Exploring Detection of Complex, Non-Literal Expressions of Cultural Motifs

Ibrahim H. Alyami[1] **Mark A. Finlayson**[2]
[1]Najran University, King Abdul Aziz Rd, 66462, Najran, KSA
[2]Florida International University, 11200 SW 8th Street, Miami, FL, 33199, USA
ihalmerdef@nu.edu.sa, markaf@fiu.edu

Abstract

Motifs are non-commonplace, recurring narrative elements, often found originally in folk stories and also in modern news, literature, and propaganda. Expressions of motifs in text can be most straightforwardly classified as *simple* or *complex*. Simple motif expressions are easy to detect because they almost always appear in a single sentence using the same words as the motif definition itself. However, complex motifs are strongly non-literal and often spread across multiple sentences, thus requiring more context to understand. We propose a baseline system to detect complex motif expressions that have challenged prior work. We used an annotated corpus that identified 992 complex motif expressions of 155 different motifs for training and testing. We tested five different generative approaches that included varying amounts of context: a single sentence baseline (from prior work); a window of 3 or 5 sentences; the entire story; or the entire story with the target sentence identified. We fine-tuned four off-the-shelf open-source LLMs using LoRA under these conditions. Somewhat surprisingly, we report a negative result: our experiments show that in our generative setup more context did not reliably improve the performance of detecting complex motifs, and often hurt fine-tuned models. We speculate on why this might be so and identify directions for future research.

Keywords: folklore, motifs, the Arabian Nights, natural language processing, neural methods, large language models, linguistic annotation

1. Introduction

Motifs are non-commonplace, specific, recurring narrative elements that are often found originally in folk stories and are more generally deployed in culturally inflected materials. Motifs are interesting because they are a compact source of cultural knowledge: many motifs concisely communicate a constellation of related ideas, associations, and assumptions. For example, "troll under a bridge" is a motif common in Western cultures with roots in Scandinavia. To those familiar with the motif, it entails a number of related ideas that are not directly communicated by the surface meaning of the words: the bridge is along the critical path of the hero, and he must cross it to achieve his goal; the troll often lives under the bridge, crawling out to waylay innocent passers-by; the troll charges a toll or demands something for crossing the bridge; the troll is a squatter, not the officially sanctioned master of the bridge; the troll enforces his illegitimate claim through threat of physical violence; and the hero often ends up battling (and defeating) the troll instead of paying the toll.

While motifs usually originate in folkloristic material, they are frequently used in modern discourse, and motif expressions can be easily found in speeches, news reports, press releases, propaganda, books, and movies; indeed, in any type of language where cultural knowledge is deployed. An excellent example of such modern usage is the Islamist motif *Pharaoh*. The Pharaoh appears in stories found in the Hebrew and Christian Bibles and the Qur'an; in those stories, the Pharaoh comes into conflict with Moses and his attempts to free the Hebrews from Egyptian slavery. The Pharaoh is an arrogant and obstinate tyrant who defies the will of God and is punished for it. In modern Islamist extremist narratives, the Pharaoh is a symbol of struggles against anti-Islamic regimes and has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush of the United States, the last of whom Osama bin Laden referred to as the "pharaoh of the century" (Halverson et al., 2011). Further, one must be familiar with Islamic religious and folkloristic traditions to understand the use of this motif in modern language; its meaning is metaphoric and obscure to those not versed in the tradition of the group.

Motif expressions can be most easily classified into two types: simple and complex. Simple motif expressions match the motif definition (found in a motif index) nearly word-for-word. For example, the motif *Mermaid* is found expressed in the Arabian Nights as: *While he was doing this, the sea became disturbed and out from it came mermaids, the sea's daughters, each carrying in her hand a jewel gleaming like a lamp.* (Irwin, 2010, Volume 2, Night 491). On the other hand, the words used in a complex motif expression don't overlap those used in the definition, with the expression often being non-literal and indirect. Complex motif expressions thus presumably need more context and language understanding in order to detect and interpret.

For example, consider the motif *Magic sphere*

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 40–50
11 May 2026.

burns up country. By turning that part of the globe to the sun, one can make any place on earth burn up. This motif is found expressed in the Arabian Nights as:

> *Whoever has the globe can, if he wants, sit inspecting all lands from east to west and whatever part he wants to see, he can do so by turning the globe where he wants and looking into it. He will then have a view of the land and its people as though they were all there in front of him. If he is angry with any city and turns the globe towards the sun with the intention of burning the city to the ground, this is what will happen. As for the kohl case, whoever uses its contents on his eyelids will see all the treasures of the earth.*

Note that, aside from expression complexity, there is a difference in the complexity of the concept of the motif itself: the idea of a *mermaid*—which refers to a single, albeit imaginary, class of things—is in many ways much simpler than the motif of the magic sphere above. Thus motifs themselves can be simple in their conceptual structure, or complex. We will explore this distinction as well in the work.

While simple motif expressions are relatively easy to detect because of their simple form, automatically detecting and interpreting complex motif expressions is challenging, as has been demonstrated in prior work (Alyami and Finlayson, 2026). In particular, we used an annotated corpus developed in that work which identified 992 complex motif expressions of 155 different motifs. Using these data, we explore tackling the task of automatically detecting complex motif expressions in the folkloristic materials by using more context around the target of the non-literal expressions of cultural motifs. We tested five different generative approaches, which used different amounts of context in the prompt: (1) single sentence baseline (from prior work); (2) 3-sentence window of context; (3) 5-sentence window of context; (4) the entire story; and (5) the target sentence plus the entire story. We tested four LLMs, namely: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, google/gemma-3-4b-it, and Qwen/Qwen3-8B.

This paper is structured as follows. We first provide background on motifs and review related computational work (Related Work). We then describe the dataset (Data). We describe the methods we explored, including fine-tuned LLMs using LoRA (Methods). Finally, we discuss the results, the limitations of the work, and possible next steps (Discussion) and we conclude with a list of our contributions (Contributions).

2. Related Work

2.1. Prior Work

Alyami and Finlayson (2026) developed a model to index motifs automatically from *The Arabian Nights*. They built the model using an annotated dataset that consists of 200 motifs extracted manually from *A Motif Index Of The Thousand and One Nights*. The index contains around 5,000 motifs extracted by El-Shamy from the 207 stories of The Arabian Nights (El-Shamy, 2006). Alyami and Finlayson (2026) analyzed the conceptual complexity of the motifs (classifying them as simple or complex), and also analyzed the complexity of the motif expressions (again, simple or complex), resulting in a four-way classification of motif expressions.

First, **Simple Structure** / **Simple Expression** motif expressions are where both the conceptual structure of the motif and the way it is expressed in the narrative are straightforward. These expressions are relatively easy to detect. For example, *Seven daughters* is simply structured and usually expressed using those exact words: *The Almighty provided him with seven daughters. . .* (Irwin, 2010, Volume 3, Night 784).

Second, **Simple Structure** / **Complex Expression** motif expressions are where a conceptually simple motif is expressed in a complex way. In other words, the motif is expressed in the index in only one or a few words, but in the text the motif is expressed indirectly, obscurely, or across many pages. For example, the motif *Resourcefulness* is conceptually simple, but one expression runs *The mamluks delightedly agreed that this was a good plan, and there and then they began to cut logs for the raft and to twist ropes to bind them together. They worked on this for a month, taking back firewood for the princess's kitchen each evening and devoting the rest of the day to the raft.* (Irwin, 2010, Volume 3, Night 766).

Third, **Complex Structure** / **Simple Expression** motif expressions are where the motif is conceptually complex, but still the expression of the motif in the text closely follows what is found in the index, and so keyword retrieval usually suffices to find them. For example, the motif *Apparently dead persons revived when certain thing happens. Proper prince appears, or the like.* is not simply structured but can be expressed almost exactly as is found in the index, as in *You Muslims, you soldiers, have you ever in your lives seen a man die and then come back to life?* (Irwin, 2010, Volume 1, Night 34).

Fourth and finally, **Complex Structure** / **Complex Expression** is the most difficult type of motif expression to find. The motif is conceptually complicated and motif expression is also complex. For example, the motif *What you (deal) to others will*

be done (dealt) back to you is found in the text as follows:

> *I took hold of the horse and mounted it. It didn't move and so I kicked it, and when it still refused to move, I took the whip and struck it. It didn't move and so I kicked it, and when it still refused to move, I took the whip and struck it. As soon as it felt the blow, it neighed with a sound like rumbling thunder and, opening up a pair of wings, it flew off with me, carrying me up into the sky way above the ground. After a time, it set me down on a flat roof and whisked its tail across my face, striking out my right eye and causing it to slide down my cheek. It then left me and I came down from the roof to find the ten one-eyed youths. No welcome to you, they said. Here I am, I replied. I have become like you, and I want you to give me a tray of grime with which to blacken my face and to let me sit with you. No, by God, they said, you may not do that. Get out! (Irwin, 2010, Volume 1, Night 16).*

The motif is found in across approximately 190 words and 10 sentences, where none of the words in the motif definition are found directly in the expression.

Alyami's model was trained on single sentences that were annotated for the presence or absence of particular motifs. The most effective method for detection was a fine-tuned Llama3, achieving an overall F_1 performance of 0.90 for simple motif expressions (across both conceptually simple and complex motifs). Their study shows that this task is still challenging for state-of-the-art LLMs when trying to detect complex motifs, achieving an overall F_1 performance of 0.72.

Other work has also emphasized that automatically detecting and interpreting motif expressions in modern language is a challenging problem (Yarlott and Finlayson, 2016; Yarlott et al., 2022; Yarlott, 2022; Yarlott et al., 2024; Acharya, 2022; Acharya et al., 2024). Prior work defined the task of *motif detection*, which is finding a motif expression in non-folkloristic materials. In that task, even a word-by-word expression of a motif must be further differentiated into MOTIFIC, EPONYMIC, REFERENTIAL, or UNRELATED types (Yarlott et al., 2024). An EPONYMIC usage is the use of the motif as a name; a REFERENTIAL usage is a mention of the motif itself; while a MOTIFIC usage is intended to call to mind the implicit associations of the motif. Here we tackle a different task, which is detecting appearances of a complex motif in the original folkloristic materials, which we will call *detection of complex, non-literal expressions of cultural motifs*. Methods that address this task would allow the automatic mining of positive and negative examples to train and test other stages of motif understanding, such as discovery of novel motifs (i.e., *motif discovery*: Yarlott and Finlayson, 2016), identification of motif usage in non-folkloristic materials (i.e., *motif detection*, as above: Yarlott et al., 2022, 2024), and interpretation of the meaning of motific language (i.e., *motif interpretation*: Acharya, 2022; Acharya et al., 2024).

2.2. Culture in LLMs

Recent research has highlighted the challenges LLMs face in understanding cultural elements from text. Adilazuarda et al. (2024) explained the need for context, or called "thick description" by Geertz (1973). The concept of thick description or context is needed in order to understand a culture as insider. The context needed to analyze the internal or specific small details of that cultural text to understand not only behaviors or certain events, but also cultural, religious, and psychological norms which will make LLMs more culturally aware. On the other hand, "thin description" also coined by Geertz (1973), is not context-aware analysis, but rather when an outsider frames the text without understanding why certain behaviors or habits will be interpreted by cultural insiders differently depending on the situation. Adilazuarda et al. notes that most of the current research on the cultural understanding of LLMs focuses mainly on values and emotions. However, there are many other aspects of cultures text can and should be explored.

The challenge of figurative language understanding across cultures has also been explored more broadly. Kabra et al. (2023) created a multilingual figurative language inference dataset (MABL) across seven culturally diverse non-English languages. Their work showed that figurative expressions are deeply rooted in culturally specific concepts such as food, religion, and events with overlap between languages from the same geographic region. Liu et al. (2024) evaluates Multilingual Large Language Models (mLLMs) to reason using cultural common ground by using proverbs and sayings from many languages as an exploratory method. Proverbs are culturally specific expressions. They collect proverbs and their usage in conversational situations from six languages. They examine distinct mLLMs to assess their capacity to memorize proverbs, use proverbs and sayings to reason in varied situational circumstances, and comprehend proverbs in cross-cultural conversations. They created a dataset called Multicultural Proverbs and Sayings (MAPS) for proverb understanding with conversational context for different languages. Their study found that mLLMs possess knowledge of proverbs and sayings to varying

degrees, but memorizing a proverb does not indicate the ability to reason with it in context. They also found significant culture gaps when reasoning across languages, with performance on English data consistently stronger than other languages, particularly for lower-resource languages such as Bengali and Indonesian. Park et al. (2025) stated that LLMs often succeed in recognizing figurative expressions at the sentence level but their ability to use them coherently in conversation remains uncertain. They showed that even models that recognize figurative expressions at the sentence level fail to use them appropriately in dialogue. Our work extends this line of work to the domain of folkloric cultural motifs where expressions are not only figurative but often span multiple sentences and require deep cultural background knowledge.

With this in mind, our paper evaluates LLMs on a diverse set of motifs that cover different cultural aspects of folktales. The motif index by El-Shamy (2006), which is where we source our motif lists, categorizes motifs based on their cultural, social, and psychological contextual significance. For example, a motif can represent a cultural theme such as *Tabu: eating with left hand.* (El-Shamy, 2006). In Islamic-Arabic culture, it is taboo to eat with your left hand due to the cultural habits. In the narrative, the motif is expressed in Irwin (2010) as follows:

> *I went and prepared the necessary food, drink, and so on, which I then presented to him, inviting him to eat in the Name of God. He went to the table and stretched out his left hand, after which he ate with me. This surprised me, and when I had finished, I washed his hand and gave him something to dry it with. I then sat down to talk, after I had offered him some sweetmeats. 'Sir,' I said to him, 'you would relieve me of a worry were you to tell me why you ate with your left hand. Is there perhaps something in your other hand that causes you pain?*

Motifs may also have social significance: *No low rank person would be sitting down while addressing high rank* is an example of how motifs can represent the social theme in the narrative (El-Shamy, 2006). It is expressed in Night 620 in Irwin (2010) as follows:

> *Uthman was both stupid and conceited, and when he arrived at Judar's palace he saw a eunuch seated on a chair in front of the door. This man did not get up on his arrival, and in spite of the fact that there were fifty men with 'Uthman, it was as though no one had come. 'Uthman went up to him and said: 'Slave, where is your master?' 'In the house,' the eunuch replied, and as he spoke he continued to lounge on his chair.*

Third, a *Clothes make the man* is an example of how motifs can represent psychological themes, as follows:

> *The weaver went along and saw magnificently dressed people receiving fine foods and being treated with respect by the host because of their splendid clothes. He said to himself: 'Were I to change my trade for one that would be of less trouble, more prestigious and more rewarding, I could collect a lot of money and buy clothes like these. I would then become important; people would respect me and I would be like these other's.*

A related concern is that LLMs are biased to certain cultures, especially when prompting in English. The reason behind that is that the datasets used to train these LLMs are mostly in English and represent Western cultures and values (Tao et al., 2024; Johnson et al., 2022; Atari et al., 2023). The work we present here is an example of moving beyond the Western tradition, in that the motif index used in our data is specific to the Middle East, and contains mainly Muslim cultural material.

3. Data

We began with the data that was collected and used in prior work, by Alyami and Finlayson (2026). In that work, Alyami and Finlayson used the *Motif Index of The Thousand and One Nights* (El-Shamy, 2006), which contains around 5,000 motifs extracted by El-Shamy from the 207 stories of The Arabian Nights. These motifs overlap with motifs found in the Thompson Motif Index (TMI), while adding new motifs specific to the Arabic content of the stories (Thompson, 1955-1958). El-Shamy generally followed the TMI classification scheme (i.e., 23 themes, each identified by a letter), but he also added categories and information corresponding to dimensions of cultural, social, and psychological contextual significance. The annotated dataset comprised 58,450 annotated sentence-motif pairs, of which 2,670 were positive examples. These annotations used 200 unique motifs. There were examples of motif expressions from all four types (simple-simple, simple-complex, etc.). From these data, we selected only the complex expressions. This subset of the data comprised 992 sentences with positive examples. It represents 155 unique motifs. Note that Table 1 reports the number of unique motifs per split (112 for training, 24 for validation, and 19 for testing) and not only the number of motif expressions. The full set of 992 complex

Conceptual Complexity ↓	Uniq. Motifs	Motif Exps.	Training / Validation / Testing
Simple	108	689	81 / 17 / 10
Complex	47	303	31 / 7 / 9
Overall	155	992	112 / 24 / 19

Table 1: Number of unique motifs, number of complex motifs expressions (positive examples), and the training / validation / testing set sizes (in number of unique motifs) broken down by the conceptual complexity of the motifs.

Data	Count
Unique Stories	64
Total Sentences in Unique Stories	1,393
Total Tokens in Unique Stories	34,513
+Examples in the Index	266
Motifs w/ 1 +Example in the Index	175
Motifs w/ >1 +Example in the Index	26
Sentences-Motif pairs, +Examples	2,670
+Examples Total Tokens	88,523

Table 2: Data

motif expressions is distributed across these splits in a balanced way. We paired these with an equal number of random sentences that didn't contain any motif expression, resulting in 1,984 total examples[1]. The dataset is split into training, validation, and test sets. Table 1 shows the breakdown of motifs and motif expressions by the conceptual complexity of the motifs (simple or complex). Each training, testing, or validation comprises a unique list of motifs so that each model is tested on an unseen list of motifs. Additional details on the data are found in Table 2.

Annotation of complex motif expressions is inherently challenging, as it can be a tedious and, at times, challenging task depending on the complexity of the data. The task is to label sentences as True Positive (TP) when the motif is present and False Positive (FP) when it is not, and the sentences and motifs fall into four categories: Simple Expression, Complex Expression, Simple Structure, and Complex Structure. It is essential to ensure every element of the motif is present when annotating, as annotators might feel tempted to label something as TP because they want it to be, rather than because it truly meets the criteria — a bias that often occurs after a long string of FPs, and mislabeled data can pollute the dataset, leading to inaccurate model training. These sentences are isolated from their broader context, which can cause difficulties. For example, proper names may appear without their associated titles, leading to potential mislabeling. The difficulty of the annotation task itself provides important context for interpreting model performance if human annotators also struggle with complex motif expressions, model scores below human-level performance are expected rather than surprising.

[1]The code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/VKJEGZ. The original full dataset is available from the authors Alyami and Finlayson (2026).

4. Methods

We evaluated five different generative approaches to solving the task. In all five approaches, the task is formulated as a binary classification per text-motif pair: given a specific motif and a text. The model must predict whether that particular motif is expressed in the text. Each text is evaluated against exactly one motif at a time. The baseline is a single-sentence approach (§4.1) as demonstrated in Alyami and Finlayson (2026). Then we experimented with a three sentence context window (target sentence plus one sentence before and after), a five sentence context window, the entire story, and finally, the target sentence coupled with the entire story.

We experimented with four open-source LLMs fine-tuning using Low-Rank Adaptation (LoRA): Mistral (Mistral-7B-Instruct-v0.3) (May 2024), Llama (Llama-3.1-8B-Instruct) (July 2024) and Google (gemma-3-4b-it) (March 2025), and Qwen (Qwen3-8B) (May 2025) (Mistral AI, 2024; Meta, 2024; Gemma Team, et al., 2025; Qwen Team, et al., 2025). We used the training/validation/test dataset we showed in Table 1.

We fine-tuned all LLMs using Low-Rank Adaptation (LoRA), a parameter-efficient method (Hu et al., 2022). The goal was to see an improvement in classifying whether a text contains a motif or not with different amounts of context (3-sentence window, 5-sentence window, entire story, or target sentence plus entire story). We fine-tuned the LLMs using the same dataset we used in fine-tuning embedding models. Each training example consists of motif-sentence pairs as input, target output (Yes or No) only. We set epochs to 5, the learning rate to 10^{-4}, batch size to 16, temperature to 0, and new tokens to 1 only in order to force the models to be deterministic.

Illustrative Example

To show how each of the five context methods is applied. We walk through an example using the motif *Eblis: born as one of the fourteen children of*

Khalit and Malit. He disobeyed his father by refusing to marry one of his seven twin-sisters, and was transformed into a worm (which became Eblis). The motif is expressed in Night 493 in Irwin (2010) as follows:*Then, when it was the four hundred and ninety-third night, SHE CONTINUED:I have heard, O fortunate king, that when Buluqiya had told King Sakhr the full story of his wanderings from beginning to end, the king was filled with astonishment and ordered his servants to fetch tables, which they spread with cloths. Then they brought plates of red gold, of silver and of copper. On some of them were fifty cooked camels and on others twenty, while some contained fifty sheep. In all there were one thousand, five hundred plates, and when Buluqiya saw that he was amazed.The company then ate, as did Buluqiya, who, when he had had enough, gave thanks to Almighty God. After that, the food was removed and was replaced with fruit. When they had all finished eating, they called down praises on Almighty God and blessings on His Prophet, Muhammad, may God bless him and give him peace. Buluqiya was surprised to hear the name of Muhammad and he asked the king if he might put some questions to him. 'Ask what you want,' the king told him, and so he said: 'O king, what are you? What is your origin and how do you come to know of Muhammad, so that you call down blessings on him and love him?''Buluqiya,' replied the king, 'Almighty God created hellfire in seven layers, one on top of the other, each separated by the distance of a thousand years' journey. The first of these layers is called Jahannam and it has been prepared for those Muslims who disobey God's commands and die without having repented. The second layer is called Lazan and this is prepared for the unbelievers, while the third is Jahim, prepared for Gog and Magog. The name of the fourth layer is al-Sa'ir and this is for the people of Iblis; the fifth is called Saqar and is for those who abandon prayer. The sixth is al-Hutama and is for Jews and Christians, while the seventh is al-Hawiya, which has been prepared for the hypocrites. These are the seven layers.''I suppose,' said Buluqiya, 'that the punishments of Jahannam are easier to bear than all the others as it is the uppermost layer.' The king agreed with this, but added: 'In spite of that, Jahannam contains a thousand mountains of fire, in each of which there are seventy thousand valleys, each containing seventy thousand cities of fire.In each of these cities there are seventy thousand fiery castles, with seventy thousand fiery rooms in each, and each room contains seventy thousand couches of fire, with seventy thousand forms of torment in every one of them. None of the other layers, however, have any lighter punishments than these, as this is the first layer, while as for the other layers, only God Almighty knows the number of their torments.'When Buluqiya heard what the king had to say he collapsed in a faint, and when he recovered he burst into tears and said: 'O king, how then will it be with us?' 'Have no fear,' said the king, 'for you must know that the fire will not burn anyone who loves Muhammad, and for his sake, may God bless him and give him peace, such a man will be freed, while hellfire will flee from all who follow his religion. As for us, Almighty God created us from fire, and the first beings that He created in Jahannam were two of his host, the first called Khalit and the second Malit. Khalit was shaped like a lion and Malit like a wolf. Malit's tail was feminine, piebald in colour, while Khalit's was masculine, in the shape of a tortoise, and was a twenty-year journey in length. God then ordered these two tails to join together and copulate, and from them were born snakes and scorpions who live in hellfire and, having reproduced and multiplied, are used by God to torture those who enter it. God then ordered the two tails to copulate a second time, and when they did this, Malit's tail was impregnated by the tail of Khalit and gave birth to seven males and seven females. These were nurtured until they grew up, and when they had done so, the females were married to the males. All but one of them were obedient to their father; the one who disobeyed became a worm and this worm is Iblis, may God Almighty curse him. He had been one of the cherubim, serving God until he was raised to heaven, where he found favour with Him and became the leader of the cherubim.'Morning now dawned and Shahrazad broke off from what she had been allowed to say.* The story contains the following sentences (numbered for reference):

> s_{-2}: Malit's tail was feminine, piebald in colour, while Khalit's was masculine, in the shape of a tortoise, and was a twenty-year journey in length.
> s_{-1}: God then ordered these two tails to join together and copulate, and from them were born snakes and scorpions who live in hellfire and, having reproduced and multiplied, are used by God to torture those who enter it.
> s_0: **God then ordered the two tails to copulate a second time, and when they did this, Malit's tail was impregnated by the tail of Khalit and gave birth to seven males and seven females.**
> s_{+1}: These were nurtured until they grew up, and when they had done so, the females were married to the males.
> s_{+2}: All but one of them were obedient to their father; the one who disobeyed became a worm and this worm is Iblis, may God Almighty curse him.

Single Sentence (§4.1): Only s_0 is provided to the model without any surrounding context.
3-Sentence Window (§4.2): The model receives s_{-1}, s_0, and s_{+1}. This provides minimal local context.
5-Sentence Window (§4.2): The model receives s_{-2}, s_{-1}, s_0, s_{+1}, and s_{+2}. This provides a broader local context.
Entire Story (§4.3): The full story text is provided. In this case, the model must determine whether the motif appears anywhere in the story.
Target Sentence + Entire Story (§4.4): The full text of the story is provided as background context. We ask the model to classify the target sentence s_0 using the entire story as background.

4.1. Single Sentence

In this baseline, we fine-tuned the four LLMs on the training and validation sets shown in Table 1, with the goal of teaching the models to understand the smallest possible descriptions of motifs. We fed the motif a single positive or negative sentence and the models were asked to classify the sentence. The requested answer is *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the sentence.
Rules: Answer ONLY "Yes" or
"No". Do not explain.
Motif: <Motif>
Sentence: <Sentence>
Answer:
```

4.2. Window of Target Text (3- or 5-sentence)

Here, we add either two or four sentences of context (one or two sentences before and after), and asked the models to classify if the target sentence with this window contains the motif or not. The answer is again only *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the target text.
Context is provided ONLY to
resolve ambiguity Rules: Answer
ONLY "Yes" or "No". Do not
explain.
Motif: <Motif>
Context Before: <Pre Target
sentence>
Target Sentence: <Target
Sentence>
Context After: <Post Target
Sentence>
Answer:
```

4.3. Entire Story

For maximum context we feed the entire story in with the motif. Following previous modes where the model is asked to classify if the story contains the motif or not by only answering *Yes* or *No* using the following prompt:

```
Task: Decide if the motif is
present anywhere in the story
text.
Rules: Answer ONLY "Yes" or
"No". Do not explain.
Motif: <Motif>
Story Text: <Story>
Answer:
```

4.4. Target Sentence plus Entire Story

In this approach, we fed the models the sentence to be classified, plus the entire story for complete context, and asked the models to classify whether the story contains the motif in the target sentence or not. Again, the requested answer is *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the target sentence.
The story is background only.
Hard rule:
- Answer 'Yes' ONLY if the TARGET
sentence itself clearly expresses
the motif.
- If the motif is only elsewhere
in the story, or only implied
weakly, answer 'No'.
- When in doubt, answer 'No'.
Answer ONLY 'Yes' or 'No'. Do
not explain.
Motif: <Motif>
Story (background): <Story>
Target Sentence: <Target
Sentence>
Answer:
```

5. Discussion

The results are shown in Table 3. In our setup, increasing context does not consistently improve performance. Fine-tuned models often degrade as context length increases, especially in the entire-story mode. However, some zero-shot models benefit from added context. We suspect that this is because motifs are a relatively small part of the text, and the models have trouble focusing on such a small portion. Compared to the entire story, the LLMs, in most cases, failed to detect the complex motif expressions for both simple and complex motif conceptual complexity classes. While all models failed in most cases, Mistral-FT shows significantly weaker performance when increasing the context. This might suggest a direction when comparing

Conceptual Complexity ↓		Complex Expressions				
		Single Sentence	±1 Window	±2 Window	Entire Story	Target Sentence+ Entire Story
Simple	Mistral-Zero-shot	0.31 (0.66 / 0.20)	0.44 (0.73 / 0.31)	0.39 (0.68 / 0.27)	**0.65** (0.55 / 0.80)	0.45 (0.69 / 0.34)
	Mistral-FT	0.65 (0.74 / 0.58)	0.52 (0.67 / 0.43)	0.50 (0.71 / 0.39)	0.10 (0.33 / 0.06)	0.26 (0.56 / 0.17)
	LLama3-Zero-shot	0.64 (0.57 / 0.73)	**0.65** (0.60 / 0.70)	0.59 (0.59 / 0.59)	0.63 (0.59 / 0.67)	0.44 (0.55 / 0.37)
	LLama3-FT	**0.73** (0.81 / 0.67)	0.50 (0.73 / 0.39)	0.59 (0.82 / 0.46)	0.60 (0.63 / 0.57)	0.46 (0.44 / 0.47)
	Gemma3-Zero-shot	0.65 (0.53 / 0.83)	0.36 (0.50 / 0.29)	0.22 (0.32 / 0.17)	0.63 (0.49 / 0.86)	0.60 (0.48 / 0.81)
	Gemma3-FT	0.41 (0.59 / 0.31)	0.34 (0.51 / 0.26)	0.39 (0.61 / 0.29)	0.62 (0.53 / 0.74)	0.63 (0.55 / 0.73)
	Qwen-Zero-shot	0.45 (0.67 / 0.34)	0.62 (0.71 / 0.56)	**0.64** (0.65 / 0.63)	0.55 (0.72 / 0.44)	**0.64** (0.56 / 0.75)
	Qwen-FT	0.58 (0.72 / 0.49)	0.29 (0.65 / 0.19)	0.21 (0.53 / 0.13)	0.62 (0.70 / 0.56)	0.57 (0.49 / 0.66)
Complex	Mistral-Zero-shot	0.46 (0.91 / 0.31)	0.49 (0.85 / 0.34)	0.37 (0.73 / 0.25)	0.53 (0.67 / 0.44)	0.30 (0.100 / 0.17)
	Mistral-FT	0.72 (0.91 / 0.60)	0.68 (0.86 / 0.56)	0.72 (0.90 / 0.59)	0.06 (0.50 / 0.03)	0.14 (0.33 / 0.09)
	LLama3-Zero-shot	0.64 (0.85 / 0.51)	0.64 (0.89 / 0.50)	0.61 (0.88 / 0.47)	0.47 (0.63 / 0.38)	0.21 (0.60 / 0.13)
	LLama3-FT	0.70 (0.86 / 0.59)	0.60 (0.93 / 0.44)	0.52 (0.86 / 0.38)	0.53 (0.57 / 0.50)	0.38 (0.38 / 0.39)
	Gemma3-Zero-shot	0.57 (0.71 / 0.47)	0.36 (0.67 / 0.25)	0.36 (0.62 / 0.25)	0.62 (0.49 / 0.84)	0.57 (0.45 / 0.78)
	Gemma3-FT	0.59 (0.73 / 0.50)	0.43 (0.71 / 0.31)	0.41 (0.44 / 0.38)	0.59 (0.48 / 0.75)	0.51 (0.44 / 0.61)
	Qwen-Zero-shot	0.74 (0.91 / 0.63)	**0.81** (0.92 / 0.72)	**0.76** (0.91 / 0.66)	**0.65** (0.94 / 0.50)	**0.59** (0.67 / 0.52)
	Qwen-FT	**0.84** (0.80 / 0.88)	0.75 (0.88 / 0.66)	0.75 (0.88 / 0.66)	0.60 (0.61 / 0.59)	0.43 (0.47 / 0.39)
Overall	Mistral-Zero-shot	0.36 (0.75 / 0.23)	0.46 (0.77 / 0.32)	0.38 (0.69 / 0.26)	**0.63** (0.57 / 0.69)	0.42 (0.73 / 0.29)
	Mistral-FT	0.67 (0.78 / 0.58)	0.57 (0.73 / 0.47)	0.57 (0.78 / 0.45)	0.09 (0.36 / 0.05)	0.23 (0.50 / 0.15)
	LLama3-Zero-shot	0.64 (0.62 / 0.66)	0.65 (0.66 / 0.64)	0.60 (0.65 / 0.55)	0.59 (0.60 / 0.58)	0.39 (0.56 / 0.30)
	LLama3-FT	**0.72** (0.82 / 0.64)	0.53 (0.79 / 0.40)	0.57 (0.83 / 0.43)	0.58 (0.62 / 0.55)	0.44 (0.42 / 0.45)
	Gemma3-Zero-shot	0.63 (0.56 / 0.72)	0.36 (0.54 / 0.27)	0.26 (0.40 / 0.20)	0.62 (0.49 / 0.85)	0.59 (0.47 / 0.80)
	Gemma3-FT	0.47 (0.64 / 0.37)	0.37 (0.57 / 0.27)	0.40 (0.53 / 0.31)	0.61 (0.51 / 0.75)	0.59 (0.52 / 0.70)
	Qwen-Zero-shot	0.55 (0.76 / 0.43)	**0.68** (0.78 / 0.61)	**0.67** (0.71 / 0.64)	0.58 (0.78 / 0.46)	**0.63** (0.58 / 0.68)
	Qwen-FT	0.68 (0.76 / 0.61)	0.47 (0.77 / 0.33)	0.42 (0.73 / 0.29)	0.61 (0.67 / 0.57)	0.53 (0.49 / 0.59)

Table 3: Overall System Evaluations: F_1 (precision / recall).

two texts of different lengths. One approach could be to summarize the target text. However, it will be challenging for a system to summarize a story while ensuring that the relevant motif remains in the summary; to ensure this in all cases, the system will need to already have the ability to detect the motif. Another possible approach is to increase the length of the motif with more description, so that the model has enough indications of where the motif could be located. In general, our observation that more context actually hurts detection for a small piece of target text (at least under our experimental setup) is aligned with a prior work that discusses the 'Needle in a Haystack' tasks, where they show that smaller relevant contexts degrade LLM performance (Bianchi et al., 2025).

When looking at the precision and recall breakdown, we notice that the performance drop with increasing context is driven by a collapse in recall rather than precision. On the other hand, in the fine-tuned models, precision remains relatively stable across context windows. For instance, Mistral-FT maintains precision around 0.67–0.91 across single-sentence and windowed modes but the recall drops sharply when moving to entire-story mode (e.g., from 0.58 at single sentence to just 0.06 for simple-conceptual motifs, and from 0.60 to 0.03 for complex-conceptual motifs). This implies that fine-tuned models become too cautious with additional information. This effect is consistent across both simple and complex conceptual motifs but particularly noticeable for complex-concept motifs under the entire-story mode. Interestingly, the windowed modes (±1 and ±2) preserve much of the single-sentence performance. For instance, Mistral-FT holds F_1=0.68–0.72 across windows for complex-concept motifs. This suggests that a small amount of local context does not hurt but a large context window is harmful. By contrast, zero-shot models show the opposite pattern where the recall rises substantially at the entire-story level (e.g., Mistral-Zero-shot recall reaches 0.80 for simple-concept motifs and 0.44 for complex-concept motifs on the entire story compared to just 0.20 and 0.31 at single sentence mode). On the other hand, precision suffers. This implies that when models are fine-tuned on short-span text they learn to look for specific words in short text. However, when a motif is spread across a long story, they suffer when trying to detect these complex motif expressions.

Comparing fine-tuned and zero-shot models shows that LoRA fine-tuning with a small data size gives mixed results. For short contexts, fine-tuning helps in some cases. For instance, LLama3-FT achieved F_1=0.73 which outperforms LLama3-Zero-shot at F_1=0.64 for simple-concept motifs, and Qwen-FT achieved F_1=0.84, outperforming Qwen-Zero-shot at F_1=0.74 for complex-concept motifs at the single-sentence level. However, this is not consistent across all models. Gemma3-FT, for instance, underperforms Gemma3-Zero-shot

in nearly every mode. On the other hand, for longer contexts, fine-tuning is clearly harmful. Zero-shot models consistently outperform fine-tuned models, especially on the entire-story mode. For example, Mistral-Zero-shot reaches F_1=0.65 for simple-concept motifs versus just 0.10 for Mistral-FT, and Qwen-Zero-shot leads complex-concept motifs with F_1=0.65 versus 0.60 for Qwen-FT. Overall, LLama3-FT achieves the best single-sentence F_1 of 0.72, but Qwen-Zero-shot is the strongest model across all wider-context modes (F_1=0.68, 0.67, and 0.63 for ±1 window, ±2 window, and Target Sentence + Entire Story, respectively). Mistral-Zero-shot is leading on the entire-story mode at F_1=0.63. This may suggest that fine-tuning on limited data causes models to focus on short-span patterns, which could reduce their ability to reason over longer contexts. Collecting more training data and exploring other fine-tuning strategies could help address this gap.

6. Contributions

We fine-tuned four open-source LLMs using LoRA using 992 complex motif expressions using varying amounts of context: a single sentence baseline (from prior work); a window of 3 or 5 sentences; the entire story; or the entire story with the target sentence identified. The most effective model was Llama3, achieving an overall F_1 performance of 0.72 in a single sentence context. The study shows that this task remains challenging for state-of-the-art LLMs, and that additional context does not provide a consistent benefit across models and prompting modes (e.g., sentence, windows, or the entire story), with Qwen being the most effective model in the wider-context settings, achieving an overall F_1 score of 0.63. We release code and data for reproducing this study [2].

7. Limitations

The first limitation is that while our experiments show that increasing context does not improve performance under these specific experimental conditions, it is hard to generalize from this to *all* potential experimental conditions. Therefore, although our results are relatively straightforward and suggest that more context does not help these kinds of models, we cannot completely eliminate the possibility that some experimental setup using generative models of this kind won't respond positively to additional context.

A second limitation of this work is the amount of annotated data available to train the system. For future work, we believe that collecting more complex expression motifs would be valuable, since until now there is no source for these motif expressions except the prior work of (Alyami and Finlayson, 2026). While their data contains a small number of annotated examples (155 motifs with 992 positive examples), we believe that continuing to expand the annotated data will allow the development of even more capable models. It would also be useful to have a way of precisely and ideally automatically assessing the complexity of both motif conceptual structure and motif expressions. Right now, complexity judgments are done manually, and a more precise standard would allow more careful separation of hard and easy examples for training and testing the systems. Another limitation is that we did not evaluate the model when there is another motif in the same text. In the current methodology, each text is evaluated against only one motif. This is an interesting open problem to see how state-of-the-art models perform when a text contains motif X and also contains a related but distinct motif Y. Not only that, but also how these modes lead to clear boundaries between these related motifs. We leave this robustness analysis to future work.

[2]The code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/VKJEGZ.

8. Acknowledgments

This work was supported in part by a Saudi Arabian Cultural Mission Fellowship to Ibrahim H. Alyami from the College of Computer Science and Information Systems at Najran University, Saudi Arabia [grant number 443-16-40].

9. References

Anurag Acharya. 2022. *Integrating Cultural Knowledge into Artificially Intelligent Systems: Human Experiments and Computational Implementations*. Ph.d. dissertation, Florida International University.

Anurag Acharya, Diego Estrada, Shreeja Dahal, W Victor H Yarlott, Diana Gomez, and Mark Finlayson. 2024. Discovering implicit meanings of cultural motifs from text. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)*, pages 46–56.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Ibrahim H. Alyami and Mark A. Finlayson. 2026. Automated motif indexing on the arabian nights. *arXiv preprint:2603.19283*.

Mohammad Atari, Mona Xue, Peter Park, Damián Blasi, and Joseph Henrich. 2023. Which humans? *PsyArXiv*.

Owen Bianchi, Mathew J Koretsky, Maya Willey, Chelsea X Alvarado, Tanay Nayak, Adi Asija, Nicole Kuznetsov, Mike A Nalls, Faraz Faghri, and Daniel Khashabi. 2025. Hidden in the haystack: Smaller needles are more difficult for llms to find. *arXiv preprint:2505.18148*.

Hasan M. El-Shamy. 2006. *A Motif Index of The Thousand and One Nights*. Indiana University Press, Bloomington and Indianapolis.

Clifford Geertz. 1973. *The interpretation of cultures*. New York: Basic Books.

Gemma Team, et al. 2025. Gemma 3 technical report.

Jeffry Halverson, Steven Corman, and H Lloyd Goodall. 2011. *Master narratives of Islamist extremism*. Palgrave Macmillan, New York.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Robert Irwin. 2010. *The Arabian Nights: Tales of 1,001 Nights*, volume 1-3. Penguin, London, UK.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3.

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

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

Meta. 2024. Llama-3.1-8b-instruct. Hugging Face model repository.

Mistral AI. 2024. Mistral-7b-instruct-v0.3. Hugging Face model repository.

Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. Fluid qa: A multilingual benchmark for figurative language usage in dialogue across english, chinese, and korean. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30268–30282, Suzhou, China.

Qwen Team, et al. 2025. Qwen3 technical report.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Stith Thompson. 1955-1958. *Motif-Index of Folk-Literature, Volumes 1-6: A Classification of Narrative Elements in Folk Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*. Indiana University Press.

Victor Yarlott. 2022. *Communicating with Culture: How Humans and Machines Detect Narrative Elements*. Ph.d. dissertation, Florida International University.

W Victor Yarlott, Anurag Acharya, Diego Castro Estrada, Diana Gomez, and Mark Finlayson. 2024. Golem: Gold standard for learning and evaluation of motifs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7801–7813.

W Victor H Yarlott and Mark A Finlayson. 2016. Learning a better motif index: Toward automated motif extraction. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

W Victor H Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark A Finlayson. 2022. Finding trolls under bridges: Preliminary work on a motif detector. *arXiv preprint:2204.06085*.