The Danish Idiom Dataset: A collection of 1000 Danish idioms and fixed expressions

Nathalie Hau Sørensen and **Sanni Nimb** and **Agnes Aggergaard Mikkelsen** and **Jonas Jensen**
Society for Danish Language and Literature
Christians Brygge 1
1219 København K, Denmark

Abstract

Interpreting idiomatic expressions is a challenging task for learners and LLMs alike, as their meanings cannot be deduced directly from their individual components and often reflect nuances that are specific to the language in question. This makes idiom interpretation an ideal task for assessing the linguistic proficiency of large language models (LLMs). In order to test how LLMs handle this task, we introduce a new dataset comprising 1000 Danish idiomatic expressions sourced from the Danish Dictionary DDO (ordnet.dk/ddo). The dataset has been made publicly available at sprogteknologi.dk. For each expression, the dataset includes a correct dictionary definition, a literal false definition, a figurative false definition, and a random false definition. In the paper, we also present three experiments that demonstrate diverse applications of the dataset and aim to evaluate how well LLMs are able to identify the correct meanings of idiomatic expressions.

1 Introduction

Sagen er bøf does not make much sense in English when translated literally, i.e. *the matter is steak*, which obviously doesn't convey the Danish meaning, i.e. *the matter is settled*. When it comes to LLMs, the matter of language proficiency and cultural sensitivity is not yet settled.

In the ideal world, one should be able to get accurate and fluent responses from LLM-based chatbots, such as ChatGPT, even outside of the realm of major languages. In other words, models should be proficient on multiple levels from morphology and syntax to semantics and cultural idiosyncrasies irregardless of the languages involved. However, large tech companies train LLMs on internet data dominated by texts in English and a few other widely spoken languages, resulting in better performance for these languages.

For example, studies have shown that ChatGPT performs better when prompted in English (Zhang et al., 2023; Bareiß et al., 2024) even when the language task is related to another language. Another study suggests that Llama-type models may be internally biased towards English (Wendler et al., 2024). Furthermore, a recent study shows that ChatGPT and Llama struggle to accurately explain Danish culture-specific metaphors (Pedersen et al., 2025). Many of the Danish results seem to be generated on the basis of language transfer, and consequently, they often show a bias towards English and are far better at understanding those metaphors that have English equivalents.

A particularly difficult part of language understanding is idiomatic expressions like *sagen er bøf* where the analysis cannot be based directly on the identification and understanding of each word and where the figurative meaning is culturally specific. The precise knowledge of the meanings of such expressions in Danish reflects a high level of language proficiency among language learners, and we estimate that this is also the case for LLMs.

To facilitate the evaluation of Danish proficiency in LLMs, we have compiled a dataset based on idiomatic expressions in The Danish Dictionary, DDO (Det Danske Sprog- og Litteraturselskab, 2024). The dataset consists of 1000 expressions paired with their actual definitions from the DDO dictionary. Additionally, we have supplemented the data with three false definitions per expression: a literal misinterpretation, a figurative misinterpretation, and a random definition from another idiomatic expression. The aim is to use the combination of correct and false definitions to test LLMs in different scenarios and with different perspectives. In this paper, we present the compilation of the dataset as well as three examples of how the dataset can be used to test an LLM.

In the following section, we present related work. In Section 3, we describe the lexical foundation of the dataset, namely the multiword units in the DDO dictionary, and how the 1000 idiomatic expressions

Proceedings of the 1st NB-REAL Workshop 2025 - Nordic-Baltic Responsible Evaluation and Alignment of Language models, pages 55–63
March 2, 2025 ©2025 Association for Computational Linguistics

are selected. We also describe the process of compiling the false definitions. Finally, we demonstrate three test scenarios and discuss the different ways of using the dataset for evaluation in Section 4.

2 Related work

Our work builds on a continuous effort to make evaluation data in the Nordic languages available. Some notable examples are multilingual benchmarks like ScandEval (Nielsen, 2023) and the Scandinavian Embedding Benchmark (Enevoldsen et al., 2024). Additionally, language understanding is covered by monolingual benchmarks such as Swedish Superlim (Berdicevskis et al., 2023) and the Danish Semantic Reasoning Benchmark (Pedersen et al., 2024).

Within the area of idiomatic expressions, research has focused predominantly on idiom detection rather than comprehension (Tedeschi et al., 2022). However, there are examples of idiom and metaphor datasets in the context of language understanding. For example, ChID (Zheng et al., 2019) is a Chinese idiom dataset based on a so-called cloze task, where models are tasked with selecting the correct idiom to complete a given context. Chakrabarty et al. (2022) likewise created a cloze task inspired dataset, although the task was to select the best continuation to a narrative containing an idiom and thereby test whether the idiom was interpreted correctly. In MiQA (Comșa et al., 2022), they framed the task as selecting the best answer (literal or figurative) to a question which contains a metaphor. Our work builds upon these prior efforts by contributing a new Danish dataset that focuses on idioms and figurative meaning and includes human-written false alternatives. The aim is to facilitate a deeper analysis of figurative language understanding in LLMs and in particular for the Danish language. Our work is closely related to the work by Pedersen et al. (2024) that also explores figurative meaning in Danish. However, our focus is on creating evaluation data rather than exploring the relationship between culture-specific and cross-cultural metaphors.

3 The 1000 Danish idiomatic Expressions

The dataset is structured as a multiple-choice evaluation dataset where the task is to select the correct definition for an idiomatic expression from four options as shown in figure 1.

3.1 Background

The dataset is funded by the Danish Agency for Digital Government as part of the national language technology initiative sprogteknologi.dk, which supports the development of Danish AI and serves as a knowledge hub for Danish language technology resources. The project was launched when a similar dataset for Danish was made unavailable due to licensing issues. We decided to use idioms and their definitions from the DDO dictionary, and at the same time expand the dataset with three kinds of incorrect interpretations in order to make the task more challenging. The different types of false definitions, one of which is concrete, another figurative (but wrong), and a third randomly selected, facilitate a more detailed analysis since incorrect answers can be sorted according to the type of false answer.

For example, if a model frequently selects the literal misinterpretation, it suggests that the model does not recognize the expression as an idiom and consequently finds the literal meaning most plausible. This indicates a lack of abstraction and potentially a broader difficulty in handling Danish text. Likewise, if the model often selects the figurative misinterpretation, it shows that even though the model identifies the phrase as an idiom, it fails to understand its specific meaning. Finally, if the model chooses a random definition from an unrelated idiom, this points to more general issues with task comprehension or proficiency in Danish. This systematic approach provides valuable insight into specific areas where language models can be improved.

3.2 Idiomatic expressions in the Danish Dictionary DDO

Dictionaries generally treat multiword units whose sense is not directly deductible from the senses of the individual words as separate entries (e.g. entities with definitions). In the Danish dictionary DDO, such units constitute more than 13,000 (1/8 of all entries). Many are particle verbs (e.g.*spise op* ‘eat everything which is served’) or multiword terms (e.g. *grøn frø* ‘green frog, Rana esculenta’). In order to create a dataset with idiomatic expressions, we are only interested in multiword units with a metaphorical sense, and especially in those which we consider to be “a concise sentence, typically metaphorical or alliterative in form, stating a general truth or piece of advice; an dage or

Idiomatic expression		Multiple choice selection		Idiomatic expression		Multiple choice selection	
køre med klatten 'ride with the blob'	✓	*være suveræn til noget; brillere* 'be excellent at something; shine'	definition	**i den kolde sne** 'in the cold snow'	✓	*i en tilstand af udelukkelse og forladthed* 'in a state of exclusion and abandonment'	definition
	✗	*køre rundt med en klat* 'ride around with a blob'	literal		✗	*i en (større) bunke sne* 'in a (larger) pile of snow'	literal
	✗	*køre på en hensynsløs og uansvarlig måde* 'drive in a reckless and irresponsible manner'	figurative		✗	*på et tidspunkt forholdsvis tidligt om morgenen* 'at a time relatively early in the morning'	figurative
	✗	*den negative side af noget positivt* 'the negative side of something positive'	random		✗	*noget der er yderst let at forstå* 'something that is very easy to understand'	random

Figure 1: Examples of the correct definition and the three false definitions (literal, figurative, random) from the dataset.

maxim" (the Oxford English Dictionary OED.com: 'Proverb').

Dictionaries present information about such sentences and their metaphorical use in a variety of ways. In the Oxford English Dictionary (oed.com), the information might only be included in the definition text itself (e.g. *bread of idleness* 'bread or food that has not been earned or worked for; also figurative and in figurative contexts'). In the Swedish dictionary Svensk Ordbok (SO) we find cases where it is only hinted at in the definition and/or the example (*ta brödet ur munnen på någon:'beröva någon levebrödet', "strukturomvandlingen tog brödet ur munnen på många anställda"* ('take the bread out of someone's mouth','deprive someone of their livelihood','the structural reform took the bread out of the mouths of many employees')).

The editorial guidelines of the DDO dictionary state that metaphorical senses should be labeled as such. This also includes senses of fixed expressions (see examples A and B), where those that fulfill the criteria mentioned above (OED.com) are furthermore labeled as a specific metaphorical type, *talemåde* ('idiom'), see example C.

A. *tage brødet ud af munden på nogen (overført)* 'forhindre nogen i at arbejde og tjene penge; gøre nogen arbejdsløs'.

'take the bread out of someone's mouth (figurative) prevent someone from working and earning money; make someone unemployed'.

B. *ville give sin højre arm for noget(overført)* være parat til at bringe et meget stort offer for at opnå noget; brændende ønske sig noget

'would give his right arm for something (figurative) be prepared to make a very great sacrifice to achieve something; ardently desire something'

C. *brændt barn skyr ilden (talemåde)* hvis man én gang er kommet galt af sted med noget, undgår man at indlade sig på det igen

'burnt child avoids the fire ('idiom') if you have gone wrong with something once, you avoid getting involved in it again'

However, the information is not always included in the DDO entry, and the distinction between metaphorical sense (*ofø*) and idiomatic sense (*talemåde*) is not always easy to draw. 225 multiword units are labeled *talemåde* ('idiom'), but we find many metaphorical expressions and proverbial phrases of interest for our purpose in the dictionary. Some examples are *sætte tæring efter næring* ('only consume what you can afford'), *her hjælper ingen kære mor* ('not only your dear old mother will be able to help you now'), *blive ved til man styrter* ('keep going until you drop'), and *hver ting til sin tid* ('one thing at a time'). In order to obtain 1000 idioms, we therefore supplement the set of labeled ones in the DDO with a selection of multiword expressions that can be classified as metaphorical or proverbial.

3.3 Data Selection

To avoid having to check 13,000 multiword expressions, we selected only those that fulfilled a number of criteria. One criteria was whether they contain a central lemma, e.g. the nouns *brød* ('bread'), *mund* ('mouth'), *ild* ('fire') in the above examples. We define a central DDO lemma as one with at least one sense linked to the core concepts of Princeton WordNet via the Danish WordNet DanNet, see

COR.SEM (ordregister.dk; corsem.dsl.dk, based on and linked to the DDO). From COR.SEM we also know that central lemmas tend to occur more frequently in multiword units than the rest of the DDO vocabulary. The central five noun lemmas *dag*, *tid*, *hoved*, *hånd*, and ord *have* the largest number of multiword units (containing a noun) in the DDO, all more than 50, *hånd* by far the largest (97). Multiword expressions of central lemmas having many multiword units, i.e. at least three, therefore constitute the fundamental data. The data was extracted from the DDO xml manuscript, from where we also extracted all the 225 labeled idioms. Finally, We supplemented the list with introspectively chosen idioms which were in all cases described in the DDO. A useful way of finding these was to sort the multiword units by length. In the end, we collected around 2747 unique multiword units from DDO as well as their definitions in the dictionary. From this list, we manually selected 1000 idiomatic expressions based on whether it was possible to invent a somehow logical literal explanation and a figurative false description.

3.4 The false definitions

As explained in the above, we supplemented each idiomatic expression with three false definitions, one randomly chosen among other idioms, two which were invented. The task was carried out by four experienced DDO editors. The lexicographers were instructed to write a literal explanation (i.e., what would be the meaning of the sum of the words in the expression) as well as an alternative metaphorical one which did not correspond to how the expression is commonly used in Danish, but which should in some way be plausible.

Writing the false metaphorical definitions proved more challenging than expected. The ideal definition would pick up on a word or phrase in the idiomatic expression and metaphorically expand on that to create a new definition. An example is *gå op i sømmene* (lit. ‘come apart at the seams’,’to have a mental breakdown, to go bananas’). The translated false definition ended up being: ’to obsess unnecessarily over (insignificant) details’ and plays on the different meanings of two phrasal verbs *gå op* ‘loosen, open’ and gå op i ‘take an interest in’. The idea was to mimic how someone without detailed knowledge of Danish language and culture might plausibly misinterpret the idiom when encountering it for the first time. However, it was sometimes difficult to imagine a detailed, creative explanation of something that is essentially false.

In the process, some expressions were discarded from the final dataset if the task of coming up with alternative definitions proved too difficult. For instance, the lexicographer might have to give up writing a literal explanation that made logical sense. Some examples are the expressions *tale frit fra leveren* (lit. ’speak freely from the liver’) and *bide hovedet af al skam* (lit. ‘bite the head off all shame’); consequently these expressions were left out.

The form of the false definitions also had to resemble the style and follow the DDO guidelines of definition writing in order to make the test more challenging for the language model. Several rounds of revising and proofreading the 2000 invented definitions were necessary in order to capture the style of vocabulary and syntactic structure associated with the DDO. Furthermore, the average length of the false definitions turned out to be shorter than the length of the correct definition in the DDO dictionary. Many of these had to be expanded and sometimes even completely rewritten.

The random false definition were collected by shuffling all the correct definitions in the dataset and reassigning them. Since idiomatic expressions can be synonymous or near-synonymous, we run the risk of randomly assigning a definition which may correspond with the correct definition. Thus, part of the proofreading task was to check for potential overlaps. In such cases, we inserted another random definition.

4 Experiments

We set up three experiments using ChatGPT 4o-mini to illustrate how the evaluation dataset can be used. Our purpose was not to exhaustively evaluate the most common models used in Danish, but rather to show how flexibly the dataset can be used in different setups. We regard these experiments as pilot studies to inspire future work.

4.1 Multiple-choice benchmark

The first experiment illustrates the main purpose of the dataset: to create a multiple-choice benchmark dataset that can be evaluated automatically. In our case, we set it up as a multiple-choice task which aims to select the correct dictionary definition of an idiom from the four options described above (the correct definition, the literal false definition, the figurative false definition, and a random and

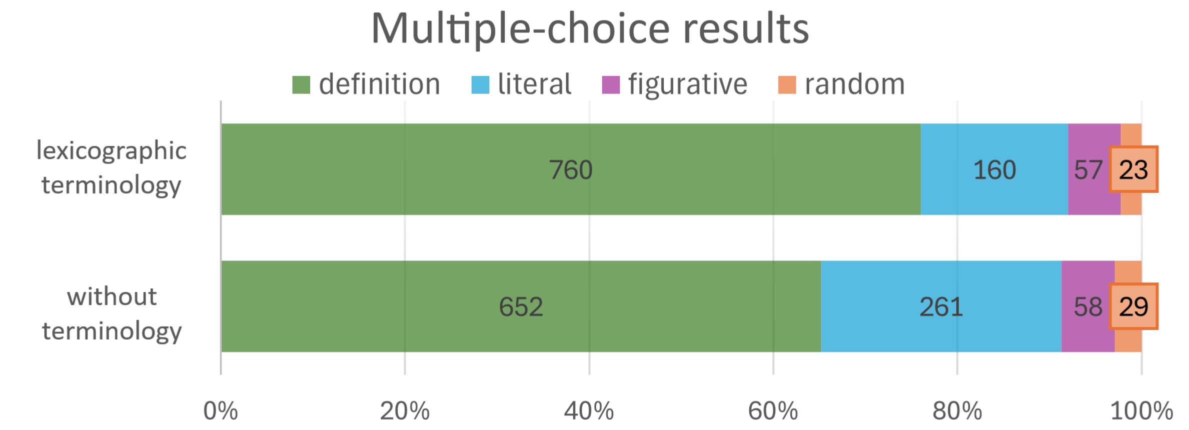

Figure 2: Results for evaluating ChatGPT-4o-mini on the dataset. We used two prompts: with lexicographic terminology (top) and without (bottom).

therefore also false definition).

The hypothesis is that the difficulty of the task will depend on the terminology used in the prompt, and that using terms like 'metaphorical', 'figurative', and 'idiom' would narrow the scope of choices (i.e., disqualify the literal option) thereby removing a step in the language understanding process. Therefore, we evaluate with two different prompts. First, we avoid the use of typical dictionary terms like 'fixed expression', 'definition' and 'idiom' and simply ask which of the four options offers the best explanation for a string of words. In the second prompt we use terms and instead ask which of the four options offers the best 'definition' of a 'fixed expression'. We still avoid terms like 'metaphorical' and 'idiom' to be able to evaluate whether the model is able to grasp the metaphorical meaning by itself.

Figure 2 shows the results for the two prompts, the one that includes lexicographic terms on top and the one that does not below. The best accuracy is achieved by using the prompt that includes lexicographic terms (75,7%), and we find a difference of approximately 10% between the two types of prompts. The difference can mainly be seen in the number of literal false definitions that are chosen while there is only a small difference in the cases of the other two types of false definitions, the figurative and the random one.

Interestingly, although the respective numbers of figurative and random false definitions selected by the model seem similar under the two conditions, the actual overlap is 60% for the figurative category and 35% for the random category. Additionally, in 52% of the cases where the non-lexicographic prompt selects the random definition, the lexicographic prompt chooses the correct definition. A similar pattern is also found for the figurative category, where 31% of the figurative non-lexicographic selections are chosen as correct definition by the lexicographic prompt. The influence of the wording of the prompt went further than the frequency of the literal category. In future work, it would be interesting to experiment with even more prompts to map out the level of influence that the prompt can have on the dataset and what the most optimal prompt could be.

Similarly, we should also test the dataset with setups other than zero-shot and with more models. In particular, it would be interesting to evaluate models aimed at the Danish or Nordic languages. They probably contain more knowledge about Danish culture, and it would be interesting to see whether this has an influence on the performance. Finally, we have to take into consideration that since the correct definitions are already published online at ordnet.dk/DDO there is a risk that they are included in the training data of the LLM's.

4.2 Generative task

Since the rise in popularity of generative models, the lexicographic community has been concerned about the future of dictionaries. If chatbots are able to satisfy the needs of the average dictionary user, it might make dictionaries obsolete and redundant. What if, for example, a chatbot is able to generate a useful explanation without influence from another language or hallucinations when a user encounters an unknown idiom? We investigate this question in experiment 2 by evaluating the quality of ChatGPT 4o-mini's output promted by the question: "*What does the Danish expression [IDIOM] mean?*". The experiment gives us the opportunity to investigate whether the model recognises the idiom and whether it can effectively convey the information to the user. We chose a subset of 100 idioms from the dataset for the experiment selected among those that the model got incorrect at least once in experiment 1. For each of the false definition types, we randomly selected 25 idioms. We expect that these idioms are challenging for the model to explain. We also randomly select 25 idioms that the model correctly classified in the previous experiment. We consider these easy idioms.

We validate the output of the model on a scale from 1-5. The scale is inspired by the evaluation method in Pedersen et al. (2025), however, we added the top category "very correct" to be able to distinguish cases of good explanations where the spelling and formulation are 100% correct from good explanations where it is not.

At the bottom of the scale, we have the grade of 1 which is given to very incorrect explanations with no relevant points mentioned, such as the example below.

> The Danish expression *køre med klatten* translated to 'drive with the pot' but its idiomatic meaning is more akin to 'go with the flow' or 'go along for the ride'.It is often used to describe a situation where someone is taking things easy or accepting something as it comes, rather than actively striving or trying to change the outcome.

This explanation does not line up with the correct definition in any way, as it is 'to be excellent at something, to shine'.

The grade 2 is given for an answer which is also incorrect, however some relevant points are mentioned. The grade 3 is given to an answer which is nearly correct but where some irrelevant points are mentioned, or some important aspects are missing. Below, we show an example of a grade 3:

> The Danish expression *have brød i ovnen* literally translates to 'to have bread in the oven'. Figuratively, it means to have something important or significant going on, such as being busy with a project or having responsibilities that require attention. It can also imply that someone is engaged in something that is developing or in progress.

The expression means 'being pregnant'. The explanation is not completely incorrect, however it is too broad to get a grade higher than 3.

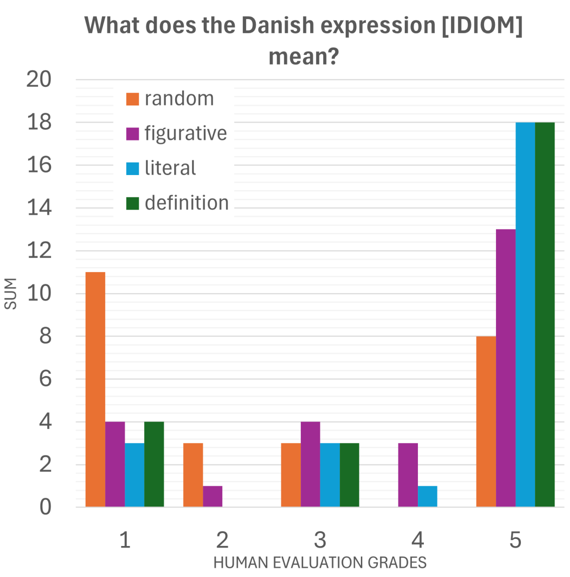

Figure 3: Human evaluation on a subset of the dataset (100 expressions).

Grade 4 covers answers that are correct but with smaller mistakes (e.g., phrasing and spelling), while grade 5 is used for the perfect answer. A good example of a grade 5 can be seen below:

> The Danish expression *for fulde gardiner* literally translates to 'with full curtains'. It is an idiomatic expression that means to go all out or to do something with great enthusiasm and without holding back. It can be used in various contexts to describe someone fully engaging in an activity or experiencing something to the fullest extent.

The results are shown in Figure 3. We see that 73 of the 100 idioms are approved, and that the majority (55) get a perfect grade. This is a high number considering that we deliberately selected the hardest cases according to experiment 1. Among the idioms that got the worst grades (1) and (2), we see a majority from the group of random definitions. These idioms can be viewed as the most challenging since the model previously chose the false definition with the least connection to the form of the idiom. But why are they challenging? Common for most of the idioms that got a low score is their lack of an exact equivalent in English. A similar result has been found in Pedersen et al. (2025) which concludes that Danish culture-specific metaphors are challenging for LLMs predominately trained on English. This could be a problem for smaller languages and cultures as some of the most culturally specific knowledge is less represented, which may not be clear to the user since the model may perform so well in other cases.

In the future, we hope to expand this experiment to other parts of the dataset. The current experiment is based on a deliberate selection of the hardest cases and it would be interesting to explore how a more representative subset would differ. However, it is expensive to manually evaluate the model output and we should thus also look into automatic methods.

4.3 Hallucinations and tone of confidence

A widely known problem with LLMs is that they hallucinate and present false information with a confident tone (Augenstein et al., 2024). Based on the three types of false information per idiom in our dataset, we explore how to use the false information to evaluate the models' ability to disregard false information in the prompt. In the experiment, we prompt the model with a question that implies that one of the false definitions is a plausible interpretation, for instance: "*Explain briefly why [IDIOM] means [FALSE_DEFINITION]*". We aim to answer the questions:

1. Does the model output correct the false information in the prompt?
2. To what degree does the model output contain true information about the meaning of the idiom in Danish?

For the experiment, we use the same subset of 100 idioms from experiment 2, but we replace the true dictionary definitions with literal, figurative, and random false definitions in equal measure. In total, we end up with 34 literal, 33 figurative, and 33 random false ones. As in experiment 2, we prompt ChatGPT 4o-mini and manually evaluate the model output.

Among the 100 answers, only one explicitly states that the expression does not mean the false definition, while another one provides a sort of correction (i.e., *"However, its meaning doesn't directly equate to"*). The next thing that comes close to a correction is the mentioning of metaphorical use, mainly in the cases of using literal false definitions in the prompt.

> The Danish expression *male fanden på væggen* literally translates to 'to paint the devil on the wall.' This phrase is used metaphorically to describe the act of exaggerating or bringing attention to a problem or danger that has not previously existed or been particularly severe.

The figurative or random false definitions sometimes influence the description in the output in such a way that an otherwise correct description of a meaning becomes too broad, specific, or changes its sentiment. In general, it seems as if a lot of effort has been put into connecting the false information from the input with the idiom, which can create quite amusing outputs, e.g. in this case (where the correct meaning of the idiom is to be poor/have no money):

> The Danish expression *ikke have salt til et æg* which translates to 'not have salt for an egg', means 'not being able to make tasty food; being a bad cook'. This idiom draws on the idea that salt is a fundamental seasoning that enhances the flavor of various dishes, including eggs.

To the question of whether the model output still retains true information, we manually grade the outputs on a scale from 1-5, similar to experiment 2. This evaluation task turned out to be much more challenging than in experiment 2. In the beginning, we had a tendency to give a higher grade to outputs with good argumentation rather than comparing the explanation to the actual meaning of the idiom. In particular, the confident tone even for the very incorrect answers was difficult not to be distracted by as a human annotator. We were also not certain on how to grade output that contained a correct definition of the idiom followed by a poor explanation of the connection between the false definition and the idiom. In the end, we attempted to disregard the sections of the answer that discuss the false definition and instead only give a grade based on whether at any point the output contains the correct definition.

The results can be seen in figure 4. Here we see that it is almost only the literal false definitions that still manage to get a good grade. At the opposite end of the scale, we see a surplus of random definitions. The probable explanation for the results is that it is possible to interpret the expression literally when we present them in isolation and the answers reflect that. The answers with top grades often mention that the expression can be used idiomatically or metaphorically and connects the false definition to a literal interpretation.

In the few examples where figurative explanations also get the highest grade, the figurative explanation resembles the correct definition to a high degree, for instance by being somewhat broader in such a manner that the figurative false definition could also cover the correct use. Considering that these false definitions were the most difficult to write, we will use the results as feedback and

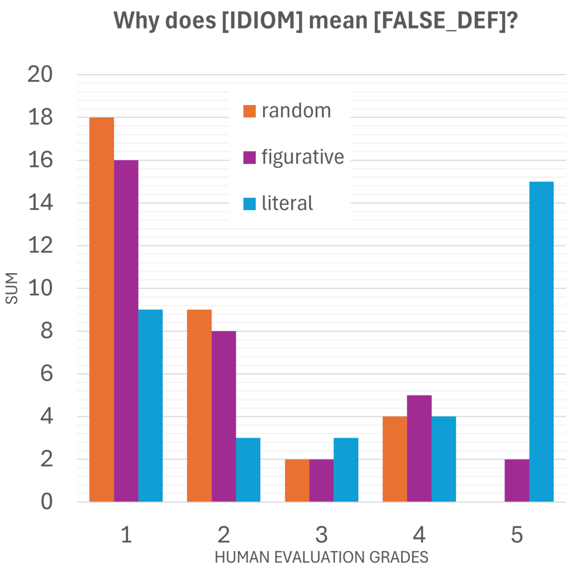

Figure 4: Human evaluation of the same subset as experiment 2 with a prompt that implies that the false definition is true.

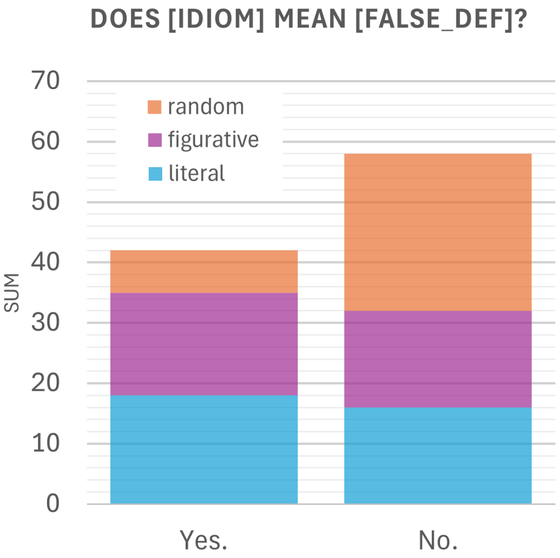

Figure 5: Results of asking ChatGPT-4o-mini directly whether a idiomatic expression can mean a false definition.

rewrite these definitions in the next version of the dataset.

We also experimented with another type of prompt to investigate whether the model would respond differently if we did not imply that the false definition was true in the prompt. Moreover, we wanted to test a prompt that did not require manual annotation. The result is the prompt: "*Does [IDIOM] mean [FALSE_DEFINITION] (yes/no)?*".

In figure 5, we see that ChatGPT 4o-mini correctly answers "no" in 58 of the cases. Considering that the same model could to some extent explain the meaning of 73 expressions in experiment 2, there is still room for improvement. In particular, there is a 29% discrepancy between the two prompts. For the "no" category, the discrepancy is caused by the random definitions which are predominately identified as false. This suggests that the model is capable of correctly identifying the very wrong (e.g. random definition) information, but is misled by false information when it's presented as correct. In the "yes" category, we see a large number of literal cases that got a high grade with the previous prompt, which is not surprising considering that this type of false definition is not necessarily wrong, but the expression is not often used with that meaning. However, we also see a similar number of literal cases in the "no" category, and a portion of these also belong to the previously correct cases (grade 5). These inconsistencies may be relevant to further exploration in the future, for instance by running the experiment on all the idiomatic expressions with each of the false definitions to see if we can find a pattern across more examples.

Conclusion

We have presented a new dataset of 1000 Danish idiomatic expressions from the Danish Dictionary DDO that includes the correct dictionary definition as well as three false definitions, namely a literal misinterpretation, a figurative misinterpretation and a random definition. The purpose of the creation of the dataset is to be able to evaluate Danish language proficiency of LLMs in one of the most challenging areas of language understanding. The dataset was more difficult to compile than anticipated; the figurative false definitions were particularly difficult to formulate. We have furthermore demonstrated three ways of using the dataset for evaluation: (1) as a benchmark dataset with multiple choice, (2) in a generative task, (3) to investigate hallucinations. The first experiment showed that the performance is influenced by the terminology used in the prompt. The second experiment supported the finding that cultural specific metaphors are challenging for LLMs, while also highlighting a problem with some of the false definitions that are broad enough to technically also cover the correct meaning. Lastly, the third experiment showed that ChatGPT struggles to correct false information provided in the prompt.

Aknowledgement

We would like to thank Ida Flörke for her contribution to the creation of the dataset.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1318–1326.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only. Association for Computational Linguistics.

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

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Bolette Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating llm-generated explanations of metaphors – a culture-sensitive study of danish. In *Proceedings of The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*. In print.

Bolette Pedersen, Nathalie Sørensen, Sussi Olsen, Sanni Nimb, and Simon Gray. 2024. Towards a Danish semantic reasoning benchmark - compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16353–16363, Torino, Italia. ELRA and ICCL.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.