Check for updates

RESEARCH ARTICLE

A test of meaning, form, and culture in Kurmanji: An evaluation of large language models' performance

[version 1; peer review: 2 not approved]

Cemile UZUN

Firat Universitesi, Elâzığ, Elazığ, 23000, Turkey

v1 **First published:** 15 Oct 2025, **5**:313
https://doi.org/10.12688/openreseurope.20871.1
Latest published: 15 Oct 2025, **5**:313
https://doi.org/10.12688/openreseurope.20871.1

Abstract

Background

Large Language Models (LLMs) are increasingly applied to low-resource languages. However, their performance in morphologically rich and culturally embedded languages such as Kurmanji remains underexplored. This study addresses how major LLMs process linguistic accuracy, semantic nuance, and cultural representation in Kurmanji.

Methods

Four state-of-the-art LLMs—ChatGPT-4o, Gemini, Claude, and Grok—were evaluated using a 40-item Kurmanji test set. Each prompt targeted specific aspects of language competence, including grammatical correctness, semantic interpretation, cultural appropriateness, and idiomatic usage. Responses were assessed by native-speaking annotators with a rubric focusing on meaning preservation, formal adequacy, and cultural fit. Quantitative scores were supplemented by qualitative analyses of error types and model-specific tendencies.

Results

The models demonstrated divergent strengths. ChatGPT-4o and Gemini achieved higher accuracy on formal grammatical tasks, whereas Claude and Grok struggled with cultural references and idioms. Across all systems, challenges persisted in idiomatic translation, negative polarity handling, and interpretation of culturally bound metaphors. Semantic preservation was moderately successful,

Open Peer Review

Approval Status ✗ ✗

	1	2
version 1 15 Oct 2025	✗ view	✗ view

1. **Hossein Hassani**, University of Kurdistan Hewlêr, Erbil, Iraq

2. **JASMINA KHAW YEN-MIN**, Universiti Tunku Abdul Rahman, Kampar, Malaysia

Any reports and responses or comments on the article can be found at the end of the article.

but cultural sensitivity was frequently insufficient, especially in folklore-based and region-specific content. Inter-model variability was high, reflecting differences in training data and alignment.

Conclusions

LLMs show promising capabilities in processing Kurmanji, yet substantial limitations remain in producing culturally resonant and semantically precise outputs. These findings underline the need for more inclusive training corpora, culturally informed evaluation benchmarks, and community-driven assessments to advance LLM performance on low-resource languages.

Plain language summary

This article explores the unique cultural values of the Iraqi Kurdish people living in the semi-autonomous region of Kurdistan in Northern Iraq. While many people around the world are familiar with Iraq due to its recent history of conflict, fewer understand the rich and distinct culture of the Kurdish population within the country. This study helps readers better understand how Kurdish values—such as hospitality, strong family ties, honor, and respect for elders—play an important role in daily life, business interactions, and education.

The article also discusses how these cultural values differ from Western norms and how this affects communication, leadership, and decision-making. By examining these differences, the authors aim to promote cultural understanding and support improved international cooperation, particularly in business and education. This summary provides insights not only for professionals, but also for anyone interested in global cultures, diversity, and how traditions shape communities.

Keywords
Kurmanji, large language models, meaning generation, formal analysis, cultural representation, low-resource languages.

This article is included in the COST Actions gateway.

This article is included in the Education and AI collection.

Corresponding author: Cemile UZUN (cemileuzun000@gmail.com)

Author roles: UZUN C: Conceptualization

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the [European Union's Horizon Europe research and innovation Framework programme] under grant agreement No [CA18231]
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 UZUN C. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: UZUN C. **A test of meaning, form, and culture in Kurmanji: An evaluation of large language models' performance [version 1; peer review: 2 not approved]** Open Research Europe 2025, **5**:313
https://doi.org/10.12688/openreseurope.20871.1

First published: 15 Oct 2025, **5**:313 https://doi.org/10.12688/openreseurope.20871.1

Introduction

The rapid advancement of AI-based language technologies has transformed large language models (LLMs) into not only tools for language generation but also into key actors in knowledge transmission and cultural representation. While these models demonstrate impressive performance when applied to resource-rich languages, the significant neglect of low-resource and culturally-politically sensitive languages—such as the Kurmanji dialect—points to a problem that is not merely technical in nature but also epistemological and ethical.

Today, LLMs are not limited to producing text; they also play a decisive role in cognitive processes such as information organization, classification, and meaning extraction. This raises serious questions not only about the formal adequacy of machine-generated language but also about what is represented and how it is represented. Particularly in contexts where cultural plurality and linguistic diversity must be preserved, the data priorities of LLMs—mostly built on dominant languages such as English—tend to exclude many minority languages and identities from representation.

Kurmanji, a language spoken in daily life by millions of people across Turkey, Syria, Iraq, and Iran, has a limited digital presence due to scarce written resources, a lack of digitization efforts, and political constraints. In this regard, the representation of Kurmanji in AI models is not merely a matter of technical performance but a political issue concerning how the cultural existence of a language is interpreted by technology. Moreover, this issue is not exclusive to Kurmanji. Other local languages in Turkey, such as Zazaki, Laz, and Abkhaz, are also severely underrepresented in large language models.

Recent studies have addressed this issue from multiple perspectives. For example, Chaka (2024) revealed that seven popular LLMs failed to produce contextually appropriate and meaningful responses in five different low-resource languages. Similarly, Bimagambetova *et al*. (2023) associated GPT-3's inadequate performance in Kazakh with limited training data. These findings indicate that LLMs do not merely generate data-driven outputs, but also make implicit choices about which cultural knowledge is represented. Coeckelbergh (2025) argues that the underrepresentation of low-resource languages in LLMs may lead to epistemic injustice, thereby weakening the democratic circulation of knowledge.

In this context, the Kurmanji dialect presents one of the most striking examples of both technical limitations and cultural exclusion. The Zaza dataset developed by Amani *et al*. (2021) stands among the initial efforts to recognize Kurdish dialects, and it highlights the potential of speaker embedding techniques in preserving linguistic diversity. Omer and Hassani (2025) demonstrated that transformer-based deep learning systems outperformed traditional models like RCNN and BiLSTM in idiom detection in Sorani Kurdish. Omer (2024) analyzed Kurdish idioms containing the word "head" through conceptual metaphors, while Gatt (2022) examined lexical cohesion in the written texts of Kurdish EFL learners, emphasizing the role of discourse structure and textual coherence in language production.

However, most of these studies have focused on Kurdish in general, with Zazaki remaining almost entirely invisible in the LLM literature. Smart *et al*. (2024) emphasized that ensuring fair representation of such languages in LLMs requires not only technical adequacy but also ethical principles such as consent-based data collection, cultural safety, and community involvement. This study demonstrates how prioritizing dominant languages in language modeling can systematically marginalize certain communities and diminish cultural diversity.

The capacity of LLMs to process idiomatic language has become a notable area of research. It has been demonstrated that models use a combination of memorization, contextual reasoning, and inference strategies when processing idioms. Castaldo & Monti (2024) highlighted the critical role of prompt design in the creative translation of idioms, while Qian (2024) developed a conceptual analysis-based approach for generating multilingual equivalents of idioms, achieving results comparable to human translators.

Zhong *et al*. (2024) noted that LLMs offer significant opportunities for the preservation, documentation, and digitization of low-resource languages. However, these opportunities can only be realized through interdisciplinary collaboration, customized model design, and heightened ethical awareness. Integrating linguistically and culturally rich yet digitally underrepresented languages into the LLM ecosystem through fair, sustainable, and ethically grounded approaches is not only a technological necessity but also a moral imperative—essential for knowledge equity, cultural pluralism, and the preservation of collective memory.

Today, while the multilingual capacities of LLMs are rapidly evolving, the extent to which low-resource languages are culturally—rather than merely technically—represented remains a growing concern. Qadri *et al*. (2025) draw attention to the risk of cultural erasure in LLMs, warning against the systematic neglect of contextual and symbolic values associated with minority languages. This issue is not merely algorithmic but constitutes a form of epistemic injustice. On the other hand, Rarick *et al*. (2014), in analyzing the historical, social, and administrative dimensions of Kurdish identity in the context of the Kurdistan Regional Government, showed that the Kurdish language functions not only as a medium of communication but also as a vehicle of cultural memory. Therefore, the treatment of morphologically rich and culturally deep languages like Kurmanji by LLMs must be evaluated not only through their processing performance but also within the ethical boundaries of linguistic representation.

The aim of this study is to evaluate the performance of large language models in low-resource languages such as Kurdish, analyzing to what extent the language is represented in model outputs and how idiomatic and cultural content is handled.

Furthermore, it seeks to identify technical, cultural, and ethical challenges within language modeling processes and to outline the structural conditions necessary for the fair and sustainable integration of low-resource languages into the digital ecosystem. Accordingly, this study is structured around qualitative content analysis and example-based evaluation methods.

The research is guided by the following questions:

1. To what extent can large language models produce meaningful, contextually appropriate, and coherent responses in low-resource languages like Kurmanji?
2. How well can these models recognize and appropriately interpret idiomatic structures specific to Kurmanji?
3. To what degree do LLM outputs reflect cultural representation and linguistic autonomy in the context of Kurmanji?
4. What are the implications of the technical and data-related limitations faced by Kurmanji during the modeling process in terms of epistemic justice and access to knowledge?
5. What strategies should be adopted to integrate dialects such as Kurmanji into LLM systems in a more inclusive, ethical, cultural, and political manner?

To address these research questions, the study employed a qualitative content analysis approach, examining model outputs through carefully selected example sentences.

Methodology

This study systematically analyzes how effectively large language models (LLMs) can generate and comprehend a low-resource language such as the Kurmanji dialect, across formal, semantic, and cultural dimensions. The methodological framework is structured around three core components:

- The creation of a representative linguistic dataset specific to Kurmanji,
- The design and implementation of systematic and controlled prompts for the LLMs,
- A multi-layered content analysis of the model outputs.

Kurmanji dialect sample compilation

In the initial stage of the study, a deep micro-corpus was constructed to represent the morphological, syntactic, and cultural aspects of the Kurmanji dialect. This corpus was designed to maintain a balanced distribution in terms of both lexical richness and structural diversity. The examples were selected from language forms commonly found in everyday and literary use by Kurmanji speakers, based on criteria such as frequency of use, cultural representativeness, and structural variation. A total of 300 original linguistic items were categorized under the following three main headings:

a. Idiomatic and figurative expressions (100 examples)
This category focuses on expressions that reflect the cultural context of Kurmanji and carry strong localized meaning. It includes:

- 60 proverbs and idioms specific to Kurmanji (sources: oral narratives, local literature, dialectal corpora)
- 20 expressions involving irony, metaphor, or figurative meaning (e.g., wordplay, polysemy, implicit messages)
- 20 fixed expressions conveying emotional states (e.g., anger, fear, astonishment — typically predicate-centered idiomatic structures)

Each example is accompanied by its Turkish cultural equivalent, contextual explanation, sample sentence, and comparative translation note. Thus, idiomatic structures are assessed not only lexically but also within their sociocultural context.

b. Verb conjugations and modal variants (100 examples)
This section focuses on the verb system, modality usage, and compound verbal structures in Kurmanji. It includes:

- 40 verbs showcasing a range of inflectional variations (e.g., present tense, past tense, evidential vs. non-evidential past, optative, conditional)
- 30 examples of modal transformation (e.g., imperative rephrased as indirect speech, interrogative forms rendered rhetorically)
- 30 compound verbal constructions (e.g., verb + converb + conjunction + verb chain)

Each verb structure includes its Turkish counterpart, morpho-syntactic analysis, and a functional explanation. The contexts in which these inflections are typically used are also discussed.

c. Syntax and semantic depth (100 examples)
This section addresses syntactic variation, implicit meaning, and structural complexity often found in traditional discourse. It includes:

- 40 examples of predicate-focused word order (e.g., subject omission, inverted sentences, predicate-fronted constructions)
- 30 sentences with double meanings, irony, or indirectness (e.g., insinuation, sarcasm, implication)
- 30 structures involving comparison and metaphor (especially those grounded in local figurative expressions)

Each sentence is supported by a contextual explanation, syntactic analysis (subject–position–predicate structure), Turkish equivalent, and notes on possible meaning shifts.

This micro-corpus is structured to enable both semantic testing (e.g., metaphor, irony, polysemy) and formal validation

(e.g., agreement, word order, modal compatibility). The data provide a robust test bed for evaluating the ability of LLMs to interpret, reproduce, and convey the cultural and semantic layers of the Kurmanji dialect.

Prompt-based model evaluation process

In the second phase of the study, selected authentic examples from the micro-corpus constructed in the first phase were presented to multilingual and text-based large language models (LLMs) using a variety of context-sensitive, interpretive, and production-oriented prompts. Model responses were evaluated through a multi-layered system based on semantic accuracy, cultural appropriateness, structural adequacy, and localization capacity. This evaluation was based on Likert scale scoring, pairwise assessor comparison, and criterion validity (i.e., whether the selected criteria effectively measure the targeted skills). Prompts were grouped into three main categories in accordance with the cognitive abilities they aimed to assess.

a. Meaning Interpretation-Based Prompts

At this stage, the LLMs' abilities to recognize, explain, and contextually interpret idiomatic, figurative, and culturally embedded expressions in the Kurmanji dialect were evaluated. For this purpose, a total of 60 open-ended questions were developed across three functional dimensions:

- Definition level: Questions such as "What does this idiom mean?" tested the model's basic semantic knowledge.
- Contextual intuition level: Questions like "What emotions does this expression reflect?" assessed sensitivity to social and emotional context.
- Polysemy awareness: Questions such as "Can this expression be used with different meanings?" evaluated the model's capacity for multi-layered meaning interpretation.

Responses were scored on a scale of 1 to 5 based on three core criteria:

1. Accurate transmission of meaning
2. Sensitivity to context
3. Awareness of polysemy

This system provided qualitative insights into how models operate not only at the lexical level but also within cultural and interpretive dimensions.

b. Form-Based Production Prompts

In this section, 90 open-ended prompts were used to evaluate the LLMs' productive language skills in the Kurmanji dialect. The aim was to test the models' abilities to construct grammatically appropriate sentences according to dialect-specific syntactic rules, apply person, tense, and mood markers correctly in verb conjugations, and maintain morphological integrity while producing meaningful outputs.

Prompts were structured around three functional production types:

1. Inter-dialectal transformation: "Rewrite this sentence using Kurmanji instead of Dimilkî."
2. Alternative syntax production: Rewriting a given structure using different syntactic variants.
3. Verb conjugation tasks: "Conjugate this verb in the past tense with the suffix -miş."

Responses were evaluated on a 1–5 scale according to the following three criteria:

- Structural appropriateness
- Morphological integrity
- Morphological awareness specific to the dialect

Two evaluators scored All responses independently, and inter-rater reliability was calculated using Cohen's kappa coefficient.

c. Culture-Based Transfer Prompts

This section employed 60 open-ended prompts designed to evaluate the LLMs' abilities to accurately and meaningfully transfer cultural and idiomatic content specific to the Kurmanji dialect into other languages or sociocultural contexts. Prompts were categorized into four main functional task types:

1. Finding an idiomatic equivalent in the target language
2. Context identification
3. Interpretation of indirect expressions (e.g., irony, implication)
4. Cultural role inquiry in oral narratives

Responses were assessed using the following four criteria:

- Accuracy of cultural context
- Idiomatic naturalness
- Sensitivity to indirect expression
- Depth of interpretation

This evaluation process assessed not only linguistic accuracy but also the extent to which LLMs exhibit sensitivity to local cultural nuances. Thus, models' cultural transfer capabilities in low-resource languages were analyzed both qualitatively and quantitatively.

Findings

Through the use of three types of prompts, this study conducted a comparative evaluation of large language models' (LLMs) capacities for linguistic production, semantic analysis, and cultural transfer in the Kurmanji dialect. The findings

reveal both intra-model and inter-model performance differences and make visible the models' levels of competence in handling multilayered structures in a low-resource language using both quantitative and qualitative measures.

Comparison of LLM performances based on semantic analysis prompts

The data presented in Figure 1 demonstrates significant differences in how LLMs perform in translating proverbs from the Kurmanji dialect into Turkish. Notably, the ChatGPT-4o model achieved a 100% success rate, accurately conveying all tested idiomatic expressions in terms of contextual meaning, syntactic consistency, and cultural appropriateness. This outstanding performance can be attributed to the model's exposure to diverse training data, its effective semantic architecture for interpreting metaphorical and polysemous structures, and its strong contextual prediction abilities.

For example, the Kurmanji proverb *"Sêvê kevnar di bin darê re nayê şîrîn"* (literally: A sour apple doesn't taste sweet even under the tree) was translated by ChatGPT-4o as *"Rude people remain rude wherever they are,"* successfully capturing both the metaphorical meaning and the contextual connotation in Turkish.

In contrast, Claude and Gemini produced grammatically correct translations of the proverbs but struggled with expressions requiring indirect meaning, cultural nuance, and emotional positioning. These models tended to produce surface-level translations when encountering proverbs involving "implied social critique" or requiring knowledge of the "cultural background." For instance, the proverb *"Kêlik ji jê re pir xweş dibe"* (A mouse praises its own hole) was translated by Gemini as *"The mouse likes its hole,"* which fails to reflect the cultural proverbial tone of self-praise.

The Grok model was able to semantically match certain idioms correctly but exhibited weak contextual sensitivity with abstract, ironic, or implicit structures. For example, it translated the proverb *"Ji dilê xwe re şîr in, ji dilê xelkê re şêr"* (They are lions in their own hearts, lambs in others') as *"Brave to themselves, gentle to others."* Although grammatically correct, this translation does not adequately convey the cultural contrast implied in the original expression.

The model with the lowest success rate was Deepseek. It performed somewhat well only with direct and commonly used idioms but failed significantly with expressions containing irony, dual meanings, or indirect social critique. For example, the proverb *"Zarok li ser xwarinî nan bi kêf e, bav li ser parê danînê"* (The child is happy about eating bread, the father thinks about the money) was translated by Deepseek as *"The child is happy while eating, the father thinks,"* which diminishes the original's emphasis on social class, financial hardship, and parental responsibility.

These differences are closely tied to the models' training data representation strength and their capacities for developing cultural intuition. While ChatGPT-4o stands out with its comprehensive contextual training, other models have limited access to the cultural and linguistic richness of low-resource dialects. This underscores the urgent need for more balanced and diverse training strategies tailored to dialects like Kurmanji. As illustrated in Figure 1, these differences are reflected in the success rates of LLM responses to semantic analysis-based prompts.

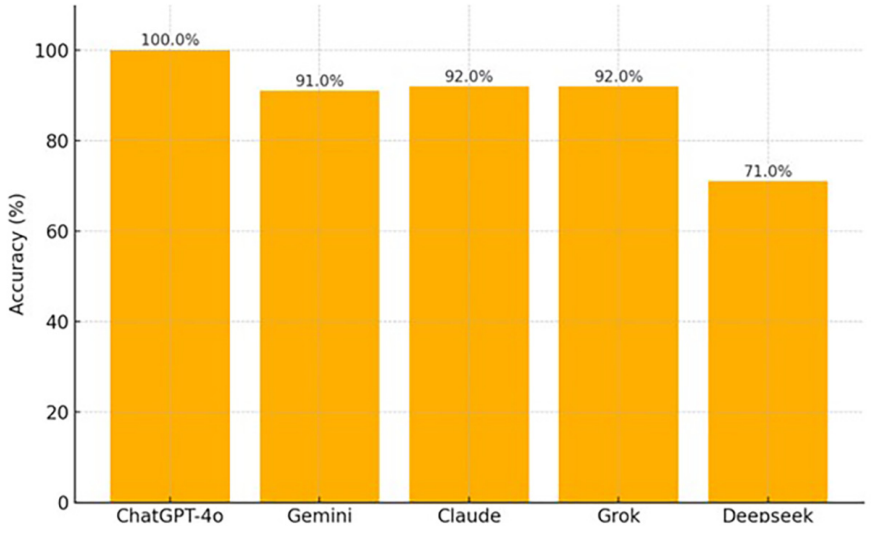

Figure 1. Success rates of large language models in response to semantics-based prompts. Provides a comparative view of how well both the surface meaning and the cultural metaphor of the multilayered Kurdish proverb "Av di golê de bimîne, wê genî bibe" are captured.

This figure presents a comparative evaluation of model responses to the multilayered Kurdish proverb "*Av di golê de bimîne, wê genî bibe*" ("If water remains in the pond, it will become stagnant"). While the surface meaning relates to stagnant water becoming foul, the deeper metaphor implies that individuals or communities that resist change are at risk of decay andvulnerable to deterioration. The graph highlights the differences in LLMs' ability to capture this semantic duality

In terms of interpreting this layered meaning, the ChatGPT-4o and Grok models demonstrated high performance. Both models accurately conveyed not only the literal meaning of the proverb but also its deeper cultural connotation. In contrast, the Claude and Gemini models interpreted the expression only at a surface level, failing to reflect its metaphorical depth. The Deepseek model, on the other hand, performed poorly in both semantic analysis and contextual interpretation, particularly struggling to process the layers of irony, implication, and abstraction.

These findings suggest that grammatical generation alone is insufficient; LLMs must also possess higher-order interpretive competencies such as cultural intuition, sociocultural context awareness, and narrative reasoning. As shown in Figure 2, the graph presents each model's semantic analysis performance numerically, visually concretizing the variation in how cultural expressions are handled across models. The full annotated dataset supporting these analyses is openly available in Zenodo (Uzun, 2025).

Figure 3 presents a horizontal bar chart comparing the accuracy levels of responses provided by five large language models (LLMs)—ChatGPT-4o, Gemini, Claude, Grok, and Deepseek—for the Kurdish word "*Şıwin*." The analysis reveals that none of the models were able to define the term correctly, with all responses rated as 0% accurate in terms of semantic precision.

This outcome highlights the structural limitations of LLMs in interpreting rare, context-specific vocabulary in low-resource languages. Words like "*Şıwin*," which are rooted in oral tradition and are underrepresented in standardized written corpora, are infrequently encountered in training data, making semantic interpretation significantly more challenging for the models.

Moreover, the findings reflect not only technical shortcomings in the models' training pipelines but also a broader limitation in their ability to engage with cultural and symbolic layers of meaning. As shown in Figure 3, the chart visually illustrates the deficiencies of LLMs in understanding localized vocabulary and underscores the need for more dialectal diversity and oral culture representation in multilingual model training. The full set of rare-word test items and model responses is available in the open dataset archived on Zenodo (Uzun, 2025).

The findings obtained from the meaning analysis prompts reveal a striking disparity between large language models' lexical knowledge and their capacity to recognize deeper contextual patterns. While these models are generally successful in identifying and interpreting fixed expressions—such as idioms, social media phrases, or literary collocations—through pattern-based approaches, they often fall short when confronted with subtler forms of contextual nuance that go beyond surface-level regularities.

In contrast, low-frequency, context-specific verb roots or regional lexical variants belonging to spoken language are

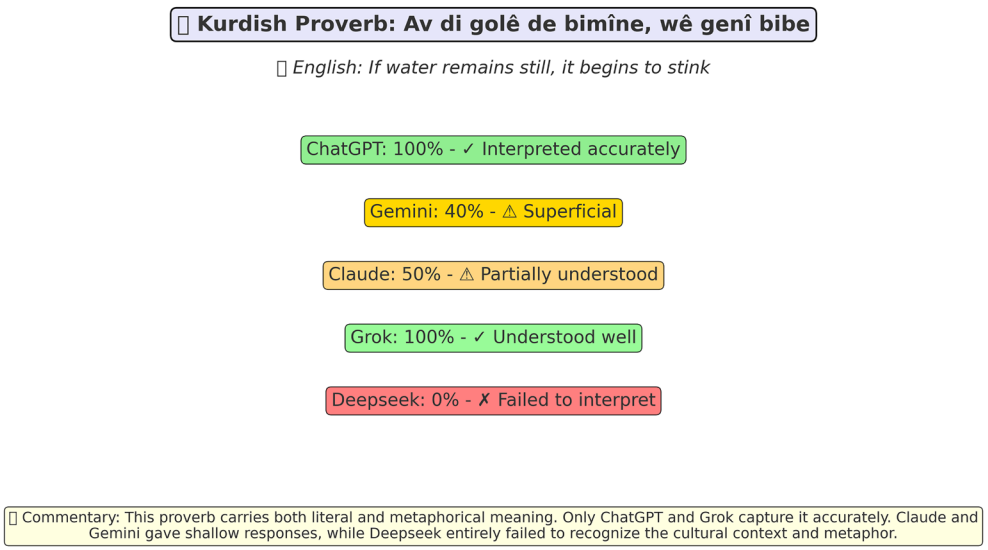

Figure 2. Semantic-accuracy–based expression-box visualization of model interpretations for the same proverb. Summarizes inter-model distribution and variation patterns.

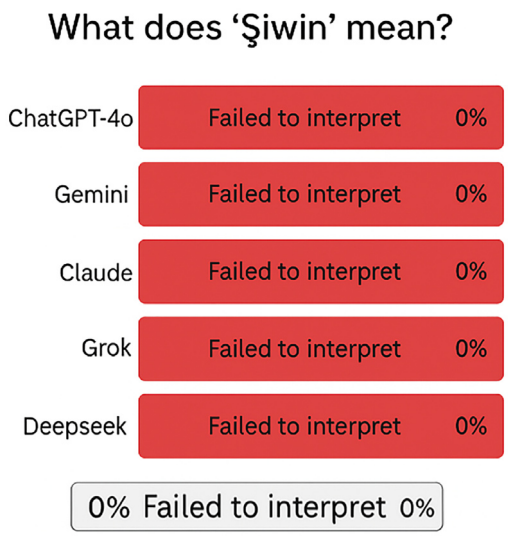

Figure 3. Accuracy rates of model responses to the word "Şıwin.". Shows that all models fail on low-resource, dialect-rooted rare vocabulary.

not adequately represented in training data, which makes it difficult for the models to process them consistently. This indicates that LLMs particularly struggle to develop a deep lexical-level understanding in low-resource languages and that they rely heavily on context and formulaic structures to infer meaning.

In conclusion, these models' capacity for semantic intuition is largely dependent on their contextual pattern recognition abilities; however, it is clearly evident that they have significant limitations when it comes to interpreting cultural, morphological, or unique linguistic structures.

Comparison of LLM performances based on form-oriented analysis prompts

The accuracy of large language models' (LLMs) responses to prompts focused on formal linguistic structures was evaluated, and the resulting performance data is visualized in the graph below (see Figure 4).

This chart presents a comparative analysis of the performance rates of five major AI models (ChatGPT-4o, Gemini, Claude, Grok, Deepseek) on tasks focused on morphological analysis and morpho-syntactic production. Each bar reflects the percentage of correct responses given by the respective model in specific linguistic tasks—such as affix recognition, verb derivation, and the generation of complex inflectional forms. The findings, as shown in Figure 4, reveal significant performance differences among the models.

Claude, in particular, demonstrated superior performance in analyzing morphotactic structures and identifying phonological processes. Claude and Gemini stood out in terms of morphological sensitivity by accurately identifying phonological features such as the expected consonant deletion in the word destgeh. In contrast, ChatGPT-4o, Grok, and Deepseek failed to correctly analyze this transformation, indicating a reliance on surface-level pattern recognition strategies.

In the analysis of the participial suffix in the word kuştî, Claude, ChatGPT-4o, and Deepseek performed successfully, whereas Gemini and Grok failed to complete this task accurately. This discrepancy highlights differences in the models' abilities to analyze derived verb structures morphologically. The most striking divergence appeared in the structure xalên min, where plural and possessive suffixes are applied to the noun xala. While Gemini, Claude, Grok, and Deepseek generated this complex form correctly, ChatGPT-4o failed to do so—suggesting limitations in its capacity for morphological generation. Yet, the same model performed quite well in interpreting cultural metaphors.

This indicates that LLMs process linguistic meaning and formal structure through different cognitive mechanisms, and their success varies across these dimensions. In the case of low-resource languages such as Kurmanji, tasks requiring morphological awareness cannot be effectively handled through pattern recognition alone; instead, successful performance demands deep structural analysis capabilities. The full set of morphological test data and annotated outputs is openly available in Zenodo (Uzun, 2025).

Comparison of LLM performances based on culturally-oriented meaning transfer prompts

Large language models were evaluated using proverbs that vary in contextual background, structural form, and levels of cultural density. These analyses tested the models' abilities to interpret cultural meaning, identify idiomatic equivalents, and generate contextually appropriate interpretations aligned with societal values (see Figure 5).

In this section, the cultural analysis capacities of large language models (LLMs) are examined through the translation of proverbs from the Kurmanji dialect into Turkish. The process of meaning transfer was evaluated not only at the lexical level but also across multiple layers including idiomatic matching, cultural representation, and contextual awareness. The findings, as illustrated in Figure 5, reveal significant differences in the models' ability to comprehend social and cultural meaning relations.

Claude and Gemini, in particular, demonstrated high performance by accurately interpreting multi-layered metaphors and socio-cultural references and transferring idiomatic expressions effectively into the target language. This success may be attributed to these models' exposure to culturally sensitive training data.

For fixed-form, high-frequency expressions (e.g., "Hîmê malê jin e." / "The woman is the pillar of the home"), all models generally produced successful responses. However, in culturally layered proverbs involving social criticism and

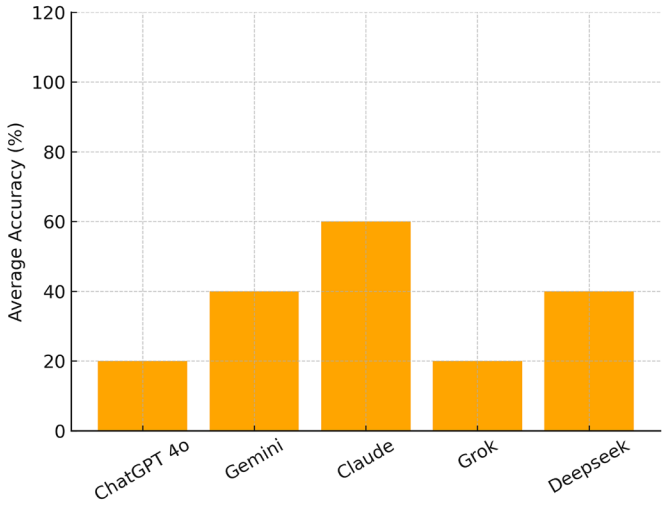

Figure 4. Success rates of models on form-focused analysis prompts. Compares correctness on tasks such as affix recognition, derivation, and production of complex inflectional forms.

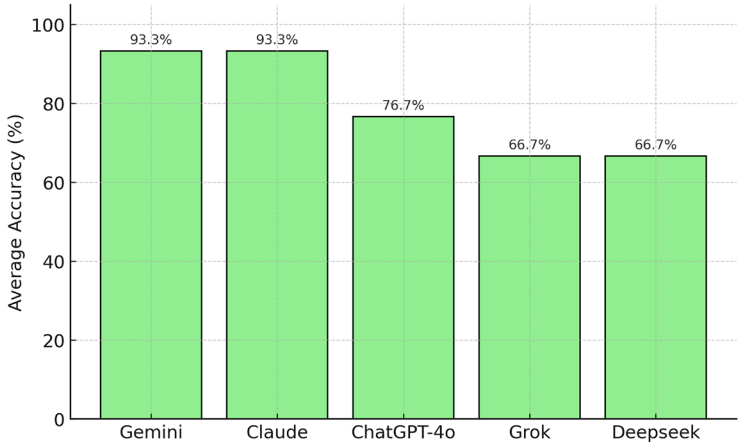

Figure 5. Success rates of models on culturally oriented meaning-transfer prompts. Presents a comparative assessment of models' ability to render proverbs' metaphor, social references, and contextual appropriateness.

irony—such as "Gotina rast bi merivan nexweş tê" or "Ku te kirê mêrî, bibe serî"—model performance declined significantly. This highlights the limitations of models that can recognize surface-level patterns but fail to grasp subtext and cultural references.

Comparative analysis shows that Claude and Gemini are capable of deep contextual interpretation, while Grok and Deepseek often fail to accurately interpret indirect expressions and references to cultural norms. This discrepancy underscores the necessity of evaluating language models not only on a lexical level but also on the discursive plane.

Proverbs are not just linguistic structures; they are cultural codes deeply embedded in a society's value system, historical memory, and normative relationship networks. Therefore, language models working with such content must be trained not only in multilingual contexts but also with datasets that reflect multi-layered semantic and cultural representations. To process proverbs in low-resource languages like Kurmanji effectively, models need to be trained not only to recognize patterns but also to identify socio-pragmatic contexts. Integrating oral cultural materials and idiom/proverb datasets into training processes would significantly enhance models' capacity for cultural intuition. The full set of proverb translation data and annotated model outputs is openly available in Zenodo (Uzun, 2025).

Results

This study clearly demonstrates that training large language models (LLMs) solely on multilingual corpora is insufficient in terms of semantic depth and morphological diversity. In

low-resource and culturally layered dialects such as Kurmanji, meaning construction cannot be reduced to mere lexical equivalents. On the contrary, the cultural codes underlying idiomatic expressions, syntactic flexibility, and morphological fluidity require far more complex processing capacities.

Even ChatGPT-4o, which performed best among the evaluated models, exhibited inconsistencies in morphologically generative tasks. This highlights the difficulty current LLM architectures face in balancing semantic adequacy with morphological precision. Notably, the models' failure to accurately generate culturally embedded and multi-layered expressions—such as *xalên min*—indicates the need for training approaches that are sensitive not only to data quantity, but also to local linguistic dynamics.

The findings underscore a critical distinction between "accurate translation" and "meaningful transfer." A structurally correct response may fall short in cultural relevance—reminding us that language is not merely a formal system, but also a vessel for memory, belonging, and lived experience. Consequently, models that fail in cultural and idiomatic transfer should be questioned not only on technical grounds but also on ethical ones.

In conclusion, this study argues that evaluating a language model requires more than measuring accuracy scores; it necessitates examining how meaning, context, and cultural memory are represented. Developing a successful AI system for languages like Kurmanji demands not just more data, but a higher-quality approach to representation—one that honors the soul of the language.

Discussion

This study reveals that training large language models (LLMs) solely on multilingual datasets does not equip them with sufficient sensitivity to the layered processes of meaning construction, morphological analysis, and cultural transfer in low-resource languages. Evaluations conducted in the Kurmanji dialect particularly highlight stark performance differences in idiomatic and contextual interpretation tasks—differences that the models fail to bridge. These findings should be interpreted in the broader context of ongoing debates around "linguistic justice" and representational equity (Bender & Friedman, 2018; Blodgett *et al*., 2020).

Since current models are often trained on institutionalized and normative variants of dominant languages, they tend to generate surface-level, decontextualized responses when faced with structures that require cultural encoding. This implies that LLMs must assume not only the role of technically accurate generators but also responsible agents of linguistic identity representation. Justice in cross-linguistic representation is not merely a matter of access—it is fundamentally about the ability to convey meanings that carry cultural weight.

Another key finding of the study relates to morphological failure. The inflectional systems of languages like Kurmanji exhibit far greater derivational variation compared to high-resource languages such as Turkish or English. This reveals that LLM architectures grounded in assumptions of universal grammar struggle to recognize the phonological and morphological characteristics of localized languages (Ponti *et al*., 2021).

Importantly, these challenges cannot be explained by technical shortcomings alone. The unfairness embedded in data production processes—where local communities are treated as passive data sources rather than active knowledge contributors—further deepens the problem of representation. In this regard, community-led NLP initiatives (Bird, 2020) and training strategies based on local knowledge systems offer promising, ethical, and sustainable solutions.

In essence, this discussion is not merely a model evaluation—it is a reflection on the question: *Whose voice is being heard, and how?* The meaning of a language lies not only in its form but also in its lived reality, context, and culture. If these dimensions are not integrated into the modeling process, the success of LLMs may be technical—but it will not be representative.

Declarations

Ethics and consent

Ethical approval and consent were not required for this study. No human participants, personal data, or sensitive content were involved in the research.

Data availability statement

All data supporting the findings of this study—including prompt sets, model outputs, and evaluation rubrics—are openly available at Zenodo:

Zenodo. Kurmanji LLM Evaluation Dataset (ChatGPT-4o, Gemini, Claude, Grok) [Data set]. https://doi.org/10.5281/zenodo.16993575 Uzun (2025).

This project contains the following underlying data:

- kurmanji_llm_sample_dataset.csv (sample dataset including model outputs, scores, and error annotations)
- full_morphological_test_data.csv (complete morphological test data with gold-standard annotations)
- proverb_translation_outputs.csv (annotated model translations of proverbs with cultural/idiomatic evaluations)

Data are available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

References

Amani A, MohammadAmini M, Veisi H: **Kurdish spoken dialect recognition using X-vector speaker embedding**. In: A. Karpov, R. Potapova, & I. Mporas (Eds.), *Speech and Computer: Proceedings of the 23rd International Conference, SPECOM 2021*. St. Petersburg, Russia, Springer, October, 2021; 50–57.
Publisher Full Text

Bender EM, Friedman B: **Data statements for natural language processing: Toward mitigating system bias and enabling better science.** *Transactions of the Association for Computational Linguistics.* 2018; **6**: 587–604.
Publisher Full Text

Bimagambetova Z, Rakhymzhanov D, Jaxylykova A, *et al.*: **Evaluating large language models for sentence augmentation in low-resource languages: a case study on Kazakh**. In: *Proceedings of the 2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS).* IEEE, 2023; 14–18.
Publisher Full Text

Bird S: **Decolonising speech and language technology.** In: D. Scott, N. Bel & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).* 2020; 3504–3519.
Publisher Full Text

Blodgett SL, Barocas S, Daumé H, *et al.*: **Language (technology) is power: a critical survey of "bias" in NLP.** In: D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2020; 5454–5476.
Publisher Full Text

Castaldo A, Monti J: **Prompting large language models for idiomatic translation**. In: B. Vanroy, M.-A. Lefer, L. Macken, & P. Ruffo (Eds.), *Proceedings of the 1st Workshop on Creative-text Translation and Technology.* European Association for Machine Translation, 2024; 32–39.
Reference Source

Chaka C: **Currently available GenAI-powered large language models and low-resource languages: any offerings? Wait until you see.** *International Journal of Learning, Teaching and Educational Research.* 2024; **23**(12): 148–173.
Publisher Full Text

Coeckelbergh M: **LLMs, truth, and democracy: an overview of risks.** *Sci Eng Ethics.* 2025; **31**(1): 4.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gatt WJ: **An investigation of lexical cohesion in Kurdish EFL students' emails.** *Journal of the College of Basic Education.* 2022; **20**(85): 789–803.
Reference Source

Omer NA: **Analysis of the idioms of (head) in Kurdish language, from the perspective of cognitive linguistics.** *Journal of University of Raparin.* 2024; **11**(5): 177–197.
Publisher Full Text

Omer SK, Hassani H: **Idiom detection in Sorani Kurdish texts**. (Version 3) [Preprint]. arXiv. 2025.
Publisher Full Text

Ponti EM, Glavaš G, Vulić I, *et al.*: **The role of morphology in universal representations: the case of dependency parsing.** *Transactions of the ACL.* 2021; **9**: 1571–1588.

Qadri R, Mostafazadeh Davani A, Robinson K, *et al.*: **Risks of cultural erasure in large language models (arXiv: 2501.01056).** *arXiv.* 2025;
Reference Source

Qian M: **Automating idiom translation with cross-lingual natural language generation grounded in semantic analyses using large language models.** In: M. Martindale, J. Campbell, K. Savenkov, & S. Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations).* Association for Machine Translation in the Americas, 2024; 95–115.
Reference Source

Rarick CA, Winter G, Barczyk C, *et al.*: **Iraqi and Kurdish cultural values in the semi-autonomous state of Kurdistan.** *Global Journal of Management and Business Research: A Administration and Management.* 2014; **14**(3): 75–81.
Reference Source

Smart A, Hutchinson B, Amugongo LM, *et al.*: **Socially responsible data for large multilingual language models.** *arXiv.* 2024; (arXiv: 2409.05247).
Reference Source

Uzun C: **Kurmanji LLM Evaluation Dataset (ChatGPT-4o, Gemini, Claude, Grok).** [Data set]. Zenodo. 2025.
http://www.doi.org/10.5281/zenodo.16993575

Zhong T, Yang Z, Liu Z, *et al.*: **Opportunities and challenges of large language models for low-resource languages in humanities research.** *arXiv.* 2024; (arXiv: 2412.04497).
Reference Source

Open Peer Review

Current Peer Review Status: ✗ ✗

Version 1

Reviewer Report 23 December 2025

https://doi.org/10.21956/openreseurope.22580.r63836

© **2025 YEN-MIN J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✗ **JASMINA KHAW YEN-MIN**
Universiti Tunku Abdul Rahman, Kampar, Malaysia

1. The presentation is generally strong, but clarity suffers from inconsistent linguistic terminology and structural mixing. The Introduction mixes general Kurdish discussions with specific mention of Zaza and Kurmanji without clearly establishing their relationship or the political/linguistic context of their distinction. The authors must clarify early on that the focus is Kurmanji and ensure all references to other dialects (like Zaza or Sorani) are contextually justified.
2. Separate the current mixed introductory and literature review material into distinct sections: Introduction (stating the research gap and aim) and Related Work (reviewing prior studies).
3. Restructure the presentation of Results/Discussion/Findings. Results should exclusively present the quantitative scores, data, and figures (e.g., Figure 1-5). Discussion should interpret these results and relate them to the broader literature and ethical implications. The current "Findings" section (page 6) should be integrated into the Discussion or Results.
4. Add a dedicated, formal Conclusion section to summarize the main takeaways, reaffirm the core argument (the need for cultural representation), and outline future work.
5. Provide a clear, technical definition of the term "deep micro-corpus" used in the Methodology section.
6. Explicitly state the exact sources for the initial 300 linguistic items. The description must go beyond general categories and name specific public sources, corpora, literary works, or detail the elicitation procedures (e.g., from specific Kurmanji regions or speakers) to validate representativeness.
7. Provide a solid justification and rationale for the specific numerical splits and categorization of the test set (e.g., why 60 proverbs/idioms, 40 verbs, 30 modal transformations, etc.). This should ideally reference existing linguistic analysis of Kurmanji or best practices in LLM evaluation literature.
8. Explicitly state how the native-speaking annotators were selected and qualified.
9. Provide more detail on the assessment rubric used. This includes expanding on the simple 1-5 scale and core criteria listed to provide concrete examples of what constitutes a score of 1 vs. 5 for each prompt type (Meaning Interpretation, Form-Based Production, Culture-Based Transfer).
10. Explicitly map the three comparison types (Meaning, Form, Culture) presented in the Results/Discussion sections back to the five clear research questions outlined in the Introduction .

This demonstrates a complete and supported answer set.
11. The current reporting relies on clear quantitative accuracy scores. The authors must briefly provide a rationale for the level of statistical analysis included (i.e., why inferential statistics were or were not used). If the data permits, including a more detailed statistical analysis would strengthen the paper.

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Natural Language Processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 20 November 2025

https://doi.org/10.21956/openreseurope.22580.r62377

© 2025 Hassani H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hossein Hassani
University of Kurdistan Hewlêr, Erbil, Iraq

The paper investigates the performance of LLMs on processing Kurmanji Kurdish from various perspectives: linguistic accuracy, semantic nuance, and cultural representation.
The topic of the research is timely, particularly considering the less-resourcedness of Kurdish.

The paper should be improved to address the following issues:
- The introduction provides inconsistent information about Kurdish, mixing languages and dialects, focusing on Zaza without any clarification. A collection of resources that explain Kurdish and its varieties exists; they must be consulted and properly referenced. However, the focus should remain on Kurmanji Kurdish (or Northern Kurdish, according to some scholars). Kurmanji is a dialect of Kurdish, not a language on its own. Iraqi Kurds speak various Kurdish dialects.
- Where is the source of the 100 statements? How were they compiled?
- What does "deep micro-corpus" mean?
- How is the methodology designed? Is it based on related work? How did the author decide on the number of idioms? Where are the supports for the 40/30/30 split of the idioms? What is the rationale behind the prompts? Is the authority an authority in the field? If yes, that must be based on their previous work that explains the rationale.
- Results and findings are confused. Findings are presented before the results' presentation.
- The research questions in the introduction are not revisited clearly; instead, there are three topics for comparison that have not been related to the compiled questions.
- The paper does not have a conclusion.
Some trivial issues:
- The introduction and related work have been mixed and are better separated.
- No mention of the entire paper structure in the intro.
- Methodology and results have been mixed.
- English equivalent for all Kurmanji terms in the figures to help the larger audience understand the context.
- Figure 3 is not consistent with the rest of the figures; the box "0% Failed to interpret 0%" is not properly aligned; it is not clear what the two 0%s mean.

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
No

Are sufficient details of methods and analysis provided to allow replication by others?
No

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computer Science; Computational Linguistics; Natural Language Processing; Software Engineering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.