# FFE-Hallu: Hallucinations in Fixed Figurative Expressions

A Benchmark of Idioms and Proverbs in the Persian Language

**Anonymous ACL submission**

## Abstract

Figurative language, especially fixed figurative expressions (FFEs) such as idioms and proverbs, poses unique challenges for large language models (LLMs). Unlike literal phrases, FFEs are culturally grounded and often non-compositional, making them vulnerable to figurative hallucination, the generation or acceptance of plausible-sounding but culturally invalid expressions. We introduce **FFE-Hallu**, the first comprehensive benchmark for evaluating LLMs' ability to generate, detect, and translate FFEs in Persian, a linguistically rich but underrepresented language. FFE-Hallu includes 600 carefully curated examples spanning three tasks: FFE generation from meaning, detection of fabricated FFEs (across four controlled categories), and FFE-to-FFE translation from English to Persian. Our evaluation of six state-of-the-art multilingual LLMs reveals persistent weaknesses in both cultural grounding and figurative competence. While models like GPT-4.1 display relative strength in rejecting fabricated FFEs and retrieving authentic ones, most systems struggle to reliably distinguish real FFEs from high-quality fabrications and often hallucinate in translation. This work shows that LLMs still have important gaps in understanding and using figurative language, and that specialized benchmarks like FFE-Hallu are needed.

## 1 Introduction

Large Language Models (LLMs) have achieved remarkable success across numerous NLP tasks, yet they often falter with figurative language—expressions whose meanings cannot be inferred literally. This work focuses on a subset of such expressions, which we term fixed figurative expressions (FFEs), encompassing idioms and proverbs. FFEs are largely non-compositional and culturally grounded, making them especially difficult for models to interpret or generate faithfully (Kabra et al., 2023; Chakrabarty et al., 2022).

In natural communication, FFEs enrich both spoken and written language by conveying deeper meaning, humor, and cultural resonance. Their interpretations often cannot be captured through direct translation (De Luca Fornaciari et al., 2024; Mi et al., 2024; Rezaeimanesh et al., 2025). Correctly handling FFEs is therefore vital for translation, dialogue systems, and cross-cultural or educational applications, as it requires both linguistic and cultural competence (Liu et al., 2024; Li et al., 2023).

While factual hallucination in LLMs has been widely examined (Ji et al., 2022), figurative hallucination—the generation or endorsement of plausible-sounding but non-existent FFEs—remains unexplored. Such hallucinations can sound convincingly authentic, spread cultural misinformation, and erode trust in model outputs. The illusory truth effect further compounds this risk: familiar-sounding statements, especially those echoing idiomatic templates, are often perceived as true even by informed readers (Fazio et al., 2015). This makes non-native or novice users particularly vulnerable to fabricated FFEs. As reliance on LLM-generated text grows, mitigating this issue becomes increasingly urgent.

To address this gap, we introduce **FFE-Hallu** (Fixed Figurative Expressions Hallucination), a benchmark designed to evaluateLLMs' figurative and cultural competence, with a focus on hallucination detection and control. FFE-Hallu comprises three tasks: (i) *FFE Generation from Meaning* tests whether a model can produce a valid idiom or proverb given a figurative definition. (ii) *Fake FFE Detection* examines the model's ability to distinguish real expressions from plausible but fabricated ones.(iii) *FFE-to-FFE Translation* evaluates whether a model can generate culturally and semantically appropriate equivalents of idiomatic expressions across languages, without defaulting

to literal or invented phrases.

The benchmark contains 600 curated examples across three settings: 200 real idioms and proverbs, 200 carefully constructed fakes simulating figurative hallucinations, and 200 bilingual English–Persian idioms with verified equivalents. Together, these tasks offer a systematic framework for probing how LLMs navigate figurative meaning, cultural grounding, and hallucination.

This work makes the following contributions: (i) We introduce **FFE-Hallu**, the first benchmark for evaluating figurative hallucination in LLMs, with a focus on FFEs in Persian. We use three complementary tasks: FFE generation from meaning, fake FFE detection, and FFE-to-FFE translation, examining different aspects of figurative competence and cultural grounding. (iii) We construct a dataset of 600 FFEs, including 200 real idioms and proverbs, 200 fabricated examples generated via four controlled strategies grounded in relevant linguistic research, and 200 English–Persian idiom pairs with verified figurative equivalence. (iv) We evaluate six multilingual LLMs (both open and proprietary) and conduct a detailed analysis of their vulnerabilities in handling figurative language, revealing core challenges in generation, detection, and cross-cultural translation.

## 2 Related Work

Research on hallucination and figurative language understanding intersects across several areas. We organize this section into two main subsections: (1) hallucination and reliability in language models, with a focus on definitions and taxonomies relevant to our work, and (2) prior efforts in figurative language processing, particularly idiom and proverb comprehension. This structure allows us to put **FFE-Hallu** at the intersection of these topics and highlight the novelty of evaluating *figurative hallucination*.

### 2.1 Hallucination and Reliability in Language Models

Hallucination, when a model generates content not grounded in input or training knowledge, has emerged as a major concern in language generation (Ji, 2023). Prior work categorizes hallucinations as intrinsic (contradicting the input) or extrinsic (unsupported by known facts) (Rawte et al., 2023). Recent research has further distinguished between factual and non-factual hallucinations (Cao et al., 2022). Factual hallucinations refer to content that is not directly inferable from the source material but is verifiable and correct according to world knowledge. Non-factual hallucinations are entities or information neither inferable from the source nor factual according to world knowledge (Bendahman et al., 2025). This distinction is particularly important as factual hallucinations may sometimes provide valuable additional context, while non-factual hallucinations are typically considered erroneous (Rawte et al., 2023).

Our definition of hallucination in the context of FFEs fits within the extrinsic category and specifically represents a form of non-factual hallucination, where models generate or accept fabricated idioms that appear plausible but are not certified in the target culture.

### 2.2 Figurative Language and Idiomaticity

Idioms are a central component of figurative language, characterized by non-compositional meanings that challenge both rule-based and data-driven systems. Recent work by De Luca Fornaciari et al. (2024) shows that idiom detection remains "a hard nut to crack" even for conversational LLMs, which still struggle to distinguish literal from figurative usage. Mi et al. (2024) introduce DICE, a framework for evaluating idiom comprehension through contextual understanding, and find that even state-of-the-art models perform poorly when figurative expressions appear in context.

The multilingual dimension of idiomaticity has been explored in studies such as MABL (Kabra et al., 2023), which evaluates figurative language understanding across eight languages, and in comparative analyses of idioms and similes in LLMs (Khoshtab et al., 2024), which report saturation effects in idiom interpretation but substantial variation by language and model.

For Persian, Rezaeimanesh et al. (2025) present PERSIANIDIOMS, a parallel corpus of 2,200 idioms—700 with usage examples—and evaluate translation accuracy and fluency using both LLM-based and hybrid systems, noting particularly strong performance from Claude 3.5 Sonnet. Gao et al. (2025) propose the CR-ST metric (Consistency Rating of Semantic Transparency) to assess idiomatic metaphor competence and introduce the prompt-based self-checking mechanism PASS to enhance model reliability.

Our work complements this research by focusing specifically on FFE generation and hallucination

resistance in Persian. It is the first benchmark designed to test whether LLMs can distinguish real FFEs from fabricated ones and generate culturally appropriate FFEs for given meanings.

## 3 FFE-Hallu Benchmark

Hallucination in LLMs has been widely studied in factual domains (Ji, 2023; Rawte et al., 2023), but here we focus on a novel subtype: *figurative hallucination* We define figurative hallucination as the generation or endorsement of a fixed FFE that is invalid within the target language or culture—that is, absent from reference dictionaries, linguistic corpora, and unrecognized by native speakers. This differs from *semantic errors*, where an existing FFE is produced that does not match the intended meaning; in our setting, such cases are treated and labeled as incorrect rather than hallucination.

Our annotation protocol enforces this distinction. While LLMs can demonstrate creative linguistic behavior, our benchmark evaluates knowledge of *real, established* idioms and proverbs. Hence, when a model produces a plausible but non-existent expression, we categorize it as a hallucination rather than creativity, as it risks misleading users into believing the expression is authentic. Unlike factual hallucinations—which can often be externally verified—figurative hallucinations demand deeper linguistic and cultural grounding to identify and assess accurately.

The goal of **FFE-Hallu** is to systematically probe figurative hallucination in LLMs through three complementary tasks. Each task targets a distinct dimension of FFE competence (*generation*, *detection*, and *translation*) while collectively measuring the rate and nature of figurative hallucinations. The following sections describe these three tasks in detail.

### 3.1 Task 1: FFE Generation from Meaning

**Motivation** Most prior work on idiom processing has focused on detection or translation rather than generation from meaning. However, generating a valid idiom from its figurative definition constitutes a stronger test of a model's internal idiomatic knowledge: it requires recalling an established, culturally fixed expression rather than constructing one compositionally.

Conventional text generation benchmarks typically assess factual recall or paraphrasing ability, but they do not evaluate whether a model can produce semantically appropriate yet fixed figurative forms. This task introduces a new diagnostic angle by measuring whether a model can retrieve authentic idioms or proverbs that match a given figurative meaning—or instead hallucinates by generating non-existent expressions. In essence, it directly tests the model's ability to balance creative linguistic output with culturally grounded recall, distinguishing genuine figurative competence from figurative hallucination.

**Task Definition** In this task, models are given a concise figurative definition in Persian and asked to produce an idiom or proverb that conveys the same meaning. The key challenge lies in generating an expression that is both culturally valid and semantically appropriate.

To minimize potential data contamination and memorization bias, we evaluate on two types of definitions: (i) human-authored figurative meanings sourced from the Abadis online dictionary (Rezaeimanesh et al., 2025), and (ii) GPT-4o-generated paraphrases verified by human annotators to ensure they convey no additional information beyond the original.

Model outputs are categorized as: (a) **correct**—the generated expression is an authentic idiom or proverb with the intended meaning; (b) **incorrect**—the output is a real FFE but semantically mismatched; and (c) **hallucinated**—the model produces a literal phrase or fabricates a non-existent expression. Table 1 presents representative failure examples.

### 3.2 Task 2: Fake FFE Detection

**Motivation** Inspired by the *NonExistentRefusal* task in HalluLens (Bang, 2023), which examines whether LLMs hallucinate information about non-existent entities, this task targets a fundamentally different phenomenon: *figurative hallucination*.

Rather than assessing factual knowledge, we test whether models can recognize and refrain from fabricating meaning for *synthetic idioms and proverbs*—expressions that are structurally and semantically plausible but do not exist in the target language. This extends hallucination evaluation beyond factual correctness toward probing figurative and cultural understanding.

In contrast to factual verification tasks, where correctness can be checked against external sources, fake FFE detection requires models to internally distinguish authentic idiomatic expressions from

| Failure Type | Real FFE but Incorrect | Hallucinated FFE |
|---|---|---|
| *Input Meaning* | Regretting a wrong action and making a firm decision not to repeat it. | Someone who considers themselves superior to others. |
| *Ground Truth Idiom* | پشت دست خود را داغ کردن<br>*Gloss:* "To burn the back of one's hand"<br>*Meaning:* To strongly vow not to repeat a mistake. | تافته‌ی جدا بافته<br>*Gloss:* "A special thread spun separately"<br>*Meaning:* A person who sees themselves as exceptional. |
| *Model Output* | توبه شکستن<br>*Gloss:* "To break a vow"<br>*Meaning:* Implies repetition of the mistake. | موش توی انبار کاه<br>*Gloss:* "A mouse in a haystack"<br>*Meaning:* - |

Table 1: Examples of model failures in Task 1: FFE Generation. Each example includes the Persian FFE, its English gloss (literal translation), and its figurative meaning.

fabricated yet convincing ones, demonstrating true competence in idiomatic form, meaning, and usage.

**Task Definition** In this task, models are prompted to assess a given FFE and determine whether it is authentic or fabricated. Each model is presented with artificially constructed FFEs designed to closely mimic the syntax, tone, and metaphorical style of real idioms and proverbs.

To evaluate the reliability of rejection behavior, we employ two complementary prompt formats: a *Fake-detection* prompt ("*Is this FFE fake in Persian?*"), where the correct answer is *Yes*, and a *Real-confirmation* prompt ("*Is this FFE real in Persian?*"), where the correct answer is *No*. Each fabricated FFE is tested under both prompt types to examine consistency across framing conditions.

We define the hallucination rate as the proportion of cases in which the model incorrectly accepts a fake FFE as real. Responses are labeled as (i) **correct** if the model detects and rejects the fabricated expression, and (ii) **hallucinated** if it endorses a fake expression as genuine.

### 3.3 Task 3: FFE Translation (English → Persian)

**Motivation** Cross-lingual idiom translation poses a significant challenge for LLMs, as the correct output cannot be achieved through literal, word-for-word translation. Instead, the model must retrieve or generate an equivalent expression that preserves the same figurative meaning while fitting the target culture.

Recent work has begun to tackle this problem; for example, Donthi et al. (2025) improve idiomatic translation by linking idioms across languages. However, such studies primarily focus on identifying correct equivalents and overlook the issue of *figurative hallucination*—cases where a model fabricates plausible but invalid idioms or proverbs.

This task reframes idiom translation as a test of both cultural and figurative competence, assessing whether models can transfer idiomatic meaning across languages without resorting to literal or invented expressions. In doing so, it bridges figurative reasoning and hallucination analysis, highlighting the gap between semantic accuracy and cultural authenticity.

**Task Definition** In this task, models are given an English idiom or proverb that has at least one verified Persian equivalent and are asked to produce a culturally appropriate FFE in Persian. A *hallucination* occurs when the model generates a non-existent idiom in the target language or produces a literal, word-for-word translation. Generating an existing FFE with an incorrect figurative meaning is instead treated as a semantic or translational error.

Model outputs are categorized as: (i) **correct** if the model produces a valid Persian FFE that accurately conveys the intended figurative meaning of the source expression; (ii) **incorrect** if it produces a real but semantically mismatched FFE; and (iii) **hallucinated** if it generates a literal translation or an invented expression that does not exist in Persian.

### 3.4 Dataset Construction

To systematically study figurative hallucination, we introduce **FFE-Hallu**, a 600-item benchmark combining authentic and adversarially constructed cases across the three tasks described in Sections 3.1–3.3.

For **Task 1**, 200 real FFEs were collected from online Persian dictionaries with their figurative

| Task | Source | # Items | Goal |
|---|---|---|---|
| FFE Generation | Authentic Persian idioms/proverbs | 200 | Cultural recall from meanings |
| Fake FFE Detection | Fabricated idioms (4 categories) | 200 | Rejection of hallucinations |
| FFE Translation | English → Persian pairs | 200 | Cross-lingual hallucination test |
| **Total** | — | **600** | — |

Table 2: Dataset composition of **FFE-Hallu**, grouped by task.

meanings (Rezaeimanesh et al., 2025).

For **Task 2**, 200 fabricated FFEs were created to mimic authentic idioms and proverbs while remaining culturally invalid (absent from all major dictionaries and corpora). They fall into four categories: (1) *Semantic Inversion*, (2) *Word Perturbation*, (3) *Structural/Syntactic Mimicry*, and (4) *Cultural/Historical Fabrication*. The first two were derived from real FFEs by reversing meaning or replacing a key word. The latter two were fully synthetic, designed using structural and cultural patterns observed in real Persian proverbs (Zolfaghari and Ameri, 2012). GPT-4o was used only to generate candidate expressions following these templates, and all outputs were manually reviewed and refined by expert annotators to ensure plausibility and non-existence.

For **Task 3**, 200 bilingual English–Persian idiom pairs with verified figurative equivalence were gathered from online resources and validated by native Persian speakers fluent in English.

### 3.5 Annotation Protocol

All data entries and model outputs were reviewed by three native Persian speakers with expertise in figurative language. For authentic FFEs in Task 1, annotators verified the accuracy of definitions and usage examples. For fabricated FFEs in Task 2, they confirmed that none appear in dictionaries or natural usage. Expressions generated by GPT-4o were additionally checked to ensure they closely resemble real idioms and proverbs in form, style, and cultural references, while remaining entirely invalid in Persian. For bilingual pairs in Task 3, annotators verified correct figurative equivalence, proposed improved translations where needed, and removed incorrect pairs. All disagreements were resolved through discussion. Detailed annotation guidelines and inter-annotator agreement scores are reported in Appendix C.

## 4 Experimental Setup and Results

**Models** We evaluate six large language models selected for their multilingual coverage and demonstrated strength on reasoning and generation tasks. We prioritized models with strong overall performance to ensure that observed failures reflect figurative competence rather than general model weakness. The evaluated models include GPT-4.1 (OpenAI, 2025, 2023), Claude 3.7 Sonnet (Anthropic, 2025), Gemma3-27B (DeepMind, 2025), DeepSeek-V3 and DeepSeek-R1 (DeepSeek, 2025), and Qwen3-235B (Group, 2025). Together, these models span both proprietary API-based systems and open-weight releases, as well as a wide range of parameter scales. The parameter counts for each model, as reported by their official documentation, are: GPT-4.1 (undisclosed), Claude 3.7 Sonnet (undisclosed), Gemma3-27B (27B), DeepSeek-V3 (236B), DeepSeek-R1 (67B), Qwen3-235B (235B).

**Computational Resources** All LLM outputs were obtained via official API endpoints. Across all experiments, total API usage amounted to approximately $25 USD. All evaluations were performed on the cloud using OpenAI and Anthropic API services; no additional GPU or local compute resources were required.

**Prompts** All models are queried with consistent, task-specific prompt templates to ensure comparability across tasks. Prompts are designed to be concise, neutral in style, and free of additional examples beyond what is necessary for task clarification. For Task 1: FFE generation, the input is always a concise figurative meaning in Persian; for Task 2: Fake FFE Detection, a fake FFE; and for Task 3: FFE to FFE translation, an English idiom or proverb that has at least a known Persian equivalent. Full prompt templates for each task are provided in Appendix A.

| Model | Human-Written Meaning | | | LLM-Paraphrased Meaning | | |
|---|---|---|---|---|---|---|
| | Correct(%) ↑ | Incorrect(%) | Hallucination(%) ↓ | Correct(%) ↑ | Incorrect(%) | Hallucination(%) ↓ |
| GPT-4.1 | **60.5** | 19.0 | 20.5 | 59.5 | 23.5 | 17.0 |
| Claude 3.7 Sonnet | 50.0 | 35.0 | 15.0 | 53.0 | 33.0 | **14.0** |
| Gemma3-27B | 22.5 | 38.0 | 39.5 | 30.5 | 35.5 | 34.0 |
| DeepSeek-R1 | 48.5 | 24.5 | 27.0 | 47.0 | 27.5 | 25.0 |
| DeepSeek-V3 | 17.5 | 40.5 | 42.0 | 18 | 24.5 | 47.5 |
| Qwen3-235B | **7.0** | 23.5 | **69.5** | 10.0 | 21.0 | 69.0 |

Table 3: Performance of LLMs on Task 1 described in 3.1: FFE Generation, from two input types: Human-Written Meanings and LLM-Paraphrased Meanings. Metrics include correct idiom generation percentage, incorrect generation, and hallucination rate. Higher Correct(%) and lower Hallucination(%) are desirable.

**Evaluation Protocol** To assess model performance, we employed both automatic and manual evaluation methods. For Tasks 1 and 2, described in Section 3.1 and 3.2, all model outputs are annotated by two native Persian speakers with expertise in figurative language. A third annotator is consulted in cases of disagreement, and consensus is reached through discussion. For these two task we considered three labels: correct, incorrect and hallucinated. The correct label show the accuracy of models to generate or translate an authentic FFE. Other cases are divided into two groups: incorrect for the authentic FFEs that do not reserve the intended meanings and hallucinated for non-existent expressions. For Task 2: Fake FFE Detection, the outputs of models are automatically inferable based on their answers that were "NO" or "YES". If the model correctly identifies a fake FFE as fake, the output is labeled as correct; if it incorrectly accepts a fake FFE as real, it is labeled as hallucinated.

### 4.1 Results

**Task 1: FFE Generation from Meaning** Table 3 shows the results for Task 1. GPT-4.1 achieves the highest accuracy, producing valid idioms for about 60% of inputs with relatively low hallucination rates (17–21%). Claude 3.7 Sonnet performs slightly worse in accuracy but remains the most cautious, with the lowest hallucination rate (14%). In contrast, Gemma3-27B and DeepSeek-V3 generate valid idioms in fewer than one third of cases and frequently hallucinate, while Qwen3-235B exhibits severe overgeneration, hallucinating in nearly 70% of paraphrased cases. Interestingly, paraphrased meanings generally reduced hallucination rates, suggesting that rephrasing helped models access relevant idiomatic knowledge, even when their outputs were incorrect.

**Task 2: Fake FFE Detection** Table 4 and Figure 1 summarize model performance on identifying non-existent FFEs, evaluated under two prompt settings: *Fake-detection* and *Real-confirmation*. The hallucination rate, defined as the proportion of false acceptances, measures how often a model incorrectly classifies a fake FFE as real.

Results reveal strong prompt sensitivity across models: performance is consistently lower under the Fake-detection prompt than under Real-confirmation. While most models perform near-perfectly when asked "Is this real?", their accuracy drops sharply for "Is this fake?", exposing instability in figurative hallucination detection. GPT-4.1 achieves the best results, with the lowest average hallucination rate (14.5%) and highest cross-prompt agreement (86%), indicating consistent rejection across question framings. In contrast, DeepSeek-V3, Claude 3.7, and Qwen3-235B show pronounced discrepancies between prompts—performing well under Real-confirmation but poorly under Fake-detection—leading to higher hallucination rates and reduced agreement.

Interestingly, Gemma3-27B shows relatively high agreement (75.5%) for undesirable reasons: it consistently misclassifies fake FFEs as real, resulting in a high hallucination rate of 83.25%. As shown in Figure 1, across all models and prompts, "Word Perturbation" and "Cultural/Historical Fabrication" types are generally handled better than "Semantic Inversion" and "Structural/Syntactic Mimicry." Models such as DeepSeek-V3 exhibit a sharp performance decline—nearly 60%—between prompt types (agreement 38%), suggesting bias toward question framing rather than genuine recognition of authenticity. These findings underscore the importance of using both prompt types and

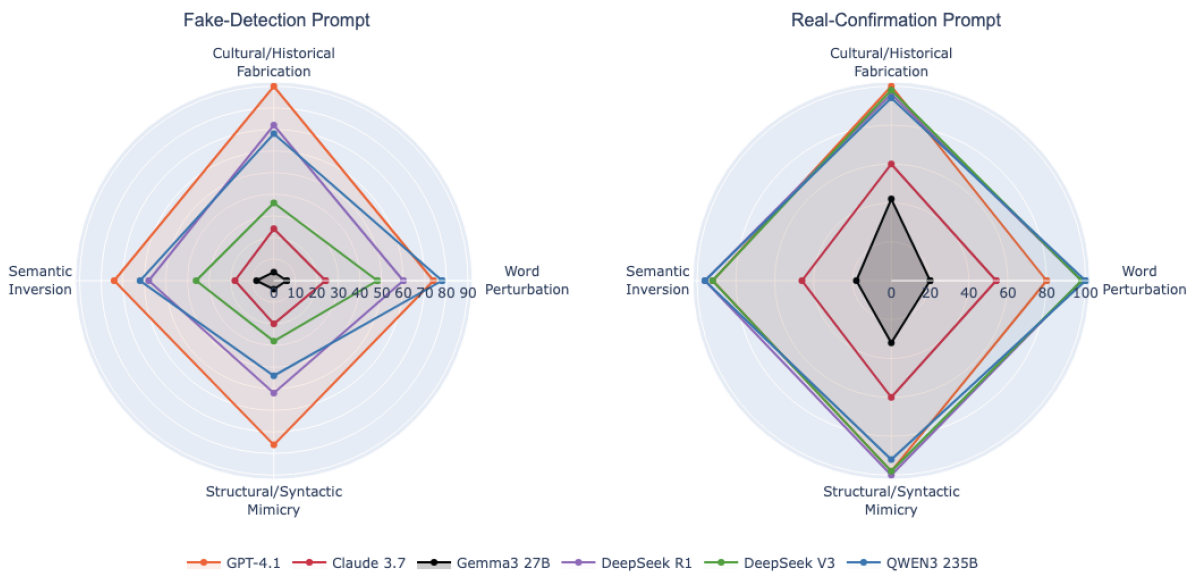


Figure 1: Fake FFE Detection accuracy by model and category for two prompt settings.

| Model | Prompt-Fake Halluc.(%) | Prompt-Real Halluc.(%) | Avg(%) | Agreement(%) |
|---|---|---|---|---|
| GPT-4.1 | **21.50** | 7.50 | **14.50** | **86.00** |
| Claude 3.7 | 78.50 | 45.00 | 61.75 | 63.50 |
| DeepSeek-R1 | 39.50 | **2.50** | 21.00 | 61.00 |
| DeepSeek-V3 | 63.00 | 3.50 | 33.25 | 38.50 |
| Gemma3-27B | 94.50 | 72.00 | 83.25 | 75.50 |
| Qwen3-235B | 37.00 | 4.50 | 20.75 | 65.50 |

Table 4: Task 2: Fake FFE Detection described in Section 3.2, Hallucination Rates and Agreement Metrics Across Prompts and Models. **Prompt-Fake**: "*Is this fake?*" (correct answer: YES). **Prompt-Real**: "*Is this real?*" (correct answer: NO). Hallucination rate is defined as the proportion of **false acceptances**, which means how many of these fake FFEs are incorrectly accepted as real. Lower is better.

cross-prompt agreement as key measures of model reliability.

| Model | Correct | Incorrect | Hallucinated |
|---|---|---|---|
| GPT-4.1 | **65.0** | 10.0 | **25.0** |
| Claude 3.7 | 60.0 | 9.0 | 31.0 |
| DeepSeek-R1 | 32.0 | 12.0 | 56.0 |
| DeepSeek-V3 | 30.5 | 7.5 | 62.0 |
| Gemma3-27B | 15.0 | 9.0 | 76.0 |
| Qwen3-235B | **4.5** | 1.5 | **94.0** |

Table 5: Results of Task 3: FFE → FFE Translation described in 3.3. Correct is the acuuracy of every model in detecting fake FFEs. Accepting any fake FFE as real is considered hallucination.

**Task 3: FFE Translation (English → Persian)** As shown in Table 5, the FFE-to-FFE translation results reveal substantial variation across models in both accuracy and hallucination behavior. GPT-4.1 achieved the highest accuracy (65%) and the lowest hallucination rate (25%), followed by Claude 3.7, which was fluent but prone to figurative distortions (31% hallucination). Both models demonstrated partial understanding of Persian figurative language but often translated idioms too literally instead of retrieving culturally equivalent expressions.

DeepSeek-R1 and DeepSeek-V3 reached moderate accuracy ( 30%) yet hallucinated in over half of their outputs, frequently merging fragments of real FFEs or fabricating plausible but invalid phrases. Gemma3-27B showed limited grasp of idiomatic structure, using culturally familiar words but mis-

ordering or substituting them, thus breaking fixed idiomatic forms. In contrast, Qwen3-235B produced mostly fabricated or ungrammatical outputs, reflecting minimal knowledge of Persian idioms and culture.

Overall, models differed not only in accuracy but also in the *types* of figurative hallucination—ranging from literal substitutions and structural distortions to fully invented expressions—highlighting persistent gaps in both linguistic and cultural competence.

**Cross-Task Analysis** Across all three tasks, GPT-4.1 achieved the highest accuracy and lowest hallucination rates, showing consistent performance marked by strong linguistic fluency and cultural awareness. Claude 3.7 Sonnet also demonstrated the ability to generate valid FFEs in both generation (Task 1) and translation (Task 3), but its performance dropped notably in fake FFE detection (Task 2), where it was more prone to accept fabricated expressions as real.

In contrast, DeepSeek-R1 and DeepSeek-V3 exhibited only partial understanding of idiomatic and cultural patterns. They frequently combined or distorted existing expressions, producing outputs that appeared idiomatic but were structurally or semantically invalid. Gemma3-27B and Qwen3-235B performed weakest overall, often generating or accepting entirely fabricated expressions and showing limited awareness of Persian figurative conventions.

Interestingly, Qwen3-235B showed relatively strong performance in rejecting fake FFEs, achieving high agreement in the detection task. However, in generation and translation, it almost never produced literal or incorrect real FFEs; instead, it consistently hallucinated plausible but non-existent ones. This pattern suggests that while the model can identify non-genuine FFEs, it lacks the cultural and lexical grounding needed to r

These results demonstrate that competence in generating FFEs does not necessarily entail strong detection ability, revealing key differences in how models approach generation versus verification of figurative language. Figurative hallucination thus appears systematic rather than random, reflecting each model's depth of cultural and semantic understanding. Models with stronger cultural grounding achieve higher accuracy and consistency across tasks, whereas those with weaker grounding hallucinate more frequently.

## 5 Conclusion

This work introduced FFE-Hallu, the first benchmark for evaluating figurative hallucination in LLMs, with a focus on fixed figurative expressions (FFEs) in Persian. The benchmark comprises three carefully designed datasets, each targeting a distinct aspect of figurative competence—generation, detection, and translation—enabling a systematic analysis of how models handle figurative meaning across contexts.

Experimental results show that GPT-4.1 was the most reliable model overall, followed by Claude 3.7 Sonnet, though both exhibited subtle cultural and literal hallucinations. In contrast, open-weight models such as DeepSeek-R1, DeepSeek-V3, Gemma3-27B, and Qwen3-235B frequently overgenerated and failed to reject fabricated FFEs, revealing persistent gaps in cultural and linguistic grounding.

Overall, FFE-Hallu underscores that mastery of figurative language requires more than multilingual fluency—it demands genuine cultural understanding. We hope this benchmark facilitates future research toward hallucination-resistant and culturally aware language models.

## 6 Limitations

A central limitation of this study is that detecting figurative hallucinations often requires the intuition and judgment of native speakers. Many of the fabricated FFEs in our dataset were intentionally designed to closely resemble authentic expressions in structure, rhythm, or appearance. Distinguishing these subtle hallucinations from genuine FFEs was only possible for native Persian speakers with strong cultural grounding. Because our annotators were all native Persian speakers, we restricted the benchmark to Persian. This choice ensured reliable annotation quality but necessarily limits the scope of our conclusions to a single language.

The dataset size is also modest (600 items), reflecting the difficulty of curating and validating idioms at scale. While sufficient for diagnostic evaluation, broader coverage is needed to capture the full range of figurative phenomena.

## References


Anthropic. 2025. Claude 3.7 sonnet release notes.

Yejin et al. Bang. 2023. Hallulens: A lens for measuring hallucination in instruction-tuned llms. In *NAACL*.

Nesrine Bendahman, Karen Pinel-Sauvagnat, and Mohand Boughanem. 2025. Not all hallucinations are good to throw away when it comes to legal abstractive summarization. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Google DeepMind. 2025. Gemma 3: Multimodal multilingual models up to 27b parameters. *arXiv preprint*.

DeepSeek. 2025. Deepseek v3 and r1 model overview.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving llm abilities in idiomatic translation. *Preprint*, arXiv:2407.03518. Preprint for LoResLM Workshop at COLING 2025.

Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993–1002.

Hui Gao, Jing Zhang, Peng Zhang, and Chang Yang. 2025. Consistency rating of semantic transparency: An evaluation method for metaphor competence in idiom understanding tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE. Association for Computational Linguistics.

Alibaba Group. 2025. Qwen 3 technical report: Large-scale multilingual models with moe. *arXiv preprint*.

Zhewei et al. Ji. 2023. Survey of hallucination in natural language generation. In *TACL*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2024. Comparative study of multilingual idioms and similes in large language models. *arXiv preprint arXiv:2410.16461*.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *arXiv preprint arXiv:2308.13961*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. *Preprint*, arXiv:2410.16069.

OpenAI. 2025. Gpt-4.1 technical overview.

Steven Adler Sandhini Agarwal Lama Ahmad Ilge Akkaya Florencia Leoni Aleman Diogo Almeida Janko Altenschmidt Sam Altman Shyamal Anadkat Red Avila Igor Babuschkin Suchir Balaji Valerie Balcom Paul Baltescu Haiming Bao Mohammad Bavarian Jeff Belgum Irwan Bello Jake Berdine Gabriel Bernadett-Shapiro Christopher Berner Lenny Bogdonoff Oleg Boiko Madelaine Boyd Anna-Luisa Brakman Greg Brockman Tim Brooks Miles Brundage Kevin Button Trevor Cai Rosie Campbell Andrew Cann Brittany Carey Chelsea Carlson Rory Carmichael Brooke Chan Che Chang Fotis Chantzis Derek Chen Sully Chen Ruby Chen Jason Chen Mark Chen Ben Chess Chester Cho Casey Chu Hyung Won Chung Dave Cummings Jeremiah Currier Yunxing Dai Cory Decareaux Thomas Degry Noah Deutsch Damien Deville Arka Dhar David Dohan Steve

Dowling Sheila Dunning Adrien Ecoffet Atty Eleti Tyna Eloundou David Farhi Liam Fedus Niko Felix Simón Posada Fishman Juston Forte Isabella Fulford Leo Gao Elie Georges Christian Gibson Vik Goel Tarun Gogineni Gabriel Goh Rapha Gontijo-Lopes Jonathan Gordon Morgan Grafstein Scott Gray Ryan Greene Joshua Gross Shixiang Shane Gu Yufei Guo Chris Hallacy Jesse Han Jeff Harris Yuchen He Mike Heaton Johannes Heidecke Chris Hesse Alan Hickey Wade Hickey Peter Hoeschele Brandon Houghton Kenny Hsu Shengli Hu Xin Hu Joost Huizinga Shantanu Jain Shawn Jain et al. (181 additional authors not shown) OpenAI, Josh Achiam. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774. ArXiv preprint arXiv:2303.08774.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M. Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

Hassan Zolfaghari and Hayat Ameri. 2012. Persian proverbs: Definitions and characteristics. *Journal of Islamic and Human Advanced Research*, 2:93–108. Accessed 2025-10-05.

# A Prompt Templates

## 1.1 Task 1: FFE Generation from Meaning

**System Prompt**

You are an expert on Persian fixed figurative expressions (FFEs)—idioms and proverbs. When given a figurative meaning in Persian, output *one* culturally valid Persian idiom or proverb that best matches the meaning.

**User Prompt Template**

**Definition of FFE (for the model):** Fixed Figurative Expressions are fixed, often indirect combinations of words whose meanings cannot be directly inferred from their parts. They include idioms and proverbs.
**Examples: Examples:** 1) Idiom: *[IDIOM_EX1]* → Meaning: *[MEANING_EX1]*
2) Proverb: *[PROVERB_EX2]* → Meaning: *[MEANING_EX2]*
**Your task:** The following phrase is the figurative meaning of a valid Persian idiom or proverb. Output a Persian idiom or proverb that aligns with this meaning. Output only the expression.
**Meaning (Persian):** *[PUT_MEANING_HERE]*

## 1.2 Task 2: Fake FFE Detection

**System Prompt**

You are a careful verifier of Persian FFEs (idioms/proverbs). Given a candidate expression, decide if it is an authentic, attested FFE in Persian. If it is real, output:
`Yes`
`Meaning: <concise Persian gloss>`
`Example: <a short Persian sentence>.`
If it is not real, output only: `No`.

**User Prompt Template**

**Question:** Is the following an official and real idiom/proverb in Persian? If yes, provide its meaning and an example. If not, just say `No`.
**Candidate expression:** *[PUT_EXPRESSION_HERE]*

## 1.3 Annotation Prompt (for Human Guidelines)

**Annotator Instructions (summary)**

Label each model output for Task 1 as: `Correct` (valid FFE matching the meaning), `Incorrect` (valid FFE but mismatched meaning, or literal phrase), or `Hallucinated` (not an attested Persian FFE). For disagreements, discuss and reach consensus.

# B Dataset Creation

## 2.1 Fake Idioms Dataset Creation

We created 200 fake idioms that are not real expressions in Persian but are designed to look or sound like authentic idioms in Persian. These fake idioms were manually created by 2 native Farsi speakers and also the Model in the loop method using GPT4o. These fake idioms were divided into four distinct categories based on how they were constructed and the specific reasoning vulnerabilities they are meant to test:

1. **Word Perturbation**
   We took a real idiom and changed just one important word, for example, replacing one animal or object or any entity with another. This checks if the model notices small changes that make the idiom incorrect.

   **Original Idiom**

   **Persian Idiom:** ماه پشت ابر نمی‌ماند
   The **moon** won't stay behind the clouds.

   **Idiom after Word Perturbation**

   **Fake Idiom:** خورشید پشت ابر نمی‌ماند
   The **sun** won't stay behind the clouds.

2. **Semantic Inversion**
   We kept the structure of the idiom but flipped its meaning. For example, turning "a wise enemy is better than a foolish friend" into "a foolish friend is better than a wise enemy." This tests if the model understands the logic behind the idiom.

   **Original Idiom**

   **Persian Idiom:** دشمن دانا به از دوست نادان
   A wise enemy is better than a foolish friend.

   **Semantically Inverted Idiom**

   **Fake Idiom:** دوست نادان به از دشمن دانا
   A foolish friend is better than a wise enemy.

3. **Structural/Syntactic Mimicry**
   By prompting GPT4o, we created fake idioms that structure and rhyme like real Per-

sian FFEs, but don't have any true meaning or background. This checks if the model is tricked just because the fake FFEs "sounds like real idioms/proverbs!"

4. **Cultural/Historical Fabrication**
Every language has its own unique cultural elements that are commonly used by the speakers of that language. For example, in Persian proverbs, there are so many cultural elements like "faith". Again, by prompting GPT4o, we invented FFEs that refer to well-known elements, characters or stories from Persian culture, with their presence in a completely fake event or story. These may sound to LLMs cultural and believable, but are completely fictional. This tests if the model knows Persian culture or not.

### 2.2 Why These Categories Matter

In the evaluation phase of Task 1: Idiom Generation, we observed a range of hallucination patterns in the outputs of LLMs. Notably, models would sometimes produce variants of well-known idioms by altering a single word, subtly shifting their meaning. For example, the classical Persian idiom from Golestan Saadi,

<The teacher's strictness is better than the father's love>, while the model generates: <The teacher's beating is better than the father's love>. In this case, the model changes the word "Strictness" to "Beating", distorting the intended moral nuance and cultural acceptability of the idiom. This suggests that the model may be relying on surface-level patterns or templates, rather than deeply understanding the idiom's semantics and cultural grounding.

Motivated by such observations, we designed a second task to probe how well LLMs can distinguish real idioms from plausible but fake ones. Specifically, we created four types of adversarial examples, inspired by the kinds of generation errors observed in Task 1. These were grouped into two broader categories:

- **Template-based perturbations:** First two groups, Word Perturbation and Semantic Inversion, check if the model catches small edits that break the idiom structure or meaning, or if it just relies on the template.

- **Fabricated idioms with surface plausibility:** The Other two groups, Structural/Syntactic Mimicry and Cultural/Historical Fabrication, check if the model is fooled by structure or rhythm or fake cultural Narrations.

By evaluating model performance across these categories, we can better understand whether LLMs truly grasp idiomatic meaning or simply mimic patterns. This diagnostic task sheds light on their linguistic and cultural grounding, particularly in low-resource, high-context settings like Persian FFEs.

## C Inter-Annotator Agreements

| Model | Agreement (%) | Cohen's $\kappa$ |
|---|---|---|
| GPT-4.1 | 91.5 | **0.84** |
| Claude 3.7 Sonnet | 86.5 | 0.82 |
| DeepSeek-R1 | 88.0 | 0.78 |
| DeepSeek-V3 | 83.5 | 0.72 |
| Gemma3-27B | 87.5 | 0.66 |
| Qwen3-235B | **94.0** | 0.58 |

Table 6: Task 1 Inter-annotator agreement for human evaluation across three labels (Correct, Incorrect, Hallucination). Raw agreement measures identical labels; Cohen's $\kappa$ adjusts for chance agreement

| Model | Agreement (%) | Cohen's $\kappa$ |
|---|---|---|
| GPT-4.1 | 90.16 | 0.803 |
| Claude 3.7 Sonnet | 91.19 | 0.836 |
| DeepSeek-R1 | 92.23 | **0.866** |
| DeepSeek-V3 | 88.08 | 0.773 |
| Gemma3-27B | 89.12 | 0.703 |
| Qwen3-235B | **95.85** | 0.584 |

Table 7: Task 3 Inter-annotator agreement for human evaluation across three labels (Correct, Incorrect, Hallucination). Raw agreement is the percentage of identical labels, while Cohen's $\kappa$ adjusts for chance agreement.