Evaluating LLM-Generated Explanations of Metaphors – A Culture-Sensitive Study of Danish

Bolette S. Pedersen[1], **Nathalie Sørensen**[3], **Sanni Nimb**[3],
Dorte Haltrup Hansen[1], **Sussi Olsen**[1], **and Ali Al-Laith**[1,2]
Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark[1]
Department of Computer Science, University of Copenhagen, Denmark[2]
The Danish Language and Literature Society, Denmark[3]
bspedersen@hum.ku.dk, nats@dsl.dk, sn@dsl.dk
dorteh@hum.ku.dk, saolsen@hum.ku.dk, alal@di.ku.dk

Abstract

We examine how well Danish culture-specific metaphors are explained by two of the best performing language models for Danish, ChatGPT and Llama. For comparison, the explanations are measured against how well cross-lingual (or 'universal') metaphors are explained by the models; referring here to metaphors that exist in Danish as well as *across* cultures and languages and in particular in English. To perform our study, we compile a pilot dataset of 150 Danish metaphors and idioms divided tentatively by culture specificity. We prompt the two models and perform a careful qualitative evaluation of the explanations against a four-graded scale. Our studies show that both models are heavily biased towards English since they have much more success in explaining the metaphors that also exist in English than the culture-specific ones, relying presumably on erroneous transfer from English when dealing with the latter. In particular, the *sentiment* of the culture-specific metaphors seems to be often 'lost in translation'. We further claim that this strong colouring towards English poses a serious problem in the era of LLMs with regards to developing and maintaining cultural and linguistic diversity in other languages.

1 Introduction

Metaphorical expressions are an essential part of language and offer considerable cognitive benefits in both oral and written communication by making the content much more personal and engaging (Noveck et al., 2001; Citron and Goldberg, 2014; Prabhakaran et al., 2021). Metaphorical language is therefore also highly frequent and occur with reference to both universal, culture-specific and personal aspects of life. In other words, metaphors and idiomatic expressions provide an advanced tool for humans to express themselves in abstract and complex situations with reference to highly culture-specific, personal, and opinion-oriented values (Lakoff and Johnson, 1980).

With the recent advancements of large language models (LLMs), however, using metaphors in communication is no longer exclusive to humans. Chatbots like ChatGPT produce and interpret metaphors when they communicate, and they do so with apparent fluency and equilibrism, in particular for English. A more careful look into the use of metaphors in language models, however, exposes quite a lot of serious problems and cultural biases, even if it is hard to pinpoint exactly from where these problems arise. Some may be due to unbalanced training data where some languages are prioritised over others and thereby causing erroneous language transfer and cultural hallucinations (Zhang et al., 2023; Cao et al., 2023) and (Myung et al., 2024). Others may derive from a general lack of grounding of the language models with respect to physical objects and spacial conditions, and therefore a lack of ability to 'see' which features from a concrete sense are transferred to the metaphorical meaning; a deficit that may decrease in future with language models becoming increasingly more multi-modal (Szot et al., 2024).

Under all circumstances, cultural biases in the use of and interpretation of metaphors become particularly evident and problematic when working with the models on medium-resourced languages like the Scandinavian ones. Standard techniques for evaluating the language models in terms of large-scale benchmarks that are often both rigid and simplistic in nature do not reveal a fully nuanced picture of how this complex figure of speech is dealt with by the models, as mentioned for Dan-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 470–479
March 3-4, 2025 ©2025 University of Tartu Library

ish in (Pedersen et al., 2024).

In order to gain better knowledge and understanding of the models' treatment of metaphor in our own language, we therefore aim at i) compiling a pilot metaphor dataset, which is culturally sensitive in that it is developed from Danish language resources and from the point of perspective of the Danish society, and ii) providing qualitative evaluations by Danish native speakers of the explanations given by the models on these metaphors.

For our study of model performance, we choose the currently two best performing models in Danish according to the Scandeval Benchmark, namely ChatGPT and Llama, as reported in (Nielsen, 2023, 2024). Both chatbots are based on high-performing multilingual transformer models that are well-suited for the kind of conversation on metaphors that we are interested in with our experiments. Where Llama is a partly open-source model, GPT is a proprietary model. We are however only exploring the models via prompting.

The paper is organised as follows: To position our work, we refer in Section 2 to related work on metaphors in linguistics, lexicography, and NLP. Further, in Section 3 we describe the creation of the culture-specific pilot dataset of Danish metaphors, looking into the typical source and target domains reflecting cultural aspects of the Danish society. Section 4 is devoted to our model experiments with ChatGPT and Llama and explains how we have prompted the models about Danish metaphors in both Danish and English and with and without a textual context. We describe in Section 5 our procedure for evaluating the LLM-generated explanations against a four-graded scale and discuss the annotation agreement results. In Section 6 we show and analyse the results and compare how the models deal with culture-specific vs. cross-cultural metaphors, and to which extent the two models differ in performance. All data are made freely available from github [1]. Finally, in Section 7 we conclude and sketch out how our experiments might be scaled up in future work and hopefully used for model improvement.

2 Related Work

Metaphors have been studied intensively in linguistic theory for decades and are considered an essential figure of speech that is closely related to our conceptual and cognitive system as well as to our culture. The work of Lakoff and Johnson (Lakoff and Johnson, 1980) constitutes a landmark in this line of research in stating that metaphors are fundamentally a basic means of understanding complex concepts of feelings and abstractions through mappings from more concrete and directly understandable domains. They further underline that the most fundamental values in a culture will be coherent with the metaphorical structure of the most fundamental concepts in the culture.

In recent linguistic studies, focus has further been into getting a deeper understanding of the underlying cognitive processes of metaphors (Bambini et al., 2019), as well as achieving consensus both monolingually and across languages of what constitutes a metaphor, often referred to as the Metaphor Identification Procedure (MIP) (Crisp et al., 2007; Nacey et al., 2019; Sanchez-Bayona and Agerri, 2024). Other works examine how metaphor relates to other figurative figures of speech such as irony, sarcasm, and hyperbole (Badathala et al., 2023), (Burgers et al., 2018).

In lexicography, conventionalised metaphors are typically described as specific word senses and most often also labelled explicitly in the dictionary as figurative/metaphor. In many cases the metaphor is also structurally related to its concrete sense in the form of a subsense/main sense relation. One example is the verb *to splice* in the Oxford English Dictionary (OED.com), where a subsense to the first sense of the verb is described as *In various transferred and figurative uses: To unite, combine, join, mend.* Also Svensk Ordbok (Swedish Dictionary, svenska.se) marks figurative subsenses (as in *fönster* ('window') with the label *äv. bildligt* ('also figurative')). In other cases, the dictionaries simply mention the figurative meaning as part of the concrete sense description. In the Danish Dictionary (Det Danske Sprog- og Litteraturselskab, 2024), however, metaphors are almost always described as subsenses labelled 'metaphorical' or 'slang' making them thereby easy to identify and extract for our present study.

Further, recent wordnet studies suggest a lexical metaphor representation, called ChainNet, where

---

[1] `https://github.com/kuhumcst/danish-semantic-reasoning-benchmark/tree/main/metaphors`

the link to the concrete meaning is highly explicit and where features from the concrete sense that are transferred to the metaphorical sense are described in a systematic way in terms of so-called **feature transformations** (Maudslay et al., 2024).

In NLP, metaphors are also a topic of interest since understanding and representing them is one of the most challenging tasks to deal with in the field. In particular, it has been questioned to which extent LLMs generalise over the metaphorical meanings and represent the reference to the source domain, or whether they memorise them (Pedinotti et al., 2021; Aghazadeh et al., 2022; Wachowiak and Gromann, 2023). Knowledge graphs of metaphorical facts have further been studied as a means to represent the metaphor relations in the models in order to improve performance (Peng et al., 2021).

3 The Danish Pilot Metaphor Dataset

3.1 Single Word Metaphors and Metaphorical Idioms

For our study, we have compiled a pilot dataset comprising 150 Danish metaphors of which 75 are single word metaphors (as in *sejle* ('to sail') and 75 are metaphorical idioms (as in *høste frugterne* meaning 'reap the fruits'). All are extracted from The Danish Dictionary facilitated by the aforementioned main/subsense structure and by information on metaphorical use. In the editing process, most of the senses in the dictionary were assigned a (not published) value from a set of 152 different domain labels established as part of the dictionary project. This underlying information allows us to identify figurative senses within similar source domains such as agriculture and nautical terms. In the case of the metaphorical multiword expressions - which contain no information on domain, neither on the relation to a concrete sense - we rely on the domain information of the central lemmas in the expression.

3.2 Culture-specific vs. Cross-cultural Metaphors

A central aim of our experiment is to develop and test a culturally sensitive dataset of metaphors in Danish since we hypothesise that these may cause specific problems and expose specific weaknesses and bias in the language models. For each of the two types of metaphors (single-word or multi-word), 50 words/idioms were therefore selected for being *culture-specific* to Danish (compared to English). In addition, a smaller set of 25 words/idioms that *do* exist correspondingly in English were selected for comparison. The datasets were validated by two informants who tested the (translated) metaphorical expressions in a network of English native speakers in order to confirm to which extent they are used also in their mother tongue. As commented on by our informants, the task of deciding whether an expression is culture-specific vs. cross-cultural was in fact not always truly binary since several grey-zone examples exist. In several cases approximate expressions do exist in English but not with the *exact* same selection of words from the source domain. In all such cases, however, we chose a restrictive approach and labeled the Danish expressions as culture-specific since there were no exact matches in English.

Such a grey-zone example is the metaphorical use of *studehandel* in Danish referring to a (political) agreement where two parties give a bargain on their overall ideological principles in order to each achieve short-term benefits. The concrete literal translation to English is 'stud trading'; however, a translation of the Danish metaphor into English would rather be 'horse trading' since 'horse' is the animal typically used in English to convey the same kind of agreement. Likewise, *myreflittig* in Danish has the literal translation 'ant diligent', but the corresponding metaphor in English would be 'busy as a bee', i.e. using another insect from the source domain to express a similar if not exactly the same meaning.

3.3 Typical Source and Target Domains of Danish Metaphors

We aim towards representing a selection of specific Danish traits of culture through a number of typical source and target domains of metaphors, i.e. domains which represent central aspects of the Danish society. As for the **source domains**, these include in particular the domain of farming and agriculture as found in examples like *håndplukning* (lit. 'handpicking', fig. 'carefully selecting a specific person for something, for instance a professional position'), *malkning* (lit. 'milking', fig. 'to achieve money or information in a reckless manner'), *gøde jorden*, (lit. 'to fertilise the soil', fig. 'to provide the prerequisites for something to happen'), and *tærske langhalm*, (lit. 'thresh long straw', lit. 'to speak too much about the same topic without providing new information') to give just a few. Also related to the old farming community are needle work metaphors as in *rendemaske* ('running stitch') referring in a derogative way to a 'roving person'.

The shipping domain is also central to the Danish self-understanding as represented by a long list of nautical metaphors such as *kæntre* (lit. 'capsize'), *ballast* (lit. 'ballast'), and *sikker havn* (lit. 'safe harbour') just to mention a few. Interestingly enough, however, our informants made clear that these metaphors have many direct equivalents in English, probably due to the inherent cross-cultural nature of shipping. This goes for the ones mentioned above; an exception though is the term *splejse*, (lit. 'to splice (a rope)') which in Danish refers very specifically to sharing a bill.

Last but not least, a set of miscellaneous domains are represented in our dataset, referring to e.g. animals as in *haj* (lit. 'shark') referring to someone with good skills or *kylling* (lit. 'chicken') referring to someone with a cowardly behaviour. Many animal metaphors exist in a similar way in English, however, often with a slightly different connotation. More clearly culture-specific for Danish are different kinds of miscellaneous foods and artifacts, e.g. using *klejne* (lit. 'twisted cookie') as a reference to money or *koks* (lit. 'coke') as a reference to disorder and chaos.

Finally, it should be noted that in spite of their frequency, bodily anchored metaphors like *tage hånd om noget* (lit. 'take hand around something' meaning 'deal with something') and *få fod på* (meaning 'get a foothold on something') are not prioritised in our dataset since we overall consider them as being quite universal in nature and thus not particularly specific to Danish culture even if the specific lexical choices may differ in many cases.

Regarding the **target domains**, a majority of the selected metaphors are typical conceptual metaphors in the sense that concrete concepts map onto more abstract ones conveying an abstract or mentally complex meaning as seen in e.g. *hønsegård* (lit. 'chicken coop') which in its metaphorical sense reflects an environment characterized by indifferent talk, gossip, pecking order etc. corresponding approximately to the metaphorical meaning of 'barnyard' in English.

Several of the metaphors selected, however, map an artifact to another artifact, often resulting in a *negative sentiment* of the concrete target as in *havelåge* (lit. 'garden gate') or *skærveknuser* (lit. 'shard crusher') both referring to old creaking bikes; relating indirectly to the fact that bikes are a very common means of transport in Denmark, and that they are not always in a good shape.

In fact, several of the selected metaphors convey a somewhat negative sentiment, presumably referring indirectly to the concept of the famous 'Law of Jante'[2]. This 'law' refers to a strong cultural norm existing particularly in Denmark and Norway that emphasizes humility and collective equality. It basically states that no one should think they are better than others. Examples of metaphors referring to this norm and with a clear negative sentiment are *højbenet* and *højpandet* (lit. 'long-legged' and 'with a high brow') meaning 'being knowledgeable in an arrogant way', *tågehorn* (lit. 'fog horn') referring to somebody who talks a lot in an arrogant and unclear manner, and *flødebolle* ('chocolate candy with a filling of egg white whipped with sugar') referring to a person who is smug and has (too) high thoughts about himself, probably referring back to the fluffy egg-white foam with little substance. *Høj cigarføring* (lit: 'high holding of one's cigar)' meaning being self-conscious and arrogant is another such expression.

4 Experiments

4.1 The Models Selected for Experiments

In the following, we describe the two models selected for the experiments.

ChatGPT: To represent ChatGPT, we use the ChatGPT-4o mini model which became available to the public in the ChatGPT web-interface[3] in July 2024. The model is trained on data up to October 2023. In the experiments, we used a combination of the web-interface and the API to access the model.

Llama: Llama is represented by the Llama 3.1 405B model. This model has 405 billion parameters and like ChatGPT-4o mini, is trained on data

---

[2] The norm was formulated by the Danish-Norwegian author Aksel Sandemose in his novel "A Fugitive Crosses His Tracks" from 1933.

[3] `https://chatgpt.com`

	Danish	English
Isolated	Hvad er den overførte betydning af ordet/udtrykket X, og hvad har det med X i den almindelige betydning af gøre? Har man den samme metafor på engelsk?	What is the metaphorical meaning of the Danish word/ expression X, and what has it got to do with the basic meaning of X? Does the same metaphor exist in English?
In context	Hvad betyder ordet/disse ord X i følgende eksempel: [CONTEXT]?	What does the Danish word/words X mean in the following example: [CONTEXT]?

Table 1: Examples of experimental prompts.

until late 2023. In the experiments, we access the model through the Llama API [4].

For both models it holds that information on the exact training data is not public knowledge. From information on previous model versions, we assume, however, that the vast majority of the training data stems from English texts[5].

4.2 Prompting the Models

We prompt the models using templates. The same templates are used for prompting both models regardless of prompting through a web-interface or API.

We experiment with two types of templates: (1) prompting to explain the metaphor in isolation and (2) prompting to explain the metaphor within a context. Inspired by recent work (Myung et al., 2024), we hypothesise that prompts in English about Danish might provide higher quality explanations than prompts in a less-resourced language like Danish. Thus, we ask both types of prompts in both Danish and English, resulting in four different prompts. The templates for the four prompts are shown in Table 1.

In this setup, we collect a total of 600 responses per model, covering (1) culture-specific and cross-cultural metaphors, (2) metaphors in isolation and within context, (3) prompts (about Danish) in Danish and English, and finally (4) single-word and multi-word metaphors.

5 Qualitative Evaluation of LLM-Generated Explanations

5.1 Evaluation Procedure

All LLM-generated explanations of metaphors were carefully evaluated by a human expert. We used a four-graded scale of evaluation, where a score of '1' denotes a completely wrong explanation of the metaphor with hardly any relevant points well explained, '2' denotes a wrong explanation but with some relevant points explained, '3' denotes a generally correct explanation but with some irrelevant points mentioned, and '4' is given for a correct explanation comprising a reference to all relevant aspects.

Consider below the example of an explanation of *sejle* ('to sail') given by ChatGPT to which the experts have assigned a score of 1:

> *The metaphorical use of "sejle" reflects its basic meaning of sailing. Just as sailing involves steering a vessel through different conditions, the metaphor suggests managing one's course through various life situations or challenges. It implies that, like a skilled sailor, a person is successfully navigating through complexities or adversities with a sense of control and confidence.*

In fact, the metaphorical meaning in Danish is rather the complete opposite, meaning that everything is chaos as shown in the example *hele huset sejler* (lit. 'the entire house is sailing' meaning that it is a complete mess). The nautical idiom *tage bestik* (lit: 'calculate the course (of a ship)') is another example evaluated with 1, since the explanation is based on a completely wrong literal translation of *bestik*, which can also mean 'cutlery' in Danish.

An example of an explanation evaluated with a 2 is the following with the shipping verb *splejse* ('to splice'):

> *Metaphorically, "splejse" is used to describe the act of combining or merging different elements, ideas, or people in a way that creates a cohesive unit. For example, if someone talks about "at splejse ideer" (splicing ideas), they mean integrating or combining different ideas to form a more comprehensive or effective concept. Similarly, "at splejse folk" (splicing people) can refer to bringing people together, such as in networking or forming partnerships.*

In this case, the merging aspect is correctly explained as a transfer from the concrete handicraft meaning, however, ChatGPT misses the actual meaning of *splejse* in Danish which is used when you pool your money together as a group to buy a gift together for instance, or to collectively pay for a dinner.

---

[4] `https://www.llama-api.com/`

[5] For ChatGPT 3.5 it was reported that only 0,5 % of the training material was in Danish and more than 50 % in English (www.lex.dk).

Further, an evaluation of 3 is given to the following explanation of *mursten* ('brick', referring in Danish to a big, heavy book):

> *Metaphorically, "mursten" (brick) can be used to describe something fundamental or essential, akin to how bricks are fundamental components in construction. For example, if someone refers to a concept or an idea as a "mursten," they are suggesting that it is a crucial building block or a key component in a larger structure or system. Additionally, "mursten" can be used in a more specific sense to describe something that is heavy or burdensome. For instance, referring to a lengthy report or a difficult problem as a "mursten" might imply that it is a substantial, challenging task to handle, much like lifting or working with bricks.*

Where part of this explanation is completely off (and seems to be inspired by the metaphorical meaning of the English *cornerstone*), ChatGPT does in fact capture the essence of the metaphorical meaning in the last part of the explanation, referring to the example 'a lengthy report'.

Finally, a top score of 4 is given by the experts to the following explanation of metaphorical expression *brænde alle broer* (lit: 'burn all bridges'):

> *Brænde alle broer bag sig literally means "to burn all bridges behind oneself." The basic meaning is quite straightforward: it refers to the act of destroying or severing all connections, paths, or relationships behind you, making it impossible to return or go back. The metaphorical meaning of this expression is that someone is severing all ties or relationships with their past, making a drastic change, or taking a bold step forward without the possibility of turning back. This can be applied to various situations, such as leaving a job, ending a relationship, or making a significant life decision.*

It should be noted that we do not as such evaluate the part of the explanation concerning whether or not the model finds that a similar metaphor exists in English. This part was rather used to get an impression of where erroneous transfer might come from, and also served as an extra check of whether the informants might have overlooked something when judging the universality of the metaphor.

5.2 Annotator Agreement

Five expert evaluators were involved in the evaluation of the LLM-generated explanations. All evaluators are computational linguists, three of them with specific expertise in computational lexicography. To calculate the inter-annotator agreement (IAA), three experts annotated 20% of the explanations where the models were prompted with a metaphor in isolation (i.e. without a textual context), resulting in a Cohen's Kappa agreement score of 0.475 for the four-graded scale and 0.684 when collapsing the grading into a binary false/true task (i.e. collapsing score 1 and 2 as false and score 3 and 4 as true). Because of the discrepancies and spread in the grading during this first round, the annotation scheme was further discussed and exemplified among the annotators, and two experts subsequently annotated all explanations of metaphors in isolation, resulting this time in a Kappa score of 0.633 on the four-graded scale and 0.857 when seeing the task as binary (true/false). Overall, this can be considered substantial agreement and suggests that despite the inherent subjectivity of the semantic task (based on a relatively open annotation scheme), the annotators demonstrated a robust consensus on the evaluation. Disagreement cases were partly due to diverging assessments when faced with e.g. Danish misspellings or invented words (most typical for Llama), partly due to different opinions on how much to 'punish' wrong or missing bits of explanations.

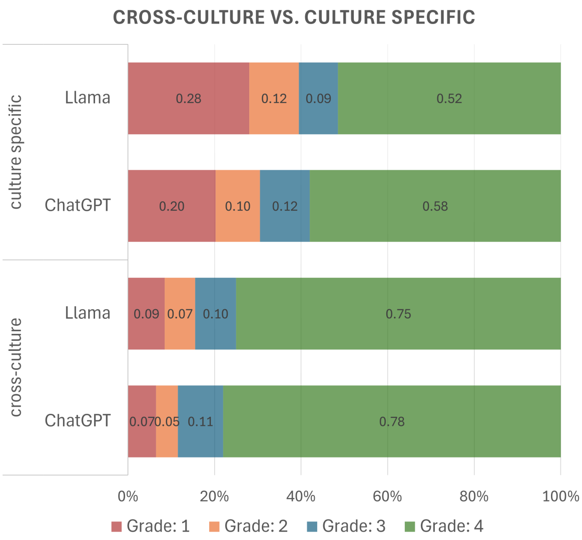

Figure 1: Explanations of Cross-cultural vs. culture-specific metaphors

6 Results and Discussion

As shown in Figure 1, both models have much more success in explaining the metaphors that also exist in English than the culture-specific ones. This indicates that they have too little information on Danish when dealing with the

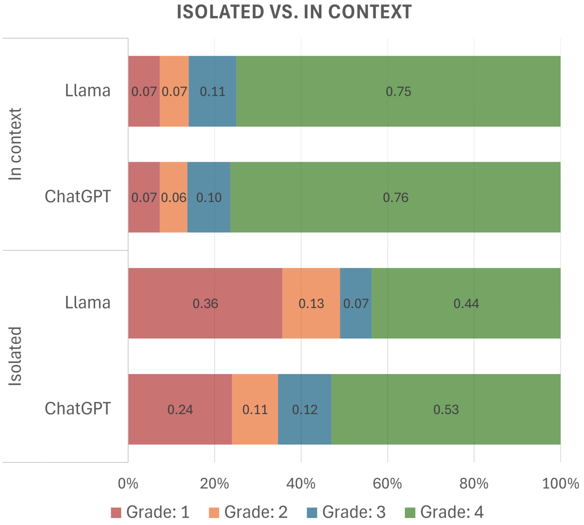

Figure 2: Isolated vs. in context prompts

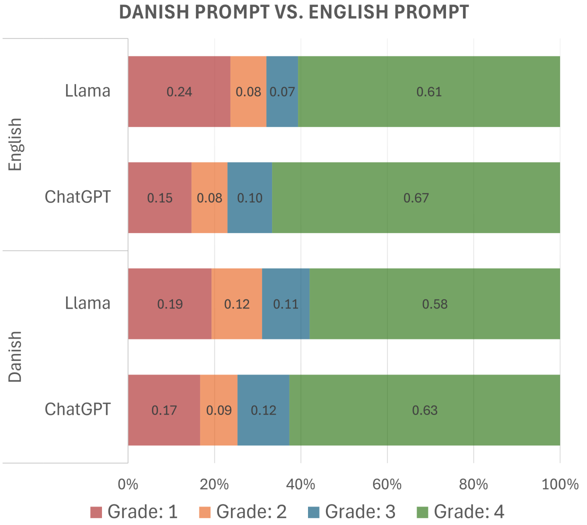

Figure 3: Danish prompts vs. English prompts about Danish metaphors

culture-specific metaphors and therefore hallucinate wrongly (from an English source) in many cases. It can also be seen that the two models differ in quality, with ChatGPT outperforming Llama. For both models, it is clear that when prompted with an example context, they perform remarkably better (Figure 2) than when they are prompted without a context and just asked about the metaphorical meaning. Likewise Figure 4 shows that multiword metaphors are easier to grasp than single word metaphors, presumably due to the multiple words from the expressions themselves. Figure 3 shows that both models perform marginally better when prompted in English about Danish than when prompted in Danish.

Further, we note that metaphors that map from an artifact in the source domain to another artifact in the target domain are generally very hard for the models to predict, and this goes across all experiments. In particular, Llama has a tendency of (amusing) hallucinations, some of which seem to stem from quite simple morphological errors or misspellings. As in the following example where *fis* (lit.'fart' meaning 'fuss' or 'craziness') from the expression *roterende fis i kasketten* meaning 'being crazy' is changed to *fisk* ('fish'). This error evokes a long story about fish swimming round and round in a cap and becoming stressed by it:

> *Udtrykkets oprindelse er usikker, men det er sandsynligt, at det kommer fra en gammel forestilling om, at fiskCould godt biputte sig i en kasket eller en anden lille, lukket beholder. Hvis fisken så begyndte at rotere rundt i kasketten, ville det være et tegn på, at den havde mistet sin normale adfærd og nu var i en tilstand af stress eller panik.*
>
> 'The origin of the term is uncertain, but it is likely that it comes from an old notion that fish could nestle well in a cap or other small, closed container. If the fish then began to rotate around the cap, this would be a sign that it had lost its normal behavior and was now in a state of stress or panic.'

Another example of hallucination regards the explanation of the concrete meaning of the (in fact cross-cultural) idiom *en nål i en høstak* ('a needle in a haystack') about a task which is really hard, in fact almost impossible. Here Llama hallucinates about women working too close to the haystacks with their needles and therefore missing them in the stack.

> *I den almindelige betydning af udtrykket handler det om at lede efter en nål, som er blevet tabt i en høstak, hvilket var et almindeligt problem førhen, da kvinder ofte syede eller reparerede tøj nær høstakke. Høstakke var meget store og hvis nålen faldt ud, kunne det være en stor opgave at finde den igen.*
>
> 'In the common sense of the term, it is about looking for a needle that has been dropped in a haystack, which was a common problem in the past, as women often sewed or repaired clothes near haystacks. Haystacks were very large and if the needle fell out, it could be a big task to find it again'

Another thing that typically goes wrong in both models' explanations, and which is not so well captured by the grading, is the aspect of connotation or sentiment. In quite a lot of cases,

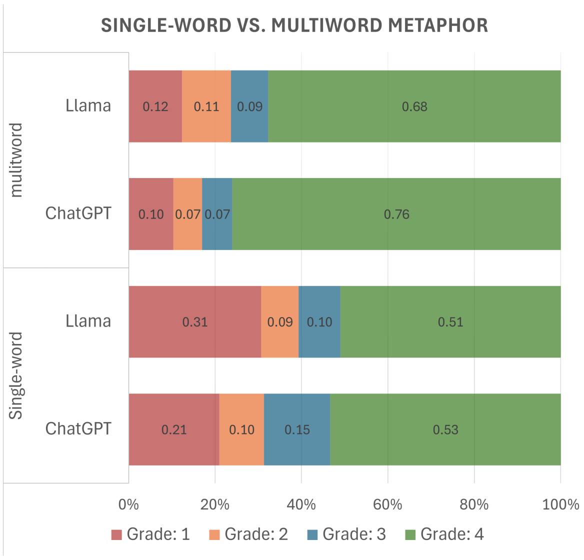

Figure 4: Single word vs. multiword metaphor

the models interpret the metaphors far too positively, missing completely the derogative or negative connotation of the concept and thereby the typical sarcastic Danish ’tone’, which can in some contexts be rather harsh. Examples of such misunderstood metaphors, of which several have already been mentioned and explained above, are: *rendemaske* (lit. ’running stitch’), *’tågehorn*, (’foghorn’), *højbenet*, (lit. ’high-legged’), *højpandet*, (’with a high brow’) *sejle*, (’to sail’), *koks* (lit. ’coke’), *hønsegård*, (lit. ’chicken coop’), *havelåge*, (lit. ’garden gate’), and *skærveknuser*, (lit.’shard crusher’). Likewise with the multiword metaphors *en sang fra de varme lande*, (lit.’a song from the warm lands’ meaning ’an evasive, bland explanation or reply’) and *sejle sin egen sø* (lit. ’sailing your own sea’ meaning ’be left to yourself; deal with your own problems (as a well-deserved punishment)’. Last but not least, similar things happen with metaphors of sexual connotations, which are completely overlooked or ignored by both models.

6.1 Limitations

Our dataset is relatively limited in size and would be improved by being scaled up. On the other hand, all 1,200 automatically generated explanations were carefully human-evaluated providing thereby an interesting set of nuanced observations regarding the performance of the models. Another limitation relates to the fact that we claim to explore culture-specific vs. cross-cultural metaphors in LLMs without going into the more ethnographic discussion of what defines a culture and a language community. We have limited ourselves to look into Danish metaphors and compare them with English because we are aware that a majority of the training material used to train the models is in English. From there on we make a general assumption regarding lack of cultural diversity in current high-performing LLMs. Further, some of the cultural characteristics described for Danish may also count for the other Scandinavian communities, while others may not. Some are reflected also in other Northern European countries, whereas some are uniquely Danish. Furthermore, our informants have only involved British native speakers. This may also have caused some unintentional biases in our dataset where some metaphors may or may not exist in American English compared to British English.

7 Conclusions and Future Work

We have compiled a dataset of culture-specific Danish metaphors supplemented with metaphors that are also found cross-culturally, or more specifically between Danish and English. Our aim was to examine how well the two leading chatbots on Danish explain the metaphors and their reference to the source domain, and to which extent we could see a pattern of decline in quality of the explanations deriving from culture-specific expressions that do not have a parallel in English. Our experiments confirm our hypothesis quite strongly. Culture-specific metaphors are highly complex for the models to interpret, and the explanations indicate that erroneous language transfer from English takes place to a large extent, leading to strongly biased and/or hallucinated explanations. In particular, the models have problems in capturing the right sentiment of the metaphors, distorting thereby the specific Danish ’tone of voice’.

Chatbots like ChatGPT are currently rolled out throughout society, in particular through Co-pilot Enterprise, and people are using them for all kinds of tasks. In this context, the strong colouring towards English that we have documented in our work, indicates that the developing and maintaining of cultural and linguistic diversity is under strong pressure, and that the development might move very fast. This tendency is reinforced by the fact that the hallucinations are very well-formulated and on the surface convincing, meaning that only the highly experienced language user can dismiss garbled output.

To extend our study, we would like to i) expand our Danish dataset, ii) include metaphor studies for the other Scandinavian language, and also iii) go deeper into the understanding of the inner wheels of the models with respect to where the tipping point is found between beneficial language transfer on the one hand and erroneous transfer that leads to cultural biases on the other. One way to proceed in improving the models (in addition to ensuring more Danish training material in the first place) is via fine-tuning or retrieval augmented generation with use of knowledge graphs or other structured information sources. Knowledge graphs can be compiled from culture-specific metaphor lists derived from existing dictionaries or corpora, or from wordnets enriched with feature transformations in a 'ChainNet'-like fashion. All in all enrichments that could potentially lead to better and more culturally diverse language interpretation and generation.

Acknowledgments

Thank you to Birte Pedersen and Guy Norman for acting as informants and for checking in their respective networks to which extent (direct translations of) Danish metaphors and idioms are commonly used in English.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Valentina Bambini, Paolo Canal, Donatella Resta, and Mirko Grimaldi. 2019. Time course and neurophysiological underpinnings of metaphor in literary context. *Discourse Processes*, 56(1):77–97.

Christian Burgers, Kiki Y Renardel de Lavalette, and Gerard J Steen. 2018. Metaphor, hyperbole, and irony: Uses in isolation and in combination in written discourse. *Journal of Pragmatics*, 127:71–83.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, United States. Association for Computational Linguistics (ACL). Publisher Copyright: © 2023 Association for Computational Linguistics.; 1st Workshop on Cross-Cultural Considerations in NLP, C3NLP 2023; Conference date: 05-05-2023.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Peter Crisp, Raymond Gibbs, Alice Deignan, Graham Low, Steen Gerard, Lynne Cameron, Elena Semino, Joe Grady, Alan Cienki, and Zoltán Köcecses. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Det Danske Sprog- og Litteraturselskab. 2024. Den Danske Ordbog. `https://www.ordnet.dk/ddo`. (September 2024).

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Rowan Hall Maudslay, Simone Teufel, Francis Bond, and James Pustejovsky. 2024. ChainNet: Structured metaphor and metonymy in WordNet. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2984–2996, Torino, Italia. ELRA and ICCL.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages.

Susan Nacey, W Gudrun Reijnierse, Tina Krennmayr, and Aletta G Dorst. 2019. *Metaphor Identification in Multiple Languages*. John Benjamins Publishing Company.

Dan Saattrup Nielsen. 2023. Scandeval: A benchmark for Scandinavian natural language processing. *Proceedings of Nodalida 2023, The Faroe Islands.*

Dan Saattrup Nielsen. 2024. Status på Scandinavian embedding benchmark (seb). *Slides from Benchmarkworkshop in The Danish Agency for Digital Government, September 20, 2024.*

Ira A Noveck, Maryse Bianco, and Alain Castry. 2001. The costs and benefits of metaphor. *Metaphor and Symbol*, 16(1-2):109–121.

Bolette Sandford Pedersen, Nathalie C Hau Sørensen, Sussi Olsen, and Sanni Nimb. 2024. Evaluering af sprogforståelsen i danske sprogmodeller - med udgangspunkt i semantiske ordbøger. *NyS, Nydanske Sprogstudier*, pages 8–40.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? Testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204.

Ciyuan Peng, Dang Thinh Vu, and Jason J Jung. 2021. Knowledge graph-based metaphor representation for literature understanding. *Digital Scholarship in the Humanities*, 36(3):698–711.

Vinodkumar Prabhakaran, Marek Rei, and Ekaterina Shutova. 2021. How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 503–512.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *arXiv preprint arXiv:2404.07053*.

Andrew Szot, Bogdan Mazoure, Harsh Agrawal, R Devon Hjelm, Zsolt Kira, and Alexander T Toshev. 2024. Grounding multimodal large language models in actions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.