, memorable situations. This tests not just knowledge of rules, but application of moral reasoning in context. For example, a question might derive from *"Little Red Riding Hood"*: *"Should Little Red Riding Hood trust a stranger in the woods offering help?"* – the expected answer would be "No" (caution with strangers), reflecting the tale's safety moral. A model consistently aligning with these folk lessons demonstrates it can navigate classic ethical scenarios.

Open-Ended Evaluation for LLM Moral Reasoning

While multiple-choice and direct Q&A give us quantitative scores, **open-ended prompts** are crucial to evaluate the depth of an LLM's moral reasoning. In this framework, we can allow the model to generate free-form answers or explanations, and then assess those for moral alignment. For example, we might present a folktale scenario and ask: *"What advice would you give the main character?"* or *"How should this story end to teach a positive lesson?"*. The LLM's response can be compared to the actual folktale's conclusion or moral.

Using open-ended questions has several advantages: it lets the model **explain its ethical reasoning**, not just pick an option. This way we can observe if the AI can articulate the principle (e.g. *"crime doesn't pay"* or *"be kind to others"*) in its own words. Previous research suggests LLMs are already quite capable in this area – for instance, GPT-4 showed a high agreement with humans in identifying the values and lessons of folktales [3]. In practice, we could automate the evaluation of open responses by using another AI or heuristic to judge similarity to the known moral. One method is a *reference answer check*: we have a reference summary of the correct moral for each tale, and we measure if the model's answer captures the same idea (via semantic similarity or by using an LLM judge prompt). Hobson et al. (2024) demonstrated a multi-step prompting technique with GPT-4 to **extract story morals** and validate them, finding that LLMs can effectively approximate human interpretations of a story's lesson [8]. We can leverage similar techniques: for example, prompt the model first to give an answer, then prompt it or another model to **critique whether that answer aligns with the folktale's moral** (in a process analogous to Anthropic's constitutional AI self-critique, but focused on story ethics).

Open-ended evaluation also uncovers nuance. Sometimes a story's moral might be interpreted in slightly different words by the model – this is acceptable as long as it's semantically aligned. We should design the scoring to give partial credit for answers that capture the spirit of the lesson even if phrased differently. Moreover, analyzing the AI's explanations can reveal *how* it is making decisions: does it cite the virtuous behavior or consequence that the folktale emphasizes? Ideally, a well-aligned model should not only get the "right answer" but also justify it with reasoning consistent with the cultural context of the tale. This provides richer insight than a simple correct/incorrect metric.

In summary, mixing **structured questions** (for clear correctness metrics) with **open-ended prompts** (for qualitative reasoning analysis) will create a comprehensive picture of the LLM's moral decision-making capabilities. We want to ensure the model isn't just guessing the right answer but truly **understands the moral dimension** of the scenario.

Criticisms and Challenges of a Folklore-Based Approach

Before embracing this folklore-driven evaluation, it's important to acknowledge potential criticisms and limitations:

- **Training Data Contamination:** Modern LLMs are trained on vast internet text, which likely includes many classic folktales and their morals. This raises the issue of *benchmark leakage*: the model might already "know" the correct answers simply by recall, rather than by reasoning. If an evaluation question closely resembles a story the model memorized, it could score highly without truly demonstrating moral deliberation. In other words, using folktales that were part of training data can lead to **inaccurate or inflated performance** on the test [9]. This is a known challenge in LLM evaluation – models sometimes inadvertently incorporate evaluation content from training, skewing results [9]. Mitigation strategies include using less-famous or region-specific folktales that the model is less likely to have seen, paraphrasing scenarios to avoid verbatim overlaps, or employing *benchmark filtering* techniques to detect and remove contaminated samples. Ensuring the evaluation truly measures ethical reasoning (and not just memory) is essential for its validity.

- **Cultural Biases in Folklore:** Folktales carry the values and biases of the cultures and eras they come from. While they teach morals, those morals might not always align with modern ethical sensibilities or universal human rights. For example, some old tales have **gender stereotypes or prejudices**

embedded in them [3], or endorse harsh punishments that today might be seen as excessive. If we use such tales as "ground truth" for morality, we risk **validating outdated or biased norms**. This needs careful curation: the framework should vet which folk narratives are used, perhaps favoring those with broadly positive lessons (honesty, kindness) and excluding or reinterpreting those with problematic messages. It's a delicate balance – we want cultural diversity, but we must avoid enshrining unjust biases as correct answers. One way to handle this is to explicitly note the presence of biases in certain stories and possibly adjust the expected answer (for instance, acknowledging in evaluation that a tale's moral might reflect historical context not a current ideal). Researchers have indeed explored how **values and biases are expressed in folktales across cultures** [3], highlighting that folklore is not value-neutral. This awareness should inform how we design the benchmark and interpret results.

- **Folktales as a Proxy for Modern Morality:** Another critique is whether success on folktale dilemmas truly translates to ethical behavior in today's complex world. Folklore tends to illustrate fairly **simplified moral situations** – clear heroes and villains, or singular virtues (e.g. "slow and steady wins the race"). Real-world ethics can be far more nuanced, involving dilemmas that folktales never imagined (data privacy, bioethics, AI-related harm, etc.). For instance, a model could learn to always choose the generous or honest action in a story context, but that doesn't guarantee it will handle a nuanced policy question or a conflict between two rights. Thus, **folklore-based evaluation may have blind spots**. It might be a good test for basic moral intuition and cultural value alignment, but not a comprehensive measure of an AI's ethical judgment in all domains. We likely need to use this folklore benchmark *alongside* other evaluations (e.g. direct ethical questionnaires, scenario-based tests from contemporary life, or formal frameworks from moral philosophy). Folklore can be a starting point – a culturally rich baseline – but not the only yardstick of AI morality.

- **Conflicting Morals and Interpretations:** With a global collection of folktales, we might encounter stories whose lessons conflict with each other or with the chosen universal standards. Different cultures might prioritize values differently (honor vs. forgiveness, individualism vs. community, etc.). Even within one story, people could interpret the moral in various ways [10]. This can complicate scoring: whose interpretation of the moral is "correct"? If an LLM gives a reasonable alternative perspective on a story's lesson, is that wrong? We need a consistent method to decide the expected answer for each tale – possibly consulting literature or majority human judgment on that folktale's meaning. Additionally, if using local benchmarks, there's a risk of **fragmentation**: a model might perform well by one culture's standards but poorly by another's. We should design the evaluation to highlight such differences without unfairly penalizing the model (one solution is to **compare the model's responses to local questions with how locals themselves answer**, to see if it aligns with the target culture's values). This challenge echoes the broader difficulty of merging *localized ethics* with *global ethics* [6].

In summary, the folklore-based evaluation is promising but **must be handled carefully**. We have to guard against the model simply parroting training data, avoid reinforcing outdated biases, and be mindful that this is a partial proxy for moral competence. Transparent documentation of the benchmark's scope and limitations will be important. These criticisms do not invalidate the approach – rather, they indicate areas where additional rigor and complementary methods will be needed.

Contrasting with Anthropic's UN-Based Guardrails

Anthropic's approach to AI safety, known as *Constitutional AI*, provides an illuminating foil to the folklore framework. Anthropic uses a **fixed set of high-level principles (a "constitution") derived from universal values** to guide model behavior. Notably, **their constitution draws from the United Nations Universal Declaration of Human Rights, along with other well-vetted sources** [11]. This means the AI is aligned by explicit rules like "do not discriminate" or "respect freedom and safety," which are broadly accepted worldwide. The focus is on **global, normative standards** – essentially a top-down imposition of ethical guidelines on the model.

Folklore vs. Constitutional AI – Key Differences:
- *Source of Morality:* Anthropic's guardrails come from formal documents and expert-informed principles (e.g. UN charters, corporate policies) [11]. In contrast, the folklore benchmark's "rules" emerge from **grassroots cultural narratives**. Instead of abstract norms, we have concrete stories with moral outcomes. This bottom-up sourcing means our evaluation covers morality as portrayed in everyday cultural content, which might include lessons on personal virtue or community values not explicitly codified in UN guidelines. It's a more **story-driven, example-based** approach versus Anthropic's rule-based approach.

- *Universality vs. Local Adaptation:* The UN Declaration and similar principles aim to be **universal** – they represent common denominators of human rights and ethics. Anthropic largely applies one global constitution to all interactions. Our folklore method, on the other hand, explicitly allows for **local variations**. By having region-specific folklore tests or guardrails, the AI's behavior could be fine-tuned to different cultural contexts. For example, Anthropic might ensure a model never violates a human right anywhere, whereas the folklore approach could additionally ensure the model honors particular cultural etiquettes or moral nuances when used in that culture. This makes the framework **more culturally pluralistic**. (Anthropic did acknowledge the need for non-Western perspectives in their principles [11], but it's still one merged set of rules. Folklore-based evaluation would treat each culture's heritage somewhat distinctly, if desired.)

- *Explicit Principles vs. Implicit Understanding:* Anthropic's model is trained to follow explicit written principles (which can be read in their "constitution"). The evaluation is then often a matter of checking if outputs violate those principles. With folklore, the "principles" are implicit in each story. The model isn't handed a list of morals; instead, it must demonstrate it **understands the moral through application**. This is a subtler test – rather than asking "Does the output comply with rule X?", we ask "In scenario Y, did the model's action align with the story's lesson?". It's possible an Anthropic-aligned model would do well in folklore scenarios if the scenarios map onto their rules (e.g. a story about not stealing aligns with the rule "don't commit wrongdoing"). However, folklore may cover **softer social virtues** (humility, hospitality, etc.) that aren't explicitly in a human rights document. This could expose differences in how a model trained on UN rules vs. one evaluated on folk morals behaves. The folklore benchmark might catch subtle cultural ethics that a straightforward UN-based checklist might overlook.

- *Evolving vs. Static Benchmark:* UN-based constitutions are relatively static and **designed by experts**. Folklore is dynamic and diverse, potentially **evolving as we include more tales or as cultures change**. Implementing folklore guardrails means periodically updating or expanding the corpus of stories considered. It's less centralized – potentially a community of users from each locale could contribute relevant tales and songs that reflect current values. This approach could be more

democratic and adaptive but also harder to manage consistently than Anthropic's centrally defined constitution.

In practical terms, Anthropic's method excels at ensuring a baseline of safety and harm-prevention aligned with *universal human rights and contemporary ethics*, which is crucial for avoiding egregious behavior. The folklore evaluation would **complement that by testing moral reasoning in relatable contexts** and checking cultural alignment. One might imagine using both: e.g., first ensure a model passes a UN-based safety filter (no violations of core principles), then also test it against folklore scenarios from the target user communities. If a discrepancy arises – say the model's output is permissible by UN standards but offends a local moral sensibility – developers could decide how to adjust (perhaps giving the model additional training on that context or adding a specific guideline).

In summary, **Anthropic's approach provides broad guardrails informed by global consensus**, whereas the folklore framework provides **granular, story-level checkpoints informed by cultural wisdom**. The latter is more about *how the AI reasons through a scenario* (in line with a traditional moral), and the former about *what the AI must never do* (according to high-level ethics). Both approaches share the goal of aligning AI with human values – they just draw those values from different wells. Highlighting this contrast underscores that our folklore-based benchmark is a novel angle: testing alignment not by formal principles alone, but by the rich, varied moral ecology found in human storytelling.

Implementation Plan and Next Steps

Designing and deploying this folklore-inspired morality benchmark will involve several steps. Below is an overview of how one could accomplish it, paving the way for a software implementation:

1. **Dataset Collection:** Gather a comprehensive set of folktales, fables, mythologies, and folk songs from as many cultures as possible. Sources might include public domain folktale collections, academic databases, and community contributions. For each tale, record the *core plot, the key moral dilemma, and the commonly understood moral lesson*. Existing research can guide this; for instance, Hobson et al. (2024) extracted morals from diverse narratives including folktales [8], and Wu et al. (2023) compiled an international folktale corpus highlighting values and biases [3]. These efforts could seed the dataset. It will be important to store metadata like the culture/region of origin for each story, so we can filter by locale.

2. **Question/Prompt Formulation:** For each tale, create one or more evaluation prompts. This includes:

3. A **scenario prompt** summarizing the tale's ethical dilemma or critical situation in a few sentences (enough for the model to understand the context).
4. A **question or instruction** for the LLM. This could be a direct question ("What should the character do?" or "Was that action right or wrong?"), a multiple-choice question, or an open-ended task ("Explain the lesson of this story.").

5. The **expected answer or scoring rubric**. For multiple-choice, mark the correct choice(s). For open-ended, have a reference answer (the moral lesson in a concise form) to compare with. In software, this could be implemented as a function that checks if the model's answer contains certain key ideas

or phrases matching the moral. In more advanced setups, one might use an automated evaluator model that was fine-tuned to judge moral alignment of answers.

6. **Global and Local Subsets:** Partition the prompts into a *Global* set and various *Local* sets. The Global set would include a balanced mix of stories from different regions, representing universally relevant morals. Then, for each region or cultural group of interest, create a subset of prompts specific to that culture's folklore. For example, one subset could be **"Japanese Folklore Morality"** (featuring dilemmas from Japanese tales), another could be **"West African Folklore"**, and so on. This allows evaluation in two modes: overall performance (all cultures) and in-culture performance (does the model particularly understand, say, Indian moral stories when evaluated on that subset?). Organizing the data this way will make it easy in software to select which benchmark variant to run (global vs. a local variant).

7. **Evaluation Engine:** Develop the evaluation harness that will pose these questions to the LLM and record its answers. This involves writing scripts or using an existing LLM evaluation toolkit to:

8. Load the prompts (possibly randomly sampling a set number per culture to ensure coverage).
9. Query the LLM (which could be done via API calls if it's an external model, or directly if you have the model locally).
10. For each response, apply the scoring logic. For multiple-choice, this is straightforward matching. For open-ended, use a comparison method: e.g., measure semantic similarity to the reference moral. In practice, one might use an embedding-based similarity or a simple heuristic to check for the presence of moral keywords. Another robust approach is to use a secondary AI model: prompt a model like GPT-4 with the story and the LLM's answer, asking it to score how well the answer aligns with the story's true moral (essentially using AI as a judge, which has been found effective in other moral evaluation research [8]).

11. Aggregate results into metrics: overall accuracy (% of correct moral decisions), and breakdowns by category (by culture, by type of moral, etc.). This will identify strengths and weaknesses of the LLM's moral reasoning. For instance, the model might excel at questions about honesty but falter on ones about courage, or do well on Western tales but struggle with East Asian ones – such patterns would be insightful.

12. **Critical Analysis of Results:** Once the evaluation is run, analyze the outcomes. If an LLM achieves, say, 90% on the global folklore benchmark, we'd interpret that as very high alignment with folk morals. However, we should cross-check *which questions were missed*. This analysis could surface systematic issues: e.g., did the model consistently make utilitarian choices when the folktale favored a deontological stance? Did it misunderstand any culturally specific context or proverb? This step loops back to improving the framework – we might find we need to reword some prompts for clarity, or that certain folktale morals were too ambiguous and require refined scoring criteria.

13. **Refinement and Iteration:** Based on the analysis and any expert feedback, refine the benchmark. This may involve removing ambiguous cases, adding more stories to cover gaps, or adjusting for any biases noticed. Because folktales can be interpreted in multiple ways, it may also help to involve human experts or native cultural insiders to ensure the "expected answers" are fair and representative. Iteratively, the benchmark will become more robust. On the software side, continuous integration of the evaluation in the model development cycle is ideal – e.g., every time a

new model version is trained, automatically run the folklore benchmark and track trends in performance.

14. **Incorporating Findings into Model Development:** The ultimate goal is not just to score models, but to use these insights to improve AI ethics and safety. If the eval reveals specific weak spots (say, the model often chooses vengeance in scenarios where the moral preaches forgiveness), developers can address this. Possible interventions include further fine-tuning the model on moral reasoning data, adding those folktale scenarios into its training (with correct answers), or adjusting the model's alignment technique to emphasize certain values. For example, one might include folklore-based prompts in a reinforcement learning from human feedback (RLHF) process, rewarding answers that match the moral and penalizing those that don't. Over time, the model should internalize these lessons – effectively **learning from folklore** in addition to learning from abstract rules.

Throughout implementation, it's crucial to **document the process and ensure transparency**. This means recording which stories were used (to avoid secrecy and to allow external critique of whether those are appropriate), and noting any cases where the "morally correct" answer might be debatable. Since this benchmark could be expanded with community input, an open-source repository (containing the prompts, answers, and evaluation code) would be beneficial. In fact, researchers have released related resources (e.g. Hobson et al.'s code for extracting story morals [12]), and our framework could similarly be shared for collaborative improvement.

By following these steps, we move from the conceptual idea of a folklore-based morality test to a concrete tool that can be run in software. The next phase would involve actually coding this system: building the dataset programmatically, writing prompt scripts, and integrating with an LLM API or model checkpoint. Once implemented, this benchmark would offer a novel lens on AI ethics – one that measures models against the timeless moral wisdom found in our collective folklore. It could serve as a **global baseline with local adaptability** for evaluating and guiding the ethical decision-making of new LLMs, complementing other evaluations and helping ensure these models act in morally acceptable ways.

---

[1] [2] [4] [8] [10] [12] aclanthology.org
https://aclanthology.org/2024.emnlp-main.723.pdf

[3] Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales | Request PDF
https://www.researchgate.net/publication/376402893_Cross-Cultural_Analysis_of_Human_Values_Morals_and_Biases_in_Folk_Tales

[5] [6] [7] Localized Ethical Frameworks: Aligning AI with Regional Cultural and Societal Norms | AIGN
https://aign.global/ai-ethics-consulting/patrick-upmann/localized-ethical-frameworks-aligning-ai-with-regional-cultural-and-societal-norms/

[9] Benchmark Data Contamination of Large Language Models: A Survey
https://arxiv.org/html/2406.04244v1

[11] Claude's Constitution \ Anthropic
https://www.anthropic.com/news/claudes-constitution