Arabic Humor as a Diagnostic Probe for Large Language Models

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Arabic humor provides a challenging diagnostic test for large language models because interpreting jokes often requires pragmatic inference, sociolinguistic awareness, and culturally grounded knowledge that standard NLP benchmarks do not evaluate. Arabic is particularly suitable for probing these abilities given its diglossic structure and dialect diversity, where humor frequently arises from register contrast, dialect-specific vocabulary, and shared cultural references. We propose a three-layer taxonomy of Arabic humor mechanisms covering pragmatic, semantic, and sociolinguistic phenomena, illustrated through thirteen curated examples spanning Egyptian, Levantine, Gulf, Tunisian, and Iraqi Arabic. Building on this taxonomy, we introduce a diagnostic evaluation framework using contrastive minimal pairs, a multi-dimensional scoring rubric, and a cultural presupposition ontology. A small proof-of-concept probing study with GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5 reveals recurring failure patterns in sarcasm interpretation, register contrast reasoning, dialectal vocabulary coverage, and cultural grounding. We position this work as a diagnostic framework and pilot, not a mature benchmark, and outline a path toward larger annotated resources.

1 Introduction

A good diagnostic test for large language model (LLM) competence should probe capabilities that standard benchmarks leave untested and produce failures that are informative about specific, identifiable gaps. We argue that Arabic humor satisfies both criteria. Interpreting a joke in Arabic requires pragmatic inference, diglossic register reasoning, and culturally grounded knowledge (Raskin, 1985), three dimensions that reasoning, knowledge, and safety benchmarks do not directly evaluate. When a model fails on Arabic humor, the failure reveals whether it lacks pragmatic inference, dialect-register awareness, or cultural background knowledge.

LLMs achieve impressive results across NLP benchmarks, yet humor remains a persistent challenge. Jentzsch and Kersting (2023) found that over 90% of ChatGPT-generated jokes were drawn from just 25 templates, revealing template reproduction rather than genuine understanding. Sarcasm detection similarly shows LLMs underperforming fine-tuned models when meaning depends on pragmatic context (Abu Farha et al., 2022).

Arabic makes the diagnostic sharper. Its diglossic structure (Ferguson, 1959) creates humor from the contrast between Modern Standard Arabic (MSA) and colloquial dialectal punchlines, a mechanism absent from English-centric humor research. Despite major Arabic NLP resources for dialect identification and sarcasm detection (Abu Farha and Magdy, 2020; Abdul-Mageed et al., 2021; Bouamor et al., 2018), humor mechanisms beyond binary sarcasm remain understudied. Arabic LLMs exhibit "pragmatic literalism" and cultural context gaps even in MSA settings (Al-Olimat and Alshareef, 2026); our work extends this into the more demanding domain of dialectal humor.

We position this paper deliberately as a diagnostic framework and pilot study rather than a mature benchmark. The thirteen examples and the small probing study should be read as a proof of concept showing how Arabic humor can surface specific LLM failure modes, not as a finalised resource for model comparison. Our contributions are: (1) a three-layer taxonomy of Arabic humor mechanisms cov-

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 39–50

ering pragmatic, semantic, and sociolinguistic phenomena; (2) thirteen curated examples across five dialect regions with explicit mechanism assignment criteria; (3) a probing study on GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5 demonstrating four failure categories; (4) a worked example of contrastive minimal pairs; and (5) a concrete evaluation protocol for future corpus development.

2 Related Work

2.1 Theories of Humor and Computational Approaches

Raskin (1985) introduced the Semantic Script Theory of Humor (SSTH): a text is humorous when compatible with two opposing semantic scripts and a trigger switches interpretation between them. Attardo and Raskin (1991) extended this into the General Theory of Verbal Humor (GTVH), adding knowledge resources for narrative strategy, logical mechanism, target, situation, and language. Our taxonomy organises mechanisms by the type of reasoning required to recover the script switch and is one operationalisation of GTVH's logical-mechanism and language resources for an NLP setting, not a claim that these three layers exhaust humor theory.

On the computational side, Mihalcea and Strapparava (2005) established humor recognition as a tractable NLP task; Miller et al. (2017) provided the first shared evaluation for pun detection in English, demonstrating that semantic ambiguity could be modelled at scale; Amin and Burghardt (2020) survey humor generation and identify the lack of theoretically grounded evaluation as a central weakness; and Jentzsch and Kersting (2023) confirmed that LLMs still rely on surface templates rather than genuine humor understanding, with over 90% of ChatGPT-generated jokes drawn from just 25 recurring patterns.

2.2 Arabic Pragmatic Benchmarks and Sarcasm Detection

Al-Olimat and Alshareef (2026) developed the Arabic Linguistic and Pragmatic Suite (ALPS), showing that Arabic LLMs exhibit systematic pragmatic literalism and fail to recover culturally specific meaning even in MSA. This directly motivates extending diagnostic evaluation to dialectal humor, where pragmatic and cultural demands compound. Abu Farha and Magdy (2020) constructed ArSarcasm by reannotating sentiment data; their BiLSTM baseline reached only F1 = 0.46, illustrating the difficulty of even binary sarcasm classification. The iSarcasmEval shared task (Abu Farha et al., 2022) confirmed the challenge persists with MARBERT (Abdul-Mageed et al., 2021), because dialectal variation interacts with ironic meaning in ways that pretraining does not address.

2.3 Dialect Identification and Multi-Task Arabic NLP

NADI (Abdul-Mageed et al., 2020) and MADAR (Bouamor et al., 2018) established dialect identification as tractable, but identifying a dialect label is a prerequisite for humor interpretation, not an end in itself: a system that labels an utterance as Egyptian Arabic but cannot reason about the pragmatic significance of a variety switch has not solved the relevant problem. Kaseb and Farouk (2022) introduced SAIDS, showing that informing sentiment analysis with dialect and sarcasm predictions improves overall sentiment performance, supporting our proposal for humor annotation that captures dialect identity, mechanism, and cultural presupposition simultaneously.

2.4 Arabic Diglossia and Code-Switching

Ferguson (1959) defined diglossia as the coexistence of a formal High (H) and informal Low (L) variety, with Arabic as the prototypical case. Code-switching humor is documented across many language pairs (Winata et al., 2023), but what is structurally distinctive in Arabic is that the H/L distinction is formalised through diglossia and available to all speakers as a shared, unmarked communicative resource. This salience makes register switching a default humor mechanism in Arabic rather than an individual identity marker.

3 Humor in Arabic: A Taxonomy

Drawing on SSTH and GTVH (Raskin, 1985; Attardo and Raskin, 1991) and on the examples in Section 4, we propose a taxonomy organising Arabic humor mechanisms into three

computationally relevant layers, defined by the type of reasoning required to interpret the humor: a pragmatic layer requiring inference about speaker intent, a semantic layer requiring resolution of lexical or conceptual ambiguity, and a sociolinguistic layer requiring shared awareness of variety norms and cultural presuppositions. Table 1 maps each mechanism to its primary linguistic trigger and computational challenge.

Theoretical grounding of mechanism names. The mechanism names in our taxonomy are not arbitrary. *Sarcasm*, *irony*, and *self-deprecation* are standard categories in the humor research literature and in Arabic-language NLP work on sarcasm (Abu Farha and Magdy, 2020; Abu Farha et al., 2022). *Expectation inversion* and *wordplay* correspond to what SSTH calls script opposition triggered by lexical or narrative pivots, where the punchline forces reinterpretation of the setup. *Metaphorical humor* and *hyperbole* are recognised cross-domain mapping and implausibility mechanisms within GTVH's logical-mechanism resource. *Dialectal lexical humor*, *register switching*, and *cultural references* correspond to what GTVH groups under the language and situation knowledge resources, refined here to reflect Arabic-specific phenomena documented in the dialect identification (Abdul-Mageed et al., 2020; Bouamor et al., 2018) and code-switching (Winata et al., 2023) literatures. This mapping makes the taxonomy traceable to established humor theory rather than emergent from the example set alone.

Mechanism assignment criteria. Each example is assigned a primary mechanism according to the dominant trigger. When multiple triggers are present (as in Example 7, where register switching co-occurs with a cultural reference), we assign the mechanism that is *necessary and sufficient* for the humorous effect: removing it while preserving propositional content destroys the humor, as Section 5 illustrates. A second trigger that amplifies but does not constitute the humor is noted as secondary. Future annotators working with broader corpora should flag items that resist single-layer assignment for adjudication.

Mechanism	**Linguistic Trigger**	**Computational Challenge**
Sarcasm / Irony	Polarity reversal	Pragmatic inference beyond literal polarity
Self-deprecation	Expectation gap	Narrative expectation tracking
Expectation inversion	Genre frame violation	Discourse-type world knowledge
Wordplay	Lexical ambiguity	Ambiguity resolution
Metaphorical humor	Cross-domain mapping	Conceptual metaphor detection
Hyperbole	Implausible quantity	Implausibility detection
Dialectal lexical	Dialect-specific cue	Dialect vocabulary coverage
Register switching	MSA to dialect shift	Sociolinguistic register modeling
Cultural references	Shared presupposition	Cultural background knowledge

Table 1: Taxonomy of Arabic humor mechanisms with primary linguistic triggers and computational challenges.

The three layers. *Pragmatic humor* arises when intended meaning differs from literal content and encompasses sarcasm and ironic praise, self-deprecating humor, and expectation inversion. *Semantic humor* exploits lexical or conceptual ambiguity and covers wordplay and lexical reframing, metaphorical humor (cross-domain mappings), and hyperbole. *Sociolinguistic humor* draws on shared awareness of language-use norms and includes dialectal lexical humor (variety-specific vocabulary), register switching (MSA setup, colloquial punchline), and cultural reference humor (food, religious practices, seasonal events, regional customs).

4 Analysis of Arabic Humor Examples

We analyse thirteen humorous expressions drawn from publicly circulating Arabic social media content (Twitter/X posts and widely shared memes from 2023 and 2024). Selection required that each example instantiate a distinct primary mechanism, be interpretable without speaker identity information, and that examples span at least five dialect regions. Each example was verified by the author and one additional Arabic-speaking colleague fa-

miliar with the relevant dialect.

4.1 Pragmatic Humor Examples

Example 1, sarcastic praise (Egyptian):
ما شاء الله يا عبقري...كسرت الكوب بعد ما قلتلك تمسكه كويس
"Mashallah, you are a genius…you broke the cup after I told you to hold it properly."
Primary mechanism: Sarcasm. *Masha'allah*, conventionally an expression of admiration, is deployed sarcastically after a careless error. Lexical polarity alone assigns positive sentiment, missing the context-dependent inversion. This example carries a secondary cultural dimension (the religious formula's broad pragmatic range in Arabic), but the primary trigger is polarity inversion rather than register switching.

Example 2, expectation inversion (Tunisian):
عملت ريجيم جمعة...نقصت 200 غرام من صبري
"I did a week's diet…I lost 200 grams of my patience."
Primary mechanism: Expectation inversion. The setup primes physical weight loss, then the punchline substitutes an emotional quantity, violating the semantic type constraint between mass entity and psychological state. All three models scored at or near zero on this example, making it the clearest instance of pragmatic inference failure in the dataset.

Example 3, ironic self-evaluation (Levantine):
اشتريت كتاب تطوير الذات...ولسه نفس الشخص
"I bought a self-improvement book…I am still the same person."
Primary mechanism: Sarcasm. The humor exploits the contrast between self-help genre promises and the complete absence of change, requiring pragmatic inference about discourse-type conventions rather than any textual cue of irony.

4.2 Semantic Humor Examples

Example 4, lexical reframing (Egyptian):
أنا مش بخيل...أنا اقتصادي
"I am not stingy…I am economical."
Primary mechanism: Lexical reframing. The negative term *bakhīl* is replaced with the positive *iqtisādī*. The humor lies in the transparent self-justification: the denial implicitly confirms the original characterisation.

Example 5, conceptual metaphor (Egyptian):
شغّل دماغك شوية
الدماغ: تم إيقاف الخدمة مؤقتاً
"Use your brain a little."
"Brain: service temporarily unavailable."
Primary mechanism: Metaphorical humor. The brain-as-tool metaphor is extended into the telecommunications service-notification register, producing cross-domain incongruity.

Example 6, hyperbole (Gulf):
الحر اليوم كأنه الشمس نازلة تشرب قهوة معنا
"The heat today is as if the sun came down to drink coffee with us."
Primary mechanism: Hyperbole. Extreme heat is personified through a domestic hospitality scenario specific to Gulf cultural norms.

4.3 Sociolinguistic Humor: Register Switching

Register switching is one of the most structurally distinctive Arabic humor sources: a formal MSA setup is deflated by a colloquial punchline, requiring models to treat language variety as a meaning-bearing dimension rather than a classification label (Abdul-Mageed et al., 2020; Bouamor et al., 2018). Current Arabic NLP systems can detect variety switches but are not equipped to reason about their pragmatic implications. The pilot results in Section 6 confirm this gap: Examples 7 and 8 produce the sharpest inter-model divergence in the dataset.

Example 7, MSA setup with Egyptian dialect punchline:
هل تفضّل تناول وجبة صحيّة؟
لا يا عم هاتلي كشري وخلاص
Setup (MSA): "Would you prefer to consume a healthy meal?"
Punchline (Egyptian): "No man, just bring me koshary and that's it."
Primary mechanism: Register switching. *Secondary:* cultural reference (koshary as Egyptian working-class street food). The register switch is necessary and sufficient; the food reference amplifies but does not constitute the humor.

Example 8, MSA setup with Levan-

tine punchline:
أرجو أن تتحلّى بالصبر
الصبر خلص من زمان
Setup (MSA): "Please adorn yourself with patience."
Punchline (Levantine): "Patience ran out long ago."
Primary mechanism: Register switching. A classical MSA rhetorical construction is undercut by colloquial Arabic treating patience as a depleted commodity.

Example 9, cultural reference (Gulf):
قالوا نعمل دايت
قلت بعد رمضان
"They said let's diet."
"I said: after Ramadan."
Primary mechanism: Cultural reference. "After Ramadan" signals indefinite deferral only if the listener knows Ramadan is immediately followed by Eid feasting, making the stated timing a humorous non-commitment rather than a concrete plan. All three models scored 1.3 here, the most consistent result among sociolinguistic examples, suggesting partial cultural knowledge of Ramadan but without full grounding in the deferral convention.

4.4 Additional Examples: Broadening Dialect Coverage

Example 10, sarcastic praise (Levantine):
يعني عنجد فكرة عبقرية...خلينا نأجل الشغل لآخر دقيقة
"Wow, truly a brilliant idea…let's postpone the work until the last minute."
Primary mechanism: Sarcasm. The positive framing introduces an obviously counterproductive plan, requiring contextual reasoning to invert surface polarity.

Example 11, expectation inversion (Egyptian):
ذاكرت طول الليل...عشان أنام في الامتحان مرتاح
"I studied all night…so I could sleep comfortably during the exam."
Primary mechanism: Expectation inversion. Academic performance is primed, then the punchline replaces the goal (passing) with its antithesis (comfortable sleeping).

Example 12, lexical reframing (Gulf):
أنا مو كسول...أنا موفّر للطاقة
"I am not lazy…I am energy efficient."
Primary mechanism: Lexical reframing. The negative *kasūl* is relabelled with the technical-environmental *muwaffir lil-ṭāqa*, an incongruous register shift that humorously rebrands a mundane failing.

Example 13, cultural reference (Iraqi):
قررت أبدأ حمية من اليوم...بس اليوم عزيمة
"I decided to start dieting today…but today we have a feast invitation."
Primary mechanism: Cultural reference. The joke depends on the norm of *'azīma*, a formal meal invitation that social obligation requires accepting. Without this knowledge the utterance reads as a simple scheduling conflict. This example scored lowest among cultural reference items (1.0 for GPT-4o and Claude, 1.3 for Gemini), with all models recognising the conflict between dieting and feasting but none articulating the specific social obligation that makes refusal culturally impossible.

Table 2 summarises all thirteen examples.

5 Contrastive Minimal Pairs: A Worked Illustration

Reviewers reasonably pointed out that minimal pairs are central to our framework but had not been worked out concretely in the original draft. Section 8 sets them out as a protocol for future corpus development. Here we present the principle in concrete form on one example so the rest of the paper can refer to a worked instance rather than an abstract proposal.

For Example 7 (the koshary register switch), three variants isolate the role of register switching:

Variant A: all-MSA, propositional content preserved.
هل تفضّل تناول وجبة صحيّة؟
لا، أفضّل تناول الكشري
Setup (MSA): "Would you prefer to consume a healthy meal?"
Punchline (MSA): "No, I prefer to consume koshary."
Effect: no register contrast, no humor. The cultural reference (koshary) is preserved, demonstrating that the cultural element alone is not sufficient.

Variant B: all-colloquial Egyptian.
عايز ماكلة صحية؟
لا يا عم هاتلي كشري وخلاص

Ex.	Arabic	English Translation	Dialect	Mechanism	Computational Challenge
1	ما شاء الله يا عبقري...كسرت الكوب	"Genius...broke the cup after being told"	Egyptian	Sarcastic praise	Polarity inversion
2	عملت ريجيم...نقصت 200 غرام من صبري	"Did a week's diet...lost 200 grams of patience"	Tunisian	Expectation inversion	Semantic type constraint
3	اشتريت كتاب تطوير الذات...ولسه نفس الشخص	"Bought self-help book...still same person"	Levantine	Ironic self-evaluation	Genre convention awareness
4	أنا مش بخيل...أنا اقتصادي	"Not stingy...economical"	Egyptian	Lexical reframing	Implicit admission via denial
5	شغّل دماغك...تم إيقاف الخدمة	"Use your brain / service unavailable"	Egyptian	Conceptual metaphor	Cross-domain metaphor
6	الحر كأن الشمس تشرب قهوة معنا	"Heat as if the sun came down for coffee"	Gulf	Hyperbole	Personification, implausibility
7	وجبة صحيّة؟...هاتلي كشري	"Healthy meal? / Just bring me koshary"	MSA + Egyptian	Register switching	Diglossia plus cultural (secondary)
8	تحلّى بالصبر...الصبر خلص	"Please be patient / patience ran out"	MSA + Levantine	Register switching	Register plus pragmatic inference
9	نعمل دايت...بعد رمضان	"Let's diet / after Ramadan"	Gulf	Cultural reference	Religious-cultural knowledge
10	فكرة عبقرية...نأجل الشغل لآخر دقيقة	"Brilliant idea...postpone work to last minute"	Levantine	Sarcastic praise	Contextual polarity inversion
11	ذاكرت طول الليل...عشان أنام في الامتحان	"Studied all night...to sleep in the exam"	Egyptian	Expectation inversion	Goal inversion tracking
12	أنا مو كسول...أنا موفّر للطاقة	"Not lazy...energy efficient"	Gulf	Lexical reframing	Register incongruity detection
13	أبدأ حمية...بس اليوم عزيمة	"Dieting today...but today there's a feast"	Iraqi	Cultural reference	Social obligation knowledge

Table 2: All analysed examples with Arabic source, English translation, dialect, primary humor mechanism, and computational challenge. Arabic text is abbreviated for the summary; full versions appear in Section 4. Example 7 carries a secondary cultural reference dimension noted in the text.

Setup (Egyptian): "Want a healthy meal?"
Punchline (Egyptian): "No man, just bring me koshary and that's it."
Effect: weaker humorous effect. The colloquial-to-colloquial pairing removes the register asymmetry; the punchline reads as ordinary preference rather than register-defying refusal.

Variant C: original (MSA setup, Egyptian punchline).
The register switch is restored, and the humorous effect returns in full.

This contrast supports the necessity and sufficiency criterion for mechanism assignment in Section 3: register switching, not the cultural reference, is the necessary trigger for the humor in Example 7. Section 8 generalises this protocol to sarcasm (where the trigger removal variant replaces the sarcastic frame with a direct statement) and to cultural references (where the loaded element is substituted with a culturally neutral equivalent). Validation of these contrasts via human funniness judgements is, as we acknowledge in the Limitations, a necessary next step before the framework can be used at scale.

6 Pilot Probing Study

To ground the four failure categories in observable model behaviour rather than intuition alone, we conducted a small-scale probing study using the thirteen examples from Section 4. We emphasise that thirteen items constitute a proof of concept rather than a sta-

tistically reliable evaluation; results are indicative and motivate larger-scale follow-up work, not direct model comparison or ranking.

We probed three widely used frontier LLMs representing different providers: GPT-4o (OpenAI, version `gpt-4o-2024-08-06`), Gemini (Google, version `gemini-2.0-flash`), and Claude (Anthropic, version `claude-sonnet-4-5`). All models were queried via their respective APIs using default decoding parameters (temperature = 1.0, top-p = 1.0) with no system prompt. Each model received the Arabic text of each example followed by the identical prompt: *"Explain why this Arabic utterance is humorous and identify the humor mechanism."* The high-temperature default was chosen to allow each model's typical generative behaviour to surface rather than to optimise for a single best response.

We deliberately focus on frontier LLMs rather than Arabic-specialised encoders such as MARBERTv2 or AraBERTv2 because our study targets open-ended explanation rather than classification. Encoder-only models are not equipped to produce free-text explanations of humor mechanisms and therefore cannot be evaluated on the same scoring dimensions; they remain an important complementary baseline for future classification-oriented tasks derived from this framework.

Scoring procedure. Responses were evaluated on three dimensions using a 0–2 scale. **Mechanism**: 0 if incorrect or absent, 1 if partially correct (for example, identifying irony without explaining the mechanism), 2 if correct with appropriate reasoning. **Register**: 0 if dialect or register is ignored, 1 if a variety difference is noted but not interpreted pragmatically, 2 if register contrast is linked to the humorous effect. **Cultural**: 0 if cultural presupposition is absent, 1 if partially recognised, 2 if clearly articulated. Two annotators (the author and one colleague familiar with Arabic NLP and humor analysis) independently evaluated all model outputs. Initial inter-annotator agreement, computed as Krippendorff's α over ordinal ratings, was $\alpha = 0.71$ (Mechanism), $\alpha = 0.68$ (Register), and $\alpha = 0.73$ (Cultural), indicating acceptable reliability before adjudication. Remaining disagreements were resolved through discussion. The reported per-example value is the mean of the three dimension scores after adjudication.

Ex.	Layer	GPT-4o	Gemini	Claude
1	Pragmatic	1.0	1.0	1.0
2	Pragmatic	0.3	0.0	0.0
3	Pragmatic	0.7	0.3	0.7
4	Semantic	0.7	0.7	0.7
5	Semantic	0.7	0.7	0.7
6	Semantic	0.7	0.7	0.0
7	Socioling.	1.7	1.7	0.7
8	Socioling.	0.0	0.0	1.3
9	Socioling.	1.3	1.3	1.3
10	Pragmatic	0.7	0.7	0.7
11	Pragmatic	0.7	0.7	0.7
12	Semantic	0.7	0.7	0.7
13	Socioling.	1.0	1.3	1.0

Table 3: Mean scores (0–2) across the three scoring dimensions (Mechanism, Register, Cultural) for the thirteen probed examples. Per-example maxima vary by layer: approximately 0.67 for semantic-only examples, 1.33 for examples requiring two dimensions, and 2.0 for examples requiring all three. Scores should be read relative to these per-example ceilings, not against a uniform 2.0 maximum.

Interpreting the scores. Because not every example exercises every dimension, the three-dimension mean has a particular interpretation. For a purely semantic example (Example 4, lexical reframing), the maximum achievable score is 2 on Mechanism and 0 on Register and Cultural, yielding a maximum mean of approximately 0.67. A score around 0.7 therefore reflects close-to-ideal performance, not mediocre performance. For a register-switching example (Example 7), a model that correctly identifies the mechanism and the register contrast but misses the cultural reference would receive (2, 2, 0), yielding a mean of approximately 1.3. The full 2.0 mean is reserved for examples whose interpretation requires all three dimensions. Results appear in Table 3.

Read relative to these per-example ceilings, semantic humor is handled close to its expected maximum: scores of 0.7 on Examples 4, 5, and 12 indicate that Mechanism was reliably captured while Register and Cultural remained at zero as expected. Pragmatic hu-

mor involving expectation inversion is interpreted inconsistently (Example 2 collapses to 0 or near 0 across all three models). Sociolinguistic humor presents the greatest challenge relative to its higher per-example ceiling: register-switching Examples 7 and 8 reveal that models sometimes detect dialectal variation but fail to connect the register contrast to the humorous mechanism, and the pattern of strengths reverses between the two examples across models. Cultural reference humor (9, 13) is sometimes recognised but explanations rely on generic social reasoning rather than explicit cultural grounding. These patterns align with ALPS (Al-Olimat and Alshareef, 2026) on pragmatic literalism and with SAIDS (Kaseb and Farouk, 2022) on dialect-sarcasm interactions in Arabic.

7 Implications for Large Language Models

The pilot results suggest four systematic categories of LLM failure, framed as hypotheses motivating larger studies rather than as conclusions established at the present scale.

7.1 Literal Interpretation of Sarcasm and Irony

Mashallah in Example 1 may be genuine admiration or deep sarcasm depending entirely on context; all three models scored 1.0, identifying irony but failing to explain the register-dependent mechanism. Example 2 scored at or near 0 across the three models, confirming that semantic type constraint violations go undetected. ArSarcasm (Abu Farha and Magdy, 2020) and iSarcasmEval (Abu Farha et al., 2022) both document this gap; even MARBERT (Abdul-Mageed et al., 2021) struggles because dialectal variation interacts with ironic meaning in ways pretraining does not address.

7.2 Dialect-Specific Vocabulary and Cultural Reference Gaps

The koshary reference (Example 7) requires knowing the food's pragmatic associations with working-class Egyptian culture; GPT-4o and Gemini scored 1.7, correctly identifying the register switch but offering only surface cultural explanation. Examples 9 and 13 scored 1.0 to 1.3, suggesting partial cultural knowledge, with explanations relying on generic social reasoning rather than the specific presuppositions involved.

7.3 Register Contrast Modeling

The register-switching examples reveal the sharpest inter-model divergence. On Example 7, GPT-4o and Gemini scored 1.7 while Claude scored 0.7. On Example 8 the pattern reversed: GPT-4o and Gemini scored 0.0 while Claude scored 1.3. This asymmetry suggests that dialectal training-data coverage shapes which register contrasts a model can reason about, and that the bottleneck is pragmatic reasoning over the shift rather than variety detection per se. Given the small sample, this is best read as a hypothesis to test on larger sets.

7.4 Pragmatic Inference Failures

Example 4 requires recognising that a denial can implicitly confirm an accusation; Examples 2 and 3 require tracking genre-specific expectations. Uniformly modest scores (0.3 to 0.7) align with documented LLM weaknesses in Arabic pragmatic inference (Al-Olimat and Alshareef, 2026). SAIDS (Kaseb and Farouk, 2022) shows that dialect and sarcasm signals improve downstream sentiment; a humor-focused multi-task architecture predicting humor type, register identity, and cultural presupposition simultaneously is a promising direction.

8 Discussion

8.1 Toward an Operationalised Evaluation Protocol

Building on the pilot study and the worked minimal-pair example in Section 5, we propose a concrete, operationalised evaluation protocol for future work.

Contrastive minimal pairs. For each mechanism, pair an original humorous utterance with a variant that removes the humor trigger while preserving propositional content. The koshary case in Section 5 illustrates the principle for register switching. For sarcasm, the trigger removal variant replaces the sarcastic frame with a direct statement (Example 1 becomes a literal complaint about a broken

cup). For cultural references, the loaded element is substituted with a culturally neutral equivalent (Example 9 becomes a generic deferral to an unspecified future date). Annotators should rate each pair for funniness so that an effect size can be reported per mechanism.

Scoring rubric. Large-scale evaluation should score Mechanism identification, Register recognition, and Cultural presupposition recovery (each 0 to 2), with worked examples per level to support inter-rater reliability and Krippendorff's $\alpha \geq 0.70$ before aggregation. As noted in Section 6, the mean across the three dimensions should be interpreted relative to a per-example ceiling determined by which dimensions are genuinely required, not against a uniform 2.0 maximum.

Cultural presupposition ontology. Presuppositions should be classified into five coarse domains: (1) food and hospitality, (2) religious practices and calendar, (3) social obligations and norms, (4) bureaucracy and institutions, and (5) generational or political references.

Predicted difficulty ordering. The pilot is consistent with the ordering sociolinguistic and cultural reference humor harder than pragmatic humor harder than semantic humor, when each is scored relative to its appropriate ceiling. Cultural reference examples (9, 13) and register-switching Example 8 fall short of their ceilings most consistently; semantic examples (4, 5, 12) reach close to their per-example maxima. Future corpora should report per-mechanism accuracy and per-mechanism ceiling-relative scores to track this ordering over time.

Annotation guidelines. Each corpus item should include: dialect label, primary and secondary mechanism categories, a non-humorous paraphrase preserving propositional content, and a cultural presupposition field. Multi-label cases where two mechanisms are jointly necessary should be flagged for adjudication.

Connections and extensions. ALPS (Al-Olimat and Alshareef, 2026) provides the template for diagnostic Arabic probing; our protocol extends it to dialectal humor, and SAIDS (Kaseb and Farouk, 2022) motivates joint evaluation of dialect, pragmatics, and humor type. Although the present study uses Arabic, the framework itself transfers naturally to other diglossic or dialect-rich language families. Future work should explore retrieval-augmented prompting with cultural knowledge snippets and multimodal extensions using Arabic ASR for prosodic register cues in spoken humor.

9 Conclusion

We argued and demonstrated that Arabic humor functions as a principled diagnostic probe for LLMs, surfacing gaps in pragmatic inference, diglossic register reasoning, and cultural grounding that standard benchmarks do not reach. A three-layer taxonomy across five dialect regions was illustrated through a small probing study on GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5. When scored relative to per-example ceilings, semantic humor reaches close to its maximum, while sociolinguistic humor produces the most variable results, with register-switching examples revealing sharp inter-model divergence tied to dialectal training-data coverage. We operationalised an evaluation protocol with contrastive minimal pairs (worked out concretely for one example), a three-dimensional scoring rubric, a cultural presupposition ontology, and annotation guidelines. We position this contribution as a diagnostic framework and pilot study, not a finished benchmark, and view the construction of a large, dialectally balanced Arabic humor corpus as the natural next step.

Limitations

This study has several limitations that should be borne in mind when interpreting its contributions.

First, the probing study in Section 6, while covering all thirteen examples across three frontier LLMs, remains limited in scale. Thirteen items constitute a proof-of-concept demonstration rather than a statistically reliable empirical evaluation. The mean scores in Table 3 should be interpreted as indicative rather than definitive, and the Krippendorff α values reported (0.68 to 0.73) fall close to the conventional 0.70 threshold, meaning the scoring dimensions require further validation

on a larger item set before they can be used for high-stakes model comparison. We also note that the stability of open-ended explanation evaluation across runs has not been measured here; future work should report variance across multiple samples per item and ideally across multiple prompt formulations.

Second, the pilot does not include Arabic-specialised encoder models such as MARBERTv2 or AraBERTv2. These models are not capable of producing free-text explanations of humor mechanisms and therefore cannot be directly evaluated under the same three-dimensional rubric. However, they represent important baselines for classification-oriented tasks, and future work should compare frontier LLM explanation quality against encoder-based classification performance to establish whether the gaps observed here are specific to general-purpose LLMs or persist for Arabic-pretrained systems.

Third, the necessity and sufficiency criterion for mechanism assignment, while operationally useful and illustrated by the worked minimal-pair example in Section 5, has not been validated through human funniness judgments at scale. Ideally, minimal-pair ablation studies would collect funniness ratings from native speakers of the relevant dialect before and after trigger removal, reporting effect sizes to demonstrate that removing the assigned trigger reliably suppresses the humorous effect. This validation is a necessary step before the taxonomy can be used to train annotation models or derive evaluation metrics.

Fourth, the thirteen examples were selected to illustrate the taxonomy rather than to constitute a statistically representative sample of Arabic humor. No formal inter-annotator agreement measurement was conducted for mechanism assignment across the full set. A large-scale corpus covering multiple dialects, humor types, and cultural reference categories, annotated by multiple native speakers with explicit reliability measures, is necessary for quantitative analysis of mechanism distribution and for training or evaluating computational models.

Fifth, while the example set spans five dialect regions, the Arab world encompasses many other varieties including Moroccan Darija, Sudanese, Yemeni, and Libyan Arabic. Moroccan Darija in particular, which draws heavily on Amazigh and French alongside Arabic, may exhibit humor dynamics that differ substantially from the MSA and colloquial contrast we foreground here.

Sixth, our treatment of cultural reference humor is necessarily incomplete. Cultural knowledge underlying humor is dynamic, generational, and community-specific. The examples represent a snapshot of humor circulating in 2023 and 2024, and the cultural presuppositions they rely on may shift over time.

Seventh, the study focuses exclusively on written text. Spoken Arabic humor involves prosodic and intonational cues that interact with register switching in ways that text alone cannot capture, as noted in the Discussion. The framework should be extended to spoken modalities as Arabic ASR resources for dialectal speech continue to develop.

Finally, the broader applicability of the framework to other languages, though plausible (and noted in Section 8), has not been tested empirically. The contrastive minimal-pair protocol and three-dimensional rubric are language-independent in design, but their utility for, say, Swiss German and Standard German diglossia, or for Mandarin and topolect contrasts, remains an open question.

Ethical Considerations

This research is primarily analytical and does not involve the collection of personal data, the construction of a new annotated dataset, or the deployment of a computational system. The humorous examples are drawn from publicly circulating informal discourse on social media platforms and do not identify specific individuals or communities in ways that could cause harm.

Humor in Arabic, as in any language, exists on a spectrum that includes not only benign wit but also content that may be offensive, stereotypical, or exclusionary. Our taxonomy focuses on humor mechanisms rather than humor targets, and we deliberately selected examples that illustrate linguistic and computational properties without targeting particular ethnic, religious, gender, or socioeconomic groups. Future work building computational systems for Arabic humor processing should

include explicit annotation guidelines for offensive and harmful humor, with careful attention to how such categories are defined across different Arabic-speaking communities.

Computational systems capable of detecting and interpreting humor in Arabic social media content could be misused, including for surveillance of political dissent, overreaching content moderation, or automated analysis of communications in ways that violate user expectations of privacy. Researchers and practitioners building on the framework proposed here should consider these downstream risks explicitly during system design and evaluation.

Finally, humor is deeply culturally situated, and interpretations of what is funny in a given context reflect the perspectives and lived experiences of individuals embedded in those cultural contexts. Annotated datasets for Arabic humor should include annotation from native speakers of the specific dialect variety in question rather than relying on annotators with general Arabic proficiency. Annotation tasks should be designed to surface disagreements rather than resolve them prematurely, since variation in humor interpretation across age groups, genders, regions, and communities is itself a scientifically meaningful signal rather than noise to be eliminated.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Hussein S. Al-Olimat and Ahmad Alshareef. 2026. ALPS: A diagnostic challenge set for Arabic linguistic and pragmatic reasoning. *Preprint*, arXiv:2602.17054.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3–4):293–347.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Abdelrahman Kaseb and Mona Farouk. 2022. SAIDS: A novel approach for sentiment analysis informed of dialect and sarcasm. In *Proceedings*

of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 22–30, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*. D. Reidel Publishing Company, Dordrecht.

Genta Indra Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.