DAMETA: An LLM Benchmark for Danish Metaphor Interpretation with Systematically Varied Distractors

Nina Skovgaard Schneidermann[1], **Sanni Nimb**[2], **Nathalie Carmen Hau Norman**[1], **Sussi Olsen**[1], **Bolette S. Pedersen**[1]
Centre for Language Technology, NorS, University of Copenhagen[1],
The Society for Danish Language and Literature[2]
Njalsgade 76, 2300 Copenhagen S[1], Christians Brygge 1, 1219 Copenhagen K[2]
{ninasc, naha, saolsen, bspedersen}@hum.dk.dk, sn@dsl.dk

Abstract

We present DAMETA, the first evaluation benchmark for Danish metaphor interpretation in language models, derived from the following sources: an annotated corpus (the Dafig Corpus), the Danish dictionary (DDO) and culture reviews in Danish newspapers. Each of the 900 data instances contains a sentence with a metaphorical target word and four human-created paraphrase options; one correct interpretation and three systematic errors or *distractors*: i) a false literal paraphrase (typically concrete), ii) a false figurative paraphrase (typically abstract), and iii) a false contradictory paraphrase. The benchmark is tested on seven language models, and 5% of the data is further tested on humans for comparison. Results show, among others, that when informed in the prompt that the target word is a metaphor, the models tend to be most distracted by the false figurative paraphrase; in contrast, when uninformed about the metaphorical setting, the models are more distracted by the false literal paraphrase. The dataset goes beyond standard by incorporating descriptive metadata regarding metaphor conventionality on a 3-graded scale (lexicalised, implicit, and ad-hoc), alongside a range of dictionary-derived source domains (military, gastronomy, health, meteorology, etc.). These metadata enable deeper analysis and potentially innovative insights of model performance regarding creativity, language change, and culture-sensitivity.

Keywords: metaphors, benchmarks, LLMs, Danish

1. Introduction

Robust evaluation and benchmarking of figurative language is a critical challenge for modern Natural Language Understanding and large language models (LLMs), particularly when it comes to the linguistic and cultural diversity, as well as generalisation in low and medium-resourced languages such as Danish.

This is because figurative language, and in particular metaphoric language, is more semantically complex than literal language, since it relies on humans' ability to simultaneously manage references to both a concrete (or literal) and a transferred, often more abstract level of meaning (Lakoff and Johnson, 1980b). For instance when we use the word *appetiser* from the gastronomy domain to describe an inspiring excerpt from a book or a movie, which should raise people's interest in enjoying the entire work. Overall, metaphors are seen as offering cognitive benefits in communication by making the content more engaging and personal (Citron and Goldberg, 2014; Noveck et al., 2001) and thereby functioning as an advanced communication tool for humans to express themselves in inspiring, engaged, and colourful ways. As a consequence of their somewhat personalised flavour, metaphors often rely on values and features specific to a certain culture and language community, making them less transferrable across languages than other figures of speech. An example of a culture-specific metaphor is the Danish word *rugbrødsarbejde* ('rye-bread work'), which describes inconspicuous but nevertheless necessary routine work - a metaphor rooted in the unpretentious but healthy ryebread that forms the nutritional backbone of a typical Danish diet.

Due to unbalanced training data, however, most large language models (LLMs) are somewhat culture-insensitive with respect to metaphors in low- and medium-sized languages. While many metaphors rely on universal human properties (Lakoff and Johnson, 1980a), a significant percentage proves to be heavily coloured by culture, norms and values of specific language communities - particularly when compared to English, the dominant language in LLM training data. This phenomenon is studied in Pedersen et al. (2025), which shows that metaphors unique to Danish culture perform remarkably worse in multilingual language models than those found across languages and cultures. An important step towards gearing future LLMs to encompass linguistic and cultural diversity of small- and medium-resourced languages is therefore to provide high-quality, human-curated interpretation benchmarks for these languages - including evaluation of complex figures of speech such as metaphors. Such benchmarks enable us

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11885–11895
11–16 May 2026.

to study strengths and weaknesses of particular models and architectures in more detail, as well as to direct and follow progress of the development of more culture-aware models.

With continuous advances in model size and capability, evaluation methods require greater sophistication which move beyond performance metrics and probe for genuine semantic reasoning and comprehension. Tedeschi et al. (2023) argue that claims of "superhuman performance" in understanding are premature, as many large-scale benchmarks predominantly rely on simple classification tasks and rapid evaluation cycles that fail to take into account the nuances of language understanding. These limitations become particularly acute in figurative language processing (FLP) research, which, in spite of its rapid increase in performance on metaphor detection tasks, still necessitates distinction between whether models actually interpret, or merely memorise, figures of speech.

Our current work seeks to address the challenges of i) constructing non-English monolingual resources and ii) moving beyond surface-level detection. Thus, in this paper, we introduce DAMETA, a novel benchmark for Danish metaphor interpretation which tests nuanced understanding through systematically designed multiple choice questions with varying distractors targeting different LLM failure modes.

The paper is organised as follows: Section 2 positions our work in the field of metaphor understanding and evaluation. Section 3 accounts for the data and methodology used to compile and annotate the dataset. In Section 4, we evaluate the dataset against a number of both proprietary and open models, as well as against a number of human non-experts, and discuss the results. Section 5 concludes and paves the way for future work.

2. Related Work

Recent work on metaphor understanding relevant to our methodology and purpose span three key areas: metaphor Interpretation benchmarks, multilingual evaluation, and Danish-specific resources.

Several benchmarks have explored metaphor understanding through paraphrase generation and binary choice tasks. Tong et al. (2024) introduce MUNCH, a dataset of over 10,000 paraphrases evaluating whether models can judge and generate appropriate literal interpretations of metaphorical expressions. The dataset includes 1500 inapt paraphrases that use literal, source domain interpretations. Across tested models, performance fell below random baseline for most tasks, with a crucial tendency to confuse source and target domains—frequently accepting multiple inapt paraphrases as correct when they should distinguish between literal and metaphorical meanings. Similarly, Liu et al. (2022) propose Fig-QA, employing a Winograd-style approach requiring figurative interpretation to resolve ambiguous references. Testing both auto-regressive and masked language models, they find above-chance zero-shot performance that nonetheless falls substantially short of human performance (26 percentage point gap), with considerable variation by model size. More recent work has also emphasised the role of common sense reasoning and inference in metaphor comprehension. For example, Comșa et al. (2022) introduce MiQA, which combines metaphor interpretation with common sense inference through a multiple-choice question-answering format, requiring models to draw on world knowledge to select correct interpretations. Crucially, although not directly related to FLP, Mozafari et al. (2025) demonstrate the value of using *plausible distractors* - incorrect options that are semantically related and superficially reasonable - to create more discriminative evaluation scenarios that better distinguish genuine understanding from pattern matching.

However, Sanchez-Bayona and Agerri (2025) demonstrate that these apparent capabilities may be misleading. Through comprehensive evaluation across multiple datasets - including MUNCH and Fig-QA - they find that LLM performance is more strongly influenced by surface-level features such as sentence length and lexical overlap than by metaphorical content itself. This suggests that any alleged emergent abilities to understand metaphorical language result from a combination of shallow pattern matching, in-context learning, and linguistic knowledge rather than genuine figurative comprehension.

These challenges are likely compounded in multilingual contexts. According to Zhang et al. (2023), models exhibit what has been termed "subordinate multilingualism", struggling with translation-varying tasks where English is not a reliable intermediary. Along the same lines, Cao et al. (2023) highlight fundamental issues with cultural alignment in multilingual models, showing that models often impose cultural frames in the source language (typically English) on the interpretations in the target language. Given that metaphors are deeply rooted in cultural conceptual systems, this suggests that the limitations observed in English benchmarks may manifest differently, and potentially more severely, in other languages: Kabra et al. (2023) create a figurative language inference dataset, MABL, for seven diverse (non-Indo-European) languages , revealing significant performance deficiencies in multilingual/multicultural contexts compared to English. Another prominent multilingual metaphor inference dataset is Meta4XNLI, Sanchez-Bayona and Agerri (2024), a cross-lingual parallel corpus for metaphor

detection and interpretation in English and Spanish, on which the authors observe that best results on multilingual pre-trained language models are obtained when the corpus in both languages is used for training. These findings demonstrate the value of cross/multilingual training data spanning over broader cultural contexts.

Systematic evaluation of metaphor understanding in non-English languages remains scarce, relatively speaking. Specifically for Danish, several lexical-semantic benchmarks have been developed in recent years and made available through the ScandEval platform (Nielsen, 2023) (now integrated into the Euroeval platform[1]) and The Scandinavian Embedding Benchmark (Enevoldsen et al., 2024). Particularly relevant to our work is the Danish Reasoning Benchmark (Pedersen et al., 2024), which provides a collection of six datasets derived from semantic dictionaries, including a wordnet, FrameNet lexicon, sentiment lexicon, and a thesaurus, establishing important infrastructure for evaluating semantic understanding. The Danish Idiom dataset (Sørensen et al., 2025) is the most methodologically adjacent dataset to DAMETA, employing a similar set of conditions with correct paraphrases and multiple distractors to allow multiple choice evaluation scenarios. In contrast, Pedersen et al. (2025) use a manually curated set of metaphors divided by culture-specificity for qualitative assessment of LLM-generated interpretations, providing complementary insights into model behavior on culturally-grounded figurative language.

To our knowledge, our dataset is the first Danish manually annotated single-word metaphor interpretation benchmark. With our work, we adapt established English metaphor interpretation methodologies to Danish, leveraging the idea of systematic distractor generation to test semantic understanding, and reveal crucial empirical results for Danish metaphor interpretation through a comprehensive analysis on metadata such as metaphor conventionality and source domain information.

3. Data and Methodology

DAMETA consists of 900 Danish metaphor interpretation instances, each containing a sentence with a metaphorical target word and four paraphrase options (one correct interpretation and three systematic distractors, see Section 3.3). All data are made freely available from github under password protection [2].

In order to capture a wide spectrum of metaphors in Danish, our dataset is constructed from multiple complementary sources; the metaphor portion of the DaFig corpus[3], the Danish Dictionary (Det Danske Sprog- og Litteraturselskab, 2024) (henceforth DDO) and a range of ad-hoc metaphors found manually in culture reviews from newspapers.

The following subsections detail the selection of source material, our metaphor typology based on lexicographic documentation, the systematic distractor design framework, the annotation procedure, and finally the included metadata.

3.1. Source Material Selection

In order to ensure coverage of a wide variety of metaphors, namely with respect to conventionality, source domains, and cultural specificity, we construct our metaphor interpretation dataset by combining existing lexicographic resources with manually annotated metaphors in context, leading to a hybrid top-down/bottom-up approach. As such, the final dataset comprises metaphors from three distinct sources, balanced to capture both conventional and emergent usage: 62% are extracted from the DDO, enabling fine-grained identification of well-documented, conventional metaphors; 30% consist of example sentences from DaFig, a manually annotated corpus of naturally occurring figures of speech; and the remaining 8% are drawn from recent Danish newspaper culture reviews to capture novel metaphors not yet documented in formal dictionaries.

The DDO portion of the metaphors were pulled from semantically related groups of lemmas across specific source domains; e.g. gastronomy, construction, military, and meteorology. As a constraint, we chose to include lemmas with only two senses identified in the dictionary manuscript, a main sense and a subsense describing a metaphorical use. Overall, DDO provides over 1600 corpus citations illustrating figurative sub-senses, plus additional citations for low-frequency metaphors that we included as a supplementary set of "ad-hoc" examples (see Section 3.2 on typology for clarification).

The DaFig portion of the dataset is derived from the Danish Figurative Language Corpus (DaFig), a manually annotated collection of approximately 40,000 words from Danish newspaper articles. This corpus was created to address the lack of non-English and multi-figurative resources by annotating metaphor, hyperbole, and verbal irony within the same texts. The annotation process followed established linguistic procedures to ensure rigor and consistency with similar corpora in other languages: For metaphors, we used the metaphor identification procedure, MIP (Group, 2007) and its extension, MIPVU (Steen et al., 2019), the core

[1] https://euroeval.com/

[2] https://github.com/kuhumcst/danish-semantic-reasoning-benchmark/tree/main/metaphor_benchmark

[3] https://github.com/NiSc91/DaFig

of which involves identifying whether a given word in context contrasts with a more basic sense of the word, which is defined as the more concrete, specific, and not necessarily most frequent, sense. We adapted the scope of the annotations such that only nouns, verbs, adjectives, and adverbs were included, and we further excluded metaphor signals (e.g., 'like,' 'as,' and other explicit markers of comparison) and so-called implicit metaphors - instances where anaphoric expressions (e.g., 'it,' 'this,' 'there') refer back to a metaphorically used word earlier in the discourse (Steen et al., 2019). We also made some adaptations recommended by Nacey et al. (2019) in adapting the procedure to Danish, particularly with respect to the treatment of phrasal verbs and multi-word expressions. For the purposes of the DAMETA dataset, the original token-level metaphor annotations from DaFig were converted to sentence-level classifications, resulting in a pool of approximately 1,500 sentences containing validated, naturally occurring metaphors in news text.

As such, this component of DAMETA is intended to serve as a supplement to the top-down, lexicographic metaphor identification approach, comprising thereby altogether a wider variety of metaphors (see section 3.2 for more details on our metaphor typology).

The fact that we utilise the MIPVU procedure allows us to consider a word's "contemporary basic sense" more broadly, enabling the identification of metaphorical uses even when a figurative sense is not explicitly listed in the dictionary, or when the word itself is absent.

The final 8% of the data was manually collected for the purpose of extending the amount of novel (ad-hoc) metaphors specifically for DAMETA. This hybrid source selection approach ensures that the benchmark is both lexically informed and robust enough to evaluate nuanced, real-world language understanding.

3.2. Metaphor Typology as Metadata

Our dataset encompasses three types of metaphorical expressions, distinguished by their degree of conventionalisation and lexicographic representation. It should be noted that a majority of our dataset consists of type 1 metaphors (N=721; 80%), with smaller proportions of type 2 (N=15; 1.6%) and type 3 (N=164; 18.2%) instances, reflecting the natural distribution of conventional vs creative metaphorical language use. Still, we include this metadata in order to allow for more granular analysis of model performance. The types are defined as follows:

Type 1: Explicit metaphorical meanings. These are lexicalised metaphors with dedicated dictionary entries, either as the primary sense or as a subsense of a lemma. For example, *kogebog* ('cookbook') has a metaphorical subsense meaning 'manual' or 'guide for doing something', while *glasloft* ('glass ceiling') exists primarily with its metaphorical meaning referring to an (often unacknowledged) barrier of norms and structure, though still clearly derived from a more basic meaning in accordance with the MIPVU definition. Other examples include *efterspil* (literally 'after-play', meaning aftermath/repercussions) and *bulldozer* (referring to a person who recklessly disregards others). Notably, a small subset of Type 1 metaphors have become so conventionalised that the metaphorical meaning functions as the primary or most frequent sense, effectively overshadowing the original literal meaning—such as *følger* ('follower' on social media).

Type 2: Implicit metaphorical meanings. These metaphors lack explicit dictionary entries but can be inferred from existing senses through systematic semantic extensions or productive patterns documented in contextual examples. For instance, the property *driftsikker* normally refers to reliability in technical operations but can be used metaphorically about a person you can always count on. Similarly, *værdifuld* ('valuable') is clearly metaphorical when applied to the value of time, but not when applied to the value of money. These are typically less conventionalised than Type 1 but still rooted in documented usage patterns.

Type 3: Novel or ad hoc metaphors. These are creative, context-dependent metaphorical uses that are not (yet) conventionalised enough to be lexicalised. They are primarily sourced from culture reviews in newspapers, from the DaFig corpus and supplemented by a DDO list collected from corpus examples of concrete dictionary senses that were marked as low-frequency figurative examples in the first edition and have not become conventionalised 30 years later. These instances represent the productive edge of metaphorical language use. An example from DaFig: *gennemtæve* ('beat heavily') referring to a sports situation where someone loses badly in a football match. These are designed to test a model's ability to generalise and interpret creative language.

3.3. Distractor Design Framework

We developed three systematic distractor types targeting specific interpretation failure modes. This framework is designed to move beyond simple accuracy and diagnose the reasoning errors models make when confronted with figurative language. The three distractor types are:

1. **Literal (mostly concrete) distractor:** This option presents a literal (mostly concrete) interpretation of the metaphorical target word

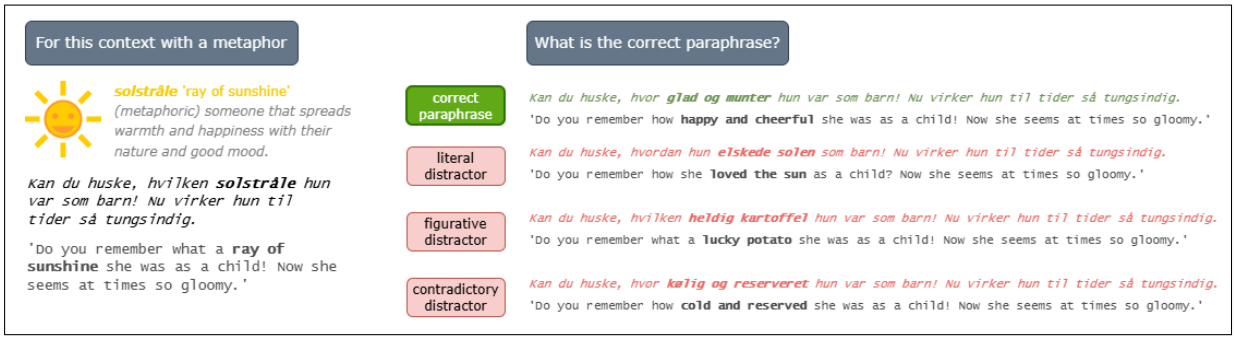

Figure 1: Examples and paraphrases for the Danish metaphor *solstråle* 'ray of sunshine'.

relating to the source domain, often resulting in a logical non-sequitur within the sentence's context. It directly tests a model's ability to recognise that a non-literal interpretation is required.

2. **Figurative (mostly abstract) distractor:** This option provides a plausible but semantically imprecise interpretation (mostly abstract) that resides in the correct target domain but misses the specific nuance of the metaphor. This "near-miss" distractor tests a model's more fine-grained semantic understanding and ability to select the most precise meaning, as opposed to a related one.

3. **Contradictory/antonymic distractor:** This option offers a meaning that is antonymic or logically opposite to the correct metaphorical interpretation. It tests the model's basic coherence checking and contextual consistency.

Figure 1 gives an example from the dataset for the metaphor *solstråle* (concrete: 'ray of sunshine'; metaphorical: 'someone that spreads warmth and happiness with their nature and good mood') with the corpus example (sometimes slightly modified or shortened) and the three systematic distractors.

3.4. Annotation Process

The annotator group consisted of five highly experienced lexicographers and computational linguists. The annotators followed a structured protocol to ensure consistency in paraphrase generation across the dataset. The process for each instance began with a thorough analysis of the sentence context to identify the metaphorical expression. Annotators then consulted the DDO to classify the metaphor according to our typology (Type 1-3).

Following this classification, annotators generated the correct paraphrase and the three distractors according to the framework. Key guidelines emphasised that paraphrases must be brief, precise, and literal wherever possible. The process was informed by Pustejovsky's qualia structure to ensure that the core entailments of the metaphor were captured.

Our validation process was iterative: Initially, a subset of 20 instances was co-annotated by at least two researchers to establish consensus and refine the guidelines. Due to the high cognitive load and complexity of the task, with annotation speeds averaging 5–10 instances per hour, the remaining dataset was annotated by individual experts. To maintain high data integrity, this was supplemented by regular peer-sparring sessions and a final curation phase to ensure lexicographic grounding against DDO definitions.

Overall, the annotators reported that the figurative distractor was the most time consuming one to generate, because providing a plausible but still imprecise paraphrase required some thinking and modification to the original corpus example. Even though the literal distractors were more easily generated, they often ended up being a little peculiar and not very natural. The annotators also reported that working from a list where the metaphors were grouped according to their source domain generally eased and sped up the annotation process , although this was only possible for the DDO examples, and not for type 2 and 3 metaphors.

3.5. Metadata Annotation

Each metaphor instance was enriched with metadata capturing linguistic and conceptual properties, allowing for detailed error analysis and a deeper understanding of model capabilities. The metadata for each instance includes:

- **Conventionality level:** The assigned metaphor type (1, 2, or 3).
- **Source domain:** A classification of the metaphor's source domain, only available from the DDO-portion of DAMETA (e.g., military, gastronomy, health).
- **DDO reference:** A link to the relevant DDO entry, where applicable.

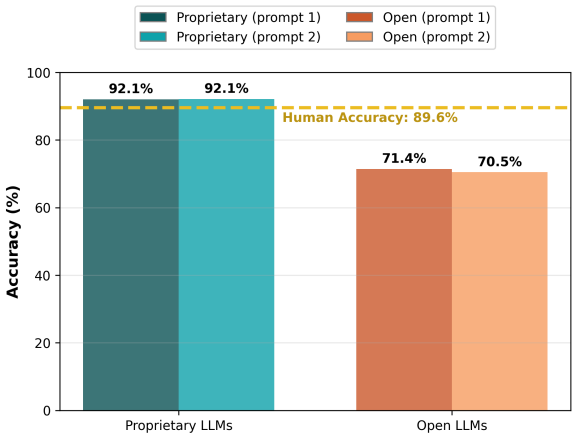

Figure 2: Overall performance across model types (proprietary vs open LLMs) and humans.

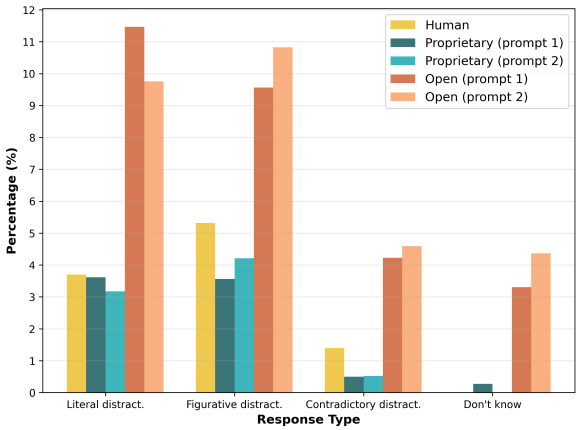

Figure 3: Overall performance across distractor response types.

- **Context summary:** An optional, brief summary for instances with particularly long or complex sentence contexts. The shortened context summary, if applicable, was replaced with the original sentence in the dataset and used in the model evaluation.
- **Uniqueness:** 75 of the metaphors currently have information on uniqueness, more specifically, on whether the same metaphor exists in English.

4. Evaluation

4.1. Experimental setup

We evaluated seven LLMs on our dataset: five open-source models run locally via Ollama (Llama 3.1; 8B, Gemma 2; 9B, Mistral; 7B, Qwen 2.5; 7B, and Phi-4; 14B), and two proprietary models accessed via OpenRouter (GPT-4o-mini and Claude 3.5 Sonnet). All models were run with default temperature settings. The range of models were selected to represent a broad spectrum of architectural approaches and resource constraints.

The evaluation was conducted under two distinct prompt conditions: an *uninformed* condition (Prompt 1), and an *informed* condition (Prompt 2). The prompts were presented in Danish, and in both versions the models were presented with a specific metaphorical lemma and the original sentence. In the uninformed condition (Prompt 1), we asked "What does the word *lemma* mean in this context?" followed by the sentence and four paraphrase options plus "Don't know." In the informed condition (Prompt 2), the prompt explicitly stated "In the following sentence, the word *lemma* is used metaphorically" before asking which interpretation best describes the metaphorical meaning. Both prompts included identical penalty scoring instructions (+1 for correct, -1 for incorrect, 0 for "don't know") to discourage guessing. This dual setup allows us to investigate whether priming a model with metalinguistic knowledge affects its performance.

In addition to the model evaluations, a subset of 48 items (5.3 % of the benchmark) was evaluated by nine human participants recruited via convenience sampling (family, friends, and colleagues with varying linguistic backgrounds). The annotators received the same scrambled multiple-choice format as prompt 1 (uninformed condition), but without penalty scoring options. However, due to the informal recruitment process, we cannot control if they were aware that this was metaphor research, making this a preliminary baseline rather than a controlled comparison.

4.2. Results, Analysis, and Discussion

Overall, models are shown to perform fairly well on the benchmark, with accuracies ranging from approximately 60% for the poorest performing model (Mistral) to around 94% for the best performing model (Claude-3.5-sonnet (see Figure 2 and Table 1). However, performance tended to stratify sharply depending on the model type: proprietary models (GPT-4o-mini: 90%, Claude 3.5 Sonnet: 94%) substantially outperformed the general class of open-source alternatives by a margin of 20 percentage points. This means that the human baseline of 89.6% is surpassed by the average proprietary model by 2 percentage points, while the class of open-source models perform under the human baseline by 18 percentage points. This quite substantial difference may partially be explained by architectural advantages such as model size as well as possible training data contamination; proprietary models may have web-crawled online DDO dictionary content and learned patterns based on definitions and examples.

Model	Prompt 1		Prompt 2	
	Acc. (%)	DK (%)	Acc. (%)	DK (%)
Human	89.58	-	-	-
Claude 3.5	**93.9**	0.3	**93.8**	0.0
GPT-4o-mini	90.3	0.2	90.6	0.0
Gemma-2 9B	83.5	0.3	83.4	0.1
Phi-4 14B	82.7	0.0	82.5	0.0
Qwen-2.5 7B	66.2	2.2	70.9	0.9
Llama-3.1 8B	64.9	7.3	57.6	19.6
Mistral 7B	60.0	6.6	59.3	0.8

Table 1: Model performance comparison across all models: Accuracy and don't-know rates (DK)

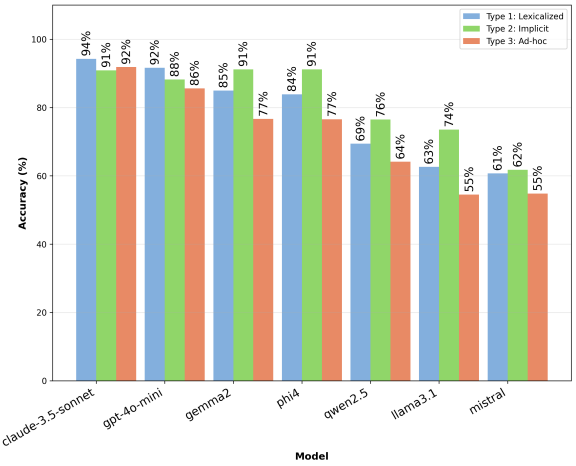

Figure 4: Performance analysis by models and types of metaphor (lexicalised, implicit or ad-hoc).

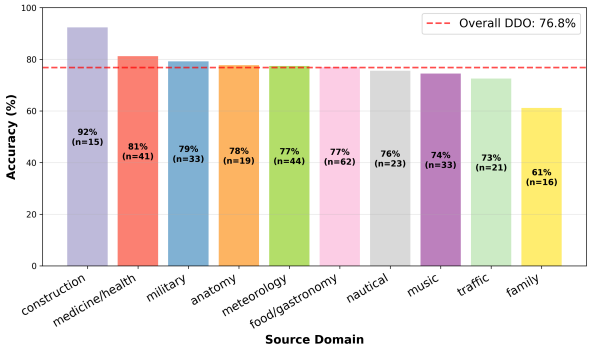

Figure 5: Performance breakdown by top-10 source domains: construction work, gastronomy, medicine/health, anatomy, meteorology, military, music, and family.

In spite of these general tendencies, table 1 shows a quite noticeable variation within open-source models of more than 20 percentage points between Mistral (7B) (60%) and Gemma-2 (9B) (83.5%). While this performance generally scales with model size, the results indicate that architectural refinements and training data quality may modulate this relationship. For instance, Gemma-2 (9B) substantially outperformed Llama-3.1 (8B) by almost 20 percentage points despite a comparable parameter count. Furthermore, Gemma-2 (9B) approached near-human performance, suggesting that high-quality reasoning-focused training sets can enable mid-sized local models to narrow the gap with much larger proprietary systems.

It is a point worth noting that a human baseline of 89.6%, i.e. a 4 percentage point gap from the best-performing proprietary model, is somewhat surprising for a task involving native-language metaphor comprehension. It suggests that the paraphrase selection task is non-trivial even for humans, particularly for less frequent metaphorical senses, and that younger participants (20s-30s) occasionally failed to recognize older or register-specific metaphors that appeared frequently in model training data. Given that humans were only tested on type 1 and type 2 metaphors, without a way to systematize whether or not humans possessed knowledge of the nature of the metaphor research, a more elaborate comparison would be premature. However, the human baselines mirror that annotators found the data creation itself to take a significant amount of effort. While we argue that high-quality data do require a manual effort in order to achieve a gold standard, it is worth investigating the benefits of supplementary automated methods of annotation or data augmentation.

With respect to prompt framing effects, as shown in figure 2, we observe that explicit metaphor framing (prompt 2) produced only small accuracy gains or losses for most of the models (below 1 percentage point). Only Qwen-2.5 (7B) improved substantially with 4.5 percentage points, when informed that the target was metaphorical. Llama-3.1 demonstrated significant instability under explicit metaphor framing (Prompt V2), with accuracy dropping from 64.9% to 57.6%, while the 'Don't Know' rate more than doubled from 7.3% to 19.6%. The divergent behaviour of the smaller models with respect to this phenomenon is intriguing, and we would like to investigate the reasons behind this more closely.

Furthermore, we observe that prompt framing significantly shifted error patterns across all models. When explicitly told that the word was metaphorical, models were substantially more likely to select the false figurative distractor over other error types (see Figure 3). This mirrors human behaviour and aligns with our expectations, in that the correct paraphrase and the false figurative distractor are semantically related, and explicit metaphor framing primes figurative interpretation.

Performance across metaphor types (lexicalised, implicit, and ad-hoc) reveals interesting hierarchical trends (see Figure 4): Ad-hoc metaphors (Type 3) seem to represent the most significant

Model	Context with metaphor	Correct paraphrase	Selected distractor	
Mistral	*Okay, jeg napper en **morfar** på sofaen i ny og næ. Med snorkelyde.*	*Okay, jeg napper en **lur** på sofaen i ny og næ. Med snorkelyde.*	*Okay, jeg lader som om **jeg er en morfar** når jeg ligger på sofaen i ny og næ. Med snorkelyde.*	Literal
	'Okay, I will take a **grandfather** (meaning a nap) on the couch now and then. With snoring noise'	'Okay, I will take a **nap** on the couch now and then. With snoring noise.'	'Okay, I will pretend **to be a grandfather** when I am lying on the couch now and then. With snoring noise.'	
GPT4o-mini	*Danmark er en lille **duksedreng** i det internationale selskab, når det gælder udviklingshjælp til verdens fattige lande.*	*Danmark er den der **altid overholder alle regler og principper** i det internationale selskab, når det gælder udviklingshjælp til verdens fattige lande.*	*Danmark er den der **altid har lavet lektierne** i det internationale selskab, når det gælder udviklingshjælp til verdens fattige lande.*	Figurative
	'Denmark is the **puppet/paragon** in international society when it comes to development aid to the world's poorest countries.'	'Denmark is the one **always following all rules and principles** in international society when it comes to development aid to the world's poorest countries.'	'Denmark is the one that has **always done their homework** in international society when it comes to development aid to the world's poorest countries'	

Table 2: Examples of incorrect choices

challenge across all tested architectures, with an accuracy drop by 2 to 8 percentage points compared to Lexicalized metaphors (Type 1), even in high-performing proprietary models. This matches our expectations and suggests that while LLMs excel at retrieving pre-learned metaphorical mappings during training, certain novel, context-dependent reasoning required for less conventionalized metaphors represents more of a struggle. Interestingly, several models—most notably Phi-4, Qwen-2.5, and Llama-3.1 showed a marginal performance uptick on Implicit metaphors (Type 2). However, this finding is modulated by two critical caveats: Firstly, the current sample size for type 2-examples is very small, representing merely 1.6% of the dataset, making any attempted interpretation cautious. Secondly, the high performance in this category may be explained by conventionalization: These metaphors may be so deeply integrated into common usage that they may function as dominant polysemous senses rather than active metaphorical mappings, thus appearing with high frequency in the model's training process. Future iterations of this benchmark will focus on expanding the Type 2 and Type 3 datasets to further decouple reasoning capabilities from frequency effects.

Irrespective of metaphor type, a brief glance into a few of the actual errors performed by the models further sheds light on particularly challenging cases. We have included two examples in table 2. In the first example (*morfar* literal: 'grandfather', figurative: 'a nap'), shows that the distractor chosen actually repeats the target word in a literal context ('pretending to be a grandfather'), making the word similarity with the original example obvious and resulting in a high lexical overlap. In the second example, the figurative distractor chosen may be explained by the fact that *lektier* ('homework') is semantically related to *duksedreng* in the sense that both words relate to the school domain in their literal meaning. The models also appears to be fairly unfamiliar with the term *duksedreng* itself.

With respect to the metadata on source domains, we observe that models showed a modest but consistent variation in accuracy (Figure 5). Notably, gastronomy-sourced metaphors yielded relatively high accuracy - an initially surprising find given the culture-specificity of this domain. We initially presumed this might challenge LLMs less attuned to Danish cultural concepts; e.g. *rugbrødsarbejde* ('ryebread work' = tedious routine work); *højtbelagt* ('highly coated', referencing Denmark's elaborate open-faced sandwiches). However, closer analysis revealed that most gastronomy metaphors in our dataset draw on universal cooking/eating scenarios rather than Denmark-specific foods: *krydre* ('to spice up'), *bismag* ('aftertaste'), *kogebog/opskrift* ('cookbook/recipe), *være sulten efter* ('be hungry for'). This universality likely explains the unexpectedly strong performance.

In contrast, metaphors derived from the family- and marriage-related source domain, such as *fornuftsægteskab* ('marriage of convenience'), *skilsmisse* ('divorce'), and *frieri* ('courtship') when used in contexts like political collaboration or crises, proved harder for the models to predict, with an accuracy of 64%. This is despite these metaphors appearing to be cross-culturally applicable. Note, however, that this part of the dataset is relatively small, and that idiosyncratic, Danish-specific metaphors like the previously mentioned *morfar* (lit. 'grandfather', fig. 'nap') disturb the general picture.

5. Conclusions and Future Work

We have presented DAMETA, a metaphor benchmark compiled to advance Danish NLP by assessing – and potentially improving – how well LLMs interpret figurative language in Danish. The dataset comprises 900 metaphor instances, each consisting of a Danish metaphor in context accompanied

by four human-created paraphrases: a correct paraphrase and three systematically incorrect distractors. The dataset is not only available at github but is incorporated in the EuroEval benchmark, too.

Evaluation results demonstrate that proprietary models perform remarkably well, even exceeding the accuracy of our nine human informants. Notably, we observe that medium-sized models (e.g., Phi-4 and Gemma-2) also achieve high performance, approaching human-level accuracy. This suggests that the capacity to interpret Danish metaphors is present even in more resource-constrained architectures, indicating that Danish metaphor interpretation may be more accessible to modern LLMs than previously assumed. When models produce errors, they are mostly confused by the figurative distractor when explicitly informed that the target word is a metaphor (Prompt 2), a pattern that coincides with human informant performance pattern. Conversely, when models are *not* informed about the nature of the target word (Prompt 1), the literal distractor is most often chosen.

Beyond the paraphrases, the dataset provides deeper categorisation through metadata regarding *metaphor conventionality* and *source domain* (currently available for lexicalised - type 1 - metaphors only). These metadata enable deeper analysis of model performance regarding creativity, language change, and culture-sensitivity, and open avenues for interesting studies of how metaphors develop over time and relate to culture-specific aspects - potentially informing more dynamic and culture-aware LLM development.

In future work, we plan to i) systematically assess metaphor *uniqueness*, i.e. whether metaphors are specific to Danish language and culture or shared with other languages like English, and add this as supplementary metadata, and ii) investigate more prompt conditions to further inspect LLM capability on our dataset. Uniqueness information currently exists for only approx. 75 instances of the dataset, but previous studies (Pedersen et al., 2025) indicate that this significantly affects model performance, suggesting that multilingual models require additional refinement to become sufficiently culture-aware for medium-resourced languages like Danish. Benchmarking such gaps is a first logical step toward improving models through extended training data, fine-tuning, or post-processing. With respect to prompt engineering, we plan to run our dataset with more systematic prompts that explore more distinct conditions - particularly whether model performance decreases if presented with a ranking task, or whether models will select multiple distractors as correct when not constrained by a single-choice requirement. More sophisticated prompt engineering may be another crucial step in evaluating the semantic capability of LLMs beyond surface-level knowledge.

Further research is planned within a newly granted project from The Independent Research Fund Denmark, starting in 2026. This project will investigate how LLMs process metaphors in their internal layers to determine whether models capture metaphors' double representation as humans do. Building on an extended version of the ChainNet resource (Maudslay et al., 2024), which describes connections between source and target domains, we will explore if and how this double reference is represented and compare with human metaphor processing through eye-tracking experiments.

6. Acknowledgements

Thank you to the Carlsberg Foundation for funding this work through The Benchmark Project: https://cst.ku.dk/english/projects/the-benchmark-project/, and to the informants performing the human evaluation.

7. Ethical Considerations and Limitations

The source data for DAMETA has been acquired with acceptance from and in collaboration with DDO and the relevant newspapers. The dataset has been compiled with respect to the cultural value it holds. All annotators were given the opportunity to flag problematic examples or cases of doubt through comments.

Regarding limitations, exposing the models to a larger variety of prompts would be fruitful and could yield different results. Several studies have shown that prompt formulation (and language) significantly impacts performance, yet our study examines only two Danish prompt conditions, both of which involve models being presented with both the metaphor word and the sentence as inputs. Furthermore, our prompt constraints the model to a single-choice response. It would be interesting and potentially enlightening to impose different constraints on the models - namely through a ranking task, in which they have to rank the paraphrases in order of plausibility, or a a task that includes a self-reported uncertainty metric.

We made a manual effort to include new ad-hoc metaphors in our dataset, not previously described in dictionaries. We hope that the 8% ad-hoc metaphors in our dataset reflect the actual metaphor distribution in natural language, even though we have not examined this.

Furthermore, one can question how well multiple-choice paraphrase selection, even with the inclusion of an abstract semantically related distractor,

truly captures metaphor interpretation. A more convincing — albeit far more complex — evaluation requiring human-in-the-loop assessment would be to examine how well a model engages with metaphors in coherent conversation and whether it can make adequate anaphoric references back to metaphors across sentences and paragraphs.

Finally, we are aware of the limitations of the human pilot study with respect to their informal recruitment through convenience sampling. Given that several of our informants were aware that we are conducting research on figurative language, it is plausible that some of them inferred that they were being asked to perform metaphor analysis, even if the prompt condition did not specify this. We also did not provide them with annotation guidelines which explicitly prohibited the use of dictionaries and other resources, which further presents a confounding variable. Thus, we cannot claim that humans were tested under the same conditions as the LLMs.

8. Bibliographical References

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.

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.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding.

Praglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

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. *Annual Meeting of the Association for Computational Linguistics*.

George Lakoff and Mark Johnson. 1980a. Conceptual Metaphor in Everyday Language. *The journal of Philosophy*, 77(8):453–486.

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

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the Ability of Language Models to Interpret Figurative Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

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.

Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025. Wrong Answers Can Also Be Useful: Plausibleqa -A Large-Scale QA Dataset with Answer Plausibility Scores. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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*.

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

Bolette S. 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 (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

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.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4XNLI: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation. (arXiv:2404.07053).

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and Large Language Models: When Surface Features Matter More than Deep Understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477. Association for Computational Linguistics.

Nathalie Hau Sørensen, Sanni Nimb, Agnes Aggergaard Mikkelsen, and Jonas Jensen. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63, Tallinn, Estonia. The University of Tartu Library.

Gerard Steen, Lettie Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Tryntje Pasma. 2019. MIPVU: A manual for identifying metaphor-related words. *Metaphor identification in multiple languages: MIPVU around the world*, 22:23.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the Meaning of Superhuman Performance in Today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor Understanding Challenge Dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536. 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.

9. Language Resource References

Det Danske Sprog- og Litteraturselskab. 2024. *Den Danske Ordbog*. (September 2024).