Of Words and Meaning: A Grammatical and Semantic Benchmark for Faroese LLM Understanding

Iben Nyholm Debess, Barbara Scalvini, Bolette Sandford Pedersen
University of the Faroe Islands, University of the Faroe Islands, University of Copenhagen
ibennd@setur.fo, barbaras@setur.fo, bspedersen@hum.ku.dk

Abstract

Evaluating language technology for low-resource languages faces a fundamental challenge: the scarcity of native benchmarks suitable for systematic assessment. For Faroese, no such evaluation frameworks exist. We address this gap by presenting the first benchmark suite for Faroese semantic understanding and grammatical competence. Our methodology transforms existing lexicographic resources, authoritative dictionaries and error corpora, into systematic evaluation tasks through computational restructuring, demonstrating a replicable approach for resource-constrained settings. The resulting benchmarks assess grammatical correctness, semantic relation classification, and metaphor comprehension. Evaluation across LLMs from compact open-source to large-scale commercial systems reveals consistent performance patterns favouring proprietary models. This work establishes a proof of concept for benchmark creation from traditional linguistic resources, and provides a methodological template for other low-resource language communities.

Keywords: low-resource evaluation, Faroese, lexicographical data, benchmark, leveraging heritage data

1. Introduction

The rapid advancement of large language models (LLMs) has demonstrated remarkable capabilities across diverse natural language understanding tasks, yet their performance on low-resource languages remains under-explored. This gap persists primarily due to the scarcity of evaluation benchmarks for such languages, as benchmark creation typically requires substantial manual annotation efforts and linguistic expertise that may be limited for smaller language communities. Existing evaluation frameworks predominantly focus on high-resource languages, leaving critical questions about LLM competence in low-resource contexts unanswered. Here, we present the first comprehensive grammatical and semantic benchmark suite for Faroese, a North Germanic language spoken in the Faroe Islands with approximately 70,000 speakers. We introduce novel datasets curated from dictionaries and a selection of other language resources that challenge models across two complementary dimensions of language understanding: grammar and semantics.

Our benchmark employs a novel methodology that combines traditional lexicographic resources with computational methods for task generation and distractor selection. By leveraging dictionary data as the source for semantic relations and metaphorical explanations, we create challenging, controlled evaluation tasks while mitigating the constraints of a low-data, low-expert setting. This approach of using carefully constructed prompts with randomized option ordering and dictionary-derived datasets creates a replicable pathway for other low-resource languages that possess lexical resources but lack language understanding benchmarks. To our knowledge, this is the first systematic curation and computational adaptation of Faroese lexicographic resources for NLP evaluation. This work contributes both methodological insights for evaluating LLMs on morphologically rich, low-resource languages and empirical findings on the extent to which current models can capture fine-grained semantic and grammatical distinctions in Faroese.

In this paper, we present three benchmark tasks for evaluating language models on Faroese: one focused on grammatical correctness and two on semantic understanding. The grammatical correctness task evaluates models' ability to identify violations of prescriptive grammatical rules. The first semantic task tests models' ability to distinguish between synonyms, antonyms, and unrelated words, evaluating their grasp of semantic polarity and lexical relations. The second semantic task assesses understanding of figurative language through metaphor explanation. We evaluate a range of proprietary and open-source LLMs of varying sizes, providing insights into how current systems handle linguistic distinctions in a morphologically rich, low-resource language.

2. Background and Related Work

2.1. Evaluation and Faroese

Evaluation plays a central role in the development and assessment of language models, providing standardized ways to measure model competence in various tasks within e.g. language, common sense, and reasoning. Large evaluation suites like GLUE and SUPERGLUE (Wang et al., 2018, 2020)

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4514–4526
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

have paved the way for evaluating language understanding in English, and have inspired the Nordic benchmark suites of SUPERLIM (Berdicevskis et al., 2023) and NorBench (Samuel et al., 2023) for Swedish and Norwegian. Closing in on Faroese, the now multi-lingual EuroEval (Nielsen, 2023) has been the only benchmarking platform to include Faroese language with currently four NLU tasks using datasets from e.g. Debess et al. (2024) for sentiment analysis and Simonsen et al. (2025) for question-answering. However, since the majority of the datasets are machine-generated with human validation, these may not fully reflect authentic Faroese language use, and possibly propagate biases related to LLM generated language (Xu et al., 2024). Other recent efforts to evaluate LLMs on Faroese revolve mainly around the task of machine translation (Scalvini and Debess, 2024; Scalvini et al., 2025b,c; Debess et al., 2025). Findings in this field have shown human assessment to be essential still, as automatic metrics and existing benchmarks are currently not suitable for evaluating Faroese in a nuanced manner. In general, there is a lack of native quality benchmarks for Faroese. Automatically translated or compiled and even fully synthetic datasets do not reflect sufficient linguistic and cultural knowledge, which is problematic, when using these as benchmarks to drive further development. In their study, (Müller-Eberstein et al., 2025) find that automatic evaluation using manually curated datasets aligns better with actual human evaluation than the setups using translated or synthetic datasets.

An important consideration for this work is Faroese's sociolinguistic context: the language exists in a diglossic situation where formal written standards, shaped by deliberate language planning and purification, diverge substantially from everyday spoken varieties (Jacobsen, 2021). Our benchmarks, derived from authoritative dictionaries and corrected school essays, therefore evaluate models' understanding of this prescriptive, institutionally sanctioned register - a limitation that also represents their value as assessments of the codified standard that remains culturally significant.

Evaluating language competence in machines is complex. Human language competence is holistic and not inherently systematic, but language models and quantitative evaluation need a structured approach. Structured tasks necessarily flatten natural language into discrete, testable dimensions that serve only as a proxy for real-world language competence. Accepting this premise, and limited by the data available, we set up three tasks that aim to test complementary facets of linguistic knowledge while remaining transparent about their inherent limitations.

2.2. Leveraging Dictionaries

The low-resource situation of Faroese and lack of native benchmarks naturally stem from data scarcity in general, especially structured types of data usually leveraged for evaluation. Developing native, hand-crafted, quality datasets is a resource-consuming task, which prompts the endeavour to leverage any existing resources for novel, evaluative purposes. Despite being a small language, Faroese has had a rich tradition for lexicography, resulting in a variety of mono- and bilingual dictionaries created by linguists. For a small language with limited digital resources as well as limited expert capacity for creating benchmarks from scratch, having these systematic, authoritative linguistic data is valuable. For Danish, projects have successfully transformed lexicographical resources into benchmarks for language understanding for testing model performance within e.g. synonymy, semantic association, conceptual knowledge, and sentiment analysis (Pedersen et al., 2024b), demonstrating the value of dictionary-based evaluation. For this experiment, we leverage the contents of two monolingual, native dictionaries to create three new gold-standard datasets for evaluation, see section 3.2.

2.3. Grammatical Correctness

Linguistic acceptability tests have been formalized in several influential benchmarks. CoLA (Warstadt et al., 2019) provides binary acceptability judgments for English sentences from linguistics literature, while BLiMP (Warstadt et al., 2020) uses minimal pair contrasts to isolate specific grammatical phenomena. MeLA (Zhang et al., 2024) extends this paradigm multilingually, evaluating acceptability across typologically diverse languages. For Scandinavian languages, ScaLa (Nielsen, 2023) covers regional varieties, including Faroese. However, the Faroese component is based on historical biblical texts (Ingason et al., 2012), limiting its utility for evaluating modern language. Recent efforts have been made to compile minimal pair datasets for evaluating linguistic skills in Faroese in a BLiMP-like setup (Kunz et al., 2025). However, sufficiently large manually annotated datasets for assessing linguistic quality remain scarce. Such benchmarks are critical in small-language contexts, as demonstrated for Icelandic, where recent work has developed comprehensive evaluation frameworks to assess models' linguistic quality across multiple dimensions (Ármannsson et al., 2025a,b). For under-resourced languages like Faroese, establishing similar benchmarks becomes essential not only for measuring model performance but also for ensuring that language technologies support rather than undermine the continued use and development of the language in digital contexts.

2.4. Semantics and Metaphors

Evaluating semantic relations and metaphor comprehension complements grammatical evaluation by probing beyond the models' surface level performance. Our task selection is driven by the semantic relations and metaphorical usage already tagged in the dictionary data. Using lexicographic resources and WordNets for evaluation has been a long standing practice in computational linguistics for specialized models, e.g. with SimLex and HyperLex (Hill et al., 2015; Vulić et al., 2017), and this type of task proves to be relevant also for the newer all-purpose LLMs. Evaluation of semantic relatedness and metaphor understanding still presents challenges for LLMs, as these tasks probe the depth of conceptual and associative meaning beyond surface similarity. Pedersen et al. (2024a) present a semantic benchmark from the Danish WordNet and other high-quality lexicographical resources with six semantic tasks, and find the performance of the best LLMs still leaving room for improvement.

Metaphor comprehension poses additional challenges, requiring models to recognize non-literal word usage and understand conceptual mappings between domains. Furthermore, metaphors can be culturally specific in some instances, which adds to the challenge. Using lexical data from resources like WordNet (Mohammad et al., 2016) provides structured semantic information, though such resources alone cannot capture the full pragmatic and contextual dimensions of metaphorical language. A lot of work has focused on metaphor detection, more so than on metaphor explanations, where recent efforts have evaluated LLM-generated interpretations of metaphorical expressions in Danish (Pedersen et al., 2025).

3. Methods

3.1. Experimental Design

Our evaluation framework comprises three tasks assessing different dimensions of semantic and linguistic understanding in Faroese. The first tests grammatical correctness through binary acceptability judgments. The second evaluates lexical semantic relations in both binary and multi-class settings. The third assesses metaphor comprehension through multiple-choice identification of correct metaphorical explanations.

3.1.1. Grammatical Correctness

The first task is designed as a binary choice for the correct sentence when presented with a minimal pair consisting of one correct and one incorrect sentence. Due to the nature of the data, this task employs a paradigm of 'correct' and 'incorrect' according to orthographic and grammar conventions. Therefore, it diverges from traditional linguistic acceptability judgments, which typically evaluate whether constructions are natural to native speakers and are often designed to target specific syntactic or morphological phenomena. Our approach focuses on surface-level conformity to prescriptive norms rather than gradient acceptability; we therefore adopt the term 'grammatical correctness'. This emphasis on observable rule conformity rather than subtle nuances reflects our data-constrained environment, though it remains informative for evaluating models' grasp of fundamental Faroese grammar.

3.1.2. Semantic Relations

The semantic task first employs a binary classification paradigm where models must identify polarity contrasts by distinguishing between semantically related words and antonyms when presented with a query word and two candidate options. Options are randomized for each instance to mitigate positional bias, with models required to respond with a single digit indicating their choice. Few-shot examples are provided to establish the task format, with the prompt explicitly defining semantically related words as having "meaning similar to the query word" and antonyms as having "opposite meaning to the query word". Models were prompted to identify semantically related words and antonyms in separate runs. We extend the semantic evaluation through a multi-class format, where models are provided with a query word alongside multiple options (3 words) that must be correctly classified as either antonymous, semantically related, or random outliers (Table 3).

3.1.3. Metaphorical Use Explanations

The third task shifts focus to figurative language comprehension, presenting models with a target word accompanied by an authentic example of its metaphorical usage, followed by four potential explanations or definitions of the metaphorical meaning - one correct and three distractors. This task specifically probes the model's ability to interpret non-literal semantic nuances and understand how lexical items acquire contextual meanings that deviate from their conventional definitions. Models must respond with a single letter (A-D), selecting the most appropriate explanation for the metaphorical usage within the given context. The task's difficulty lies in discriminating between four semantically similar alternatives rather than eliminating obviously incorrect distractors.

Sentence Pair	Label
Eitt ting, hvat er so tað, rithøvindurin vil hava framm við tí?	✗ incorrect
Eitt ting, hvat er so tað, rithøvundurin vil hava fram við tí?	✓ correct
Stuttsøgan er skriva í 3. persóni og í óbundnum máli...	✗ incorrect
Stuttsøgan er skrivað í 3. persóni og í óbundnum máli...	✓ correct

Table 1: Sample from Grammatical Correctness dataset.

3.2. Data and Dataset Construction

Faroese dictionaries have traditionally existed in analog formats and, while digitized in recent years, were primarily designed for user accessibility rather than computational analysis.

As all the data for the benchmarks, including dictionaries and error corpus, are created and curated by expert native speakers; we treat the data as a gold standard to represent 'ground truth' for Faroese. The datasets are available upon request.

3.2.1. Potential Training Data Overlap

The dictionaries exist online exclusively as searchable interfaces rather than downloadable structured datasets. Examination of the largest publicly available crawled Faroese corpus reveals no dictionary content, and our exported files constitute the first complete extraction of these resources from their proprietary editing environment. To the best of our knowledge, we assess the risk of model training data contamination as minimal. We therefore consider these data suitable for evaluation.

3.2.2. Grammatical Correctness

To evaluate grammatical correctness, we leverage a parallel error corpus of manually corrected school essays (Johannesen and Christiansen). The essays were collected for a thesis examining error correction (Johannesen, 2024) . We extract all incorrect sentences together with their corrected version and get 6628 sentence pairs labelled either 'correct' or 'incorrect', see Table 1.

3.2.3. Semantic Relations

For this experiment, we leveraged two Faroese dictionary sources:

- *Føroysk orðabók* (Faroese Dictionary) (Poulsen et al., 1998), the largest Faroese monolingual dictionary resource, descriptive in nature, approx. 70.000 lemmas
- *Føroysk samheitaorðabók* (Faroese Synonym Dictionary) (Thomsen, 2000), the only Faroese dictionary resource specifically focused on semantic relations, approx. 50.000 entries

The content of Faroese Synonym Dictionary is structured in the following way: for every entry lemma, a list of semantically connected lemmas is provided: synonyms, hyponyms, hypernyms, and antonyms. In the dictionary, only the antonyms are tagged as such explicitly, the other three are categorized together without further labelling. All lemmas are accompanied by semantically related words (i.e. synonyms, hyponyms, or hypernyms), but only a smaller number of lemmas are provided with antonyms. Due to this structure in the source data, we extracted all lemmas with antonyms and built a dataset with two categories per lemma: a semantically related word and an antonym (Table 2).

Type	Faroese	English
Lemma:	*kálvaknígggjaður*, adj	knock-kneed
Antonym:	*hjólbeintur*, adj	bow-legged
Sem. related:	*fyljaryssuknígggjaður*, adj	mare-legged

Table 2: Example lemma with antonym and semantically related word.

Lemmas with multiple antonyms were included multiple times, one entry per antonym, when sufficient semantically related words were available. For multi-class evaluation, we added five random outlier words per lemma, extracted from entries of the Faroese Dictionary with matching part-of-speech. The pos-tag is included in the dataset. Regarding the selection of outliers, we utilized uniform part-of-speech based sampling from the dictionary to establish a foundational baseline for lexical discrimination. We acknowledge that this approach may not probe the ability to distinguish between 'close neighbors' – words with high semantic or distributional similarity. However, given the lack of pre-existing benchmarks for Faroese, this setup serves as a first step in evaluating whether LLMs possess a basic grasp of Faroese lexical semantics before moving toward more adversarial, 'hard' distractor tasks in future work.

Type	Faroese	English
Lemma:	*andvekur*, n.	sleepless state
Antonym:	*svøvnur*, n.	sleep
Sem. related:	*hálvvekur*, n.	half-awake state
Random:	*slátur*, n.	slaughter
Random:	*skógarnám*, n.	timber line
Random:	*grasflingra*, n.	small grass turf
Random:	*verjumaður*, n.	protector
Random:	*talugøgn*, n.	speech organ

Table 3: Example lemma with antonym, semantically related, and random word.

3.2.4. Metaphorical Use Explanations

The Faroese Dictionary contains 624 lemmas tagged to denote metaphorical use, though this tagging has not been applied consistently. Only a subset of these entries includes both an authentic example of metaphorical usage and an accompanying explanation or definition. Extracting this subset yielded a curated dataset of 457 lemmas, each with a contextual example and explanation that serves as the ground truth in our evaluation setup. Examples and explanations vary in length from a few words to long sentences. The distractor explanations are sampled exclusively from explanations belonging to other target words in the same dataset based on semantic similarity. Specifically, we conducted a pairwise similarity analysis (208,846 comparisons) to identify explanations with high semantic similarity to the correct one, using Multilingual-E5-Large embeddings (Wang et al., 2024), a model that has previously proven effective in measuring semantic similarity for Faroese (Scalvini et al., 2025c). We applied a similarity cutoff threshold of 0.9 to ensure that distractors are sufficiently similar to challenge the models without being so identical that they essentially duplicate the correct answer. Our iterative qualitative review showed that candidates exceeding this threshold often functioned as semantic equivalents to the gold explanation, which would have compromised the 'single correct answer' protocol. On average, this resulted in 14 candidate distractor explanations per lemma being rejected for exceeding our similarity ceiling. Consequently, for each lemma, we selected the three highest-scoring candidates sitting immediately below this 0.90 threshold. This resulted in a distractor set with a high mean similarity of 0.89, ensuring that the options challenge the model's nuanced understanding without introducing ambiguity. By sourcing distractors from within the same dataset, the procedure ensures stylistic uniformity across all options, eliminating the risk of artefacts and models exploiting stylistic cues to distinguish correct answers from out-of-dataset distractors. Sample from the dataset is seen in Figure 1.

Lemma: *botnkav*
Example: *gera botnkav* (make a dive to the bottom)
Explanation choices:
a) ✗ *fyri ongum, sum eingin dugur longur er í*
(for nothing, of no use anymore)
b) ✗ *um eitthvørt ella einhvønn sum ómakar ella tarnar einhvønn*
(about something or someone that bothers or prevents someone)
c) ✓ *hvørva og síggjast ikki aftur eina tíð; um einhvønn ið ikki heldur tað hann hevur lovað*
(to disappear and not be seen again for a while; someone who doesn't keep promises)
d) ✗ *sagt við einhvønn ið brúkar seg og ikki kann tiga*
(said to someone who is nagging and cannot be quiet)

Lemma: *drekka*
Example: *báturin drekkur* (the boat is drinking)
Explanation choices:
a) ✗ *upp undir vangarnar til skiftis*
(up under the cheeks alternately)
b) ✗ *bregda teymin um agnhaldið*
(tie strands over the bait handle)
c) ✗ *ger seg til at stanga*
(prepares to stab)
d) ✓ *tekur sjógv inn um stokkin*
(takes in water over the gunwale)

Figure 1: Examples from Metaphorical Explanation dataset

3.3. Evaluation Setup

3.3.1. Prompting Strategy and Experimental Constraints

We only consider LLMs in this study, and we adapt them to the task by employing a few and zero-shot setting. However, given the small data volume, we prioritized the integrity and size of the test set over a traditional train/dev/test split. Splitting the data would have compromised statistical validity further. To mitigate the risk of implicit optimization on this single split, we employed a 'frozen' hyperparameter configuration across all experiments. This included a fixed temperature threshold to ensure deterministic outputs and a standardized prompt template for all models. This protocol establishes a stable baseline for zero- and few-shot evaluation in a low-resource context where a separate development set is currently unavailable.

The prompts present the overall task, explaining the input format, asking for output in a particular format, depending on the task. The task in grammatical correctness and the binary and multi-class tasks within semantic relation classification were prompted in a few-shot fashion (3-6 examples, selected based on lexicographical representativeness rather than performance-based tuning). For the metaphor explanation task, we used a zero-shot setting. All prompts were kept identical in content across models, but format styles were changed to adhere to the various model architectures. For the multi-class semantic task, we prompted both in English and in Faroese, to compare performance based on prompting language. All prompts can be found in Appendices.

3.3.2. Metrics

The primary evaluation metric for all tasks is accuracy. We also account for position bias when relevant, and for the multi-class task we present the F1 score. To account for dataset size and potential sampling noise, we report 95% confidence intervals for all accuracy scores using the Wilson score method for binomial proportions.

3.3.3. Model Selection

The selection of models for evaluation focused on getting a diverse suite of models with variety in multilingual capability, architectural setup, size, propriety, and licenses. Large-scale models such as Gemini-2.0-Flash (Google DeepMind, 2025a), GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025), and Claude-Sonnet-4 (Anthropic, 2025) were included due to their previously demonstrated strong performance in Faroese on other tasks (Scalvini et al., 2025b), and being popular user choices. Evaluating two models from OpenAI, GPT-4o and GPT-5, is motivated by investigating if there is a development in performance in understanding Faroese with newer models, as a slightly lower performance in Faroese generation for GPT-5 compared to GPT-4o has been observed in other studies (Scalvini et al., 2025a). The models in the smaller-to-medium parameter range, Gemma3-4B (Google DeepMind, 2025b) and Llama-3.2-3B (Meta AI, 2024), represent a practical sweet spot for deployment in resource-constrained environments while maintaining meaningful capabilities. The inclusion of specialized Nordic language model GPT-SW3-6.7b (Ekgren et al., 2024) reflects the importance of evaluating models with specific regional language exposure, as typological relevance have been found to positively impact transfer learning (Scalvini and Debess, 2024).

3.3.4. Human baseline

To establish a baseline performance measure, we asked human annotators to perform the two most challenging tasks: multi-class semantic relations and metaphor explanation. To keep the human evaluation task feasible for the pool of participants, we sampled 126 entries (11%) from the semantic relations multi-class dataset and 100 entries (22%) from the metaphor explanation dataset. The annotator group comprised both linguistically trained individuals and native speakers with general language proficiency. As native speakers, they were instructed in Faroese.

4. Results

In general, the benchmarks prove effective and together they provide an overall picture of how well the selected models understand Faroese language within the various domains. For all tasks, we see a clear division in performance between the larger proprietary models and the smaller open models. This is to be expected, as the proprietary models are larger, have larger training corpora and have overall higher ability to handle complex linguistic tasks.

4.1. Grammatical Correctness

For grammatical correctness in Table 4, the models exhibit a clear tiered performance structure, clustering into groups based on their accuracy scores and confidence intervals. GPT-5 and GPT-4o demonstrate strong grammatical competence with high accuracy scores, while Gemini-2.0-Flash and Claude-4-Sonnet achieve somewhat lower but still robust performance. Gemma-3-4b-it exhibits modest capabilities above baseline, whereas Llama-3.2-3b and GPT-SW3-6.7b-v2-instruct hover near random baseline.

4.2. Semantic Relations

Results from the binary semantic relations tasks are seen in Table 4, yielding performance patterns broadly consistent with the grammatical task. Gemini-2.0-Flash and GPT-5 perform with the highest accuracy, while Llama-3.2-3b and GPT-SW3-6.7b-v2-instruct perform slightly above or even below random baseline. The worse performing models have heavy positional biases, which explains the lower than random performance. All models perform slightly better when identifying antonyms rather than semantically related words, which could indicate the clearer nature of what constitutes an antonym versus the more muddy term of semantic relatedness.

Turning to the multi-class semantic task, results for both prompting languages can be found in Table 5. We observe approximately the same ranking pattern between the models as for the prior semantic task, but the accuracy is lower across the board, probably reflecting the more challenging nature of a multi-class task rather than a binary classification task. The models from OpenAI take the top rankings for both prompt languages, and interestingly, we see that the top performing models have a higher score when prompted in Faroese. The numbers for the highest performing model, GPT-5, show no overlap in confidence interval between the prompting languages. The lower performing models, especially Llama-3.2-3b and GPT-SW3-6.7b-v2-instruct, achieve a much lower score with

Model	Grammatical Correctness		Semantic Relation		Antonym	
	Overall (%)	Pos. Bias	Overall (%)	Pos. Bias	Overall (%)	Pos. Bias
GPT-5	**90.92 ±0.7**	**-2.46**	89.12 ±1.8	+8.61	**94.25 ±1.4**	**+1.58**
GPT-4o	90.28 ±0.7	-4.66	86.83 ±2.0	-5.95	90.89 ±1.7	-9.85
Gemini-2.0-Flash	83.77 ±0.9	+16.87	**90.27 ±1.7**	**-1.47**	92.40 ±1.5	+4.50
Claude-Sonnet-4	82.55 ±0.9	+4.23	85.41 ±2.1	+2.64	85.41 ±2.1	+4.80
Gemma-3-4b-it	62.16 ±1.2	+10.91	57.03 ±2.9	+46.07	57.91 ±2.9	+10.33
Llama-3.2-3b-Instruct	51.33 ±1.2	+6.11	50.31 ±2.9	-94.03	52.79 ±2.9	+73.05
GPT-SW3-6.7b-v2-instruct	49.95 ±1.2	-6.27	47.66 ±2.9	-82.27	50.13 ±2.9	-90.10

Table 4: Binary classification task performance across **Grammatical Correctness** (n=6628), **Semantic Relatedness** (n=1131), and **Antonym** (n=1131) tasks. Each block shows overall accuracy with confidence interval and positional bias (- indicates bias towards position 1, + indicates bias towards position 2). Random baseline: 50.0%.

Model	Overall Acc. (%)		F1 (Pr-EN)			F1 (Pr-FO)			Invalid (%)	
	Pr-EN	Pr-FO	Ant.	Sem.	Out.	Ant.	Sem.	Out.	Pr-EN	Pr-FO
GPT-5	**81.5 ±2.3**	**85.9 ±2.0**	0.87	0.83	0.82	0.90	0.85	0.83	6.13	0.21
GPT-4o	80.1 ±2.3	82.8 ±2.2	0.81	0.80	0.80	0.86	0.82	0.81	0.15	0.12
Gemini-2.0-Flash	78.9 ±2.4	79.9 ±2.3	0.81	0.78	0.78	0.82	0.79	0.79	0.03	0.09
Claude-Sonnet-4	72.4 ±2.6	72.9 ±2.6	0.73	0.73	0.71	0.75	0.72	0.71	0.03	0.15
Gemma-3-4b-it	44.2 ±2.9	41.1 ±2.9	0.06	0.50	0.55	0.33	0.37	0.51	0.09	0.24
Llama-3.2-3b-Instruct	38.6 ±2.8	26.4 ±2.6	0.08	0.48	0.41	0.31	0.34	0.26	0.97	27.82
GPT-SW3-6.7b-v2-instr.	29.8 ±2.7	21.8 ±2.4	0.13	0.42	0.31	0.22	0.31	0.27	13.09	38.20
Human	—	90.5 ±5.1	—	—	—	0.93	0.87	0.92	—	0.00

Table 5: Comparison of model performance on the **Semantic Relations multi-class** task in different prompting languages: **English** (Pr-EN) and **Faroese** (Pr-FO). Random baseline: 33.3%, n=1131.

no overlap in confidence interval, when prompted in Faroese. This lower performance can be explained by the larger percentage of invalidly formatted output. It seems that getting instructions in Faroese makes the smaller models struggle with actually solving the task in the accepted format, indicating fundamental challenges in instruction following that severely limit their practical utility for low-resource language tasks. On the other hand, it seems that getting instructions in Faroese is beneficial for the higher performing models. The Faroese instructions might serve as extra context for those models. The 6.13% invalid output from GPT-5 with English prompting is noteworthy, as this model otherwise demonstrates superior performance across all metrics. This pattern suggests that GPT-5's enhanced reasoning capabilities may lead it to provide more elaborate or nuanced responses that deviate from the strict output format requirements, prioritizing explanatory depth over format compliance. Results from human evaluation on the subset of the semantic multi-class task show an overall accuracy of 90.5%, slightly exceeding the score of the highest performing model. However, the confidence intervals of the human evaluation and GPT-5 with Faroese prompting overlap, indicating that the difference is not statistically conclusive at the 95% level.

4.3. Metaphorical Use Explanations

Model	Accuracy (%)
GPT-5	**88.4 ±2.9**
GPT-4o	81.8 ±3.5
Claude-Sonnet-4	79.0 ±3.7
Gemini-2.0-Flash	74.2 ±4.0
Gemma-3-4b-it	40.0 ±4.5
GPT-SW3-6.7b-v2-instruct	30.2 ±4.2
Llama-3.2-3b-Instruct	28.2 ±4.1
Human	91.0 ±5.7

Table 6: Model Performance on **Metaphor Explanation Task** with Semantically Similar Choices, cut-off at 0.9. Random baseline: 25%, n=457.

The metaphor explanation task reveals a similar performance hierarchy to previous tasks in Table 6. The performance gap between larger and smaller models is evident on this task. While Gemma-3-4b-it achieves a meaningful improvement over baseline - both Llama-3.2-3b and GPT-

SW3-6.7b-v2-instruct barely exceed random baseline. The confidence intervals are comparatively wide, indicating greater uncertainty in the estimates and substantial overlap among the top-performing models. While a clear separation exists between high- and low-performing tiers, most differences within tiers are not statistically relevant according to 95% confidence intervals. The exception is GPT-5, which performs significantly better than the other high-performing models, as indicated by non-overlapping confidence intervals.

For human evaluation, the metaphor explanation task yielded a 91% accuracy score. Human performance thus surpasses that of all models tested on the same task, with the best-performing model, GPT-5, reaching 88.4% accuracy. However, due to low sample volume, the performance is not statistically significant.

5. Discussion

5.1. Performance Patterns

The evaluation reveals consistent performance patterns across all tasks, with model rankings and accuracy scores remaining largely stable. This stability suggests that the underlying capabilities for semantic and linguistic understanding generalize across different types of meaning-related challenges. The proprietary models consistently outperform smaller open models across all benchmarks. Conversely, the smaller models struggle to exceed baseline performance, particularly on the metaphorical explanation task. The impact of prompting language is worth noting: top-performing models benefit from Faroese prompts, while smaller models experience substantial performance degradation. When we compare the models against human annotators on a subsets of the semantic multi-class and metaphor explanation tasks, the results indicate that current model performance, despite being competitive in many respects, has not yet reached the level of accuracy and nuanced understanding demonstrated by human participants in this experiment.

5.2. Methodological and Resource Contributions

This work represents the first attempt to construct native Faroese benchmarks for semantic and linguistic evaluation by systematically leveraging existing lexicographic resources. The methodology demonstrates that traditional dictionaries and error corpora, when computationally restructured, can serve as robust foundations for language technology evaluation without requiring extensive new data collection efforts and human resources. Furthermore, this approach is replicable: other low-resource languages with comparable dictionary resources could adopt similar strategies to develop native benchmarks tailored to their linguistic characteristics. By establishing this proof of concept, we provide a practical pathway for expanding evaluation capabilities beyond high-resource languages, enabling more equitable and linguistically grounded assessment of language models across diverse linguistic contexts. Another contribution of this work lies in making Faroese dictionary resources computationally accessible. Through collaboration with stakeholders, these materials are now available in machine-readable formats, establishing infrastructure for future research and applications beyond benchmark evaluation.

The grammatical correctness benchmark is highly applicable in a real-world setting, where Faroese currently struggles to develop high-quality spelling and grammar checkers, a task made more difficult by the language's rich inflectional morphology. The binary task of classifying correct vs. incorrect only in terms of spelling and grammar is relatively simple for an LLM, not even having to account for native acceptability in fluency, but for smaller, specialized models for spelling and grammar checking, this benchmark can be pivotal for development. The size of our dataset is non-comparable to other datasets of the same type. However, its value lies in being the only available native Faroese resource for systematic evaluation of grammatical competence, filling a critical gap in the development pipeline for Faroese language technology. Moreover, these benchmarks can be seen as a starting point for future expansion, involving possibly manual edits and additions by Faroese native speakers.

5.3. Sociolinguistic Considerations

The nature of the data introduces both constraints and insights. Using lexical items without context is a somewhat unnatural way of using language, and can explain why some models perform badly. On the other hand, some models perform surprisingly well classifying semantic relationships with no context. This result also shows that the models most likely rely on other mechanisms to solve tasks like these. In the metaphorical explanation task, more context is given, which is necessary for classifying the more abstract linguistic dimension of figurative or metaphorical language.

Dictionaries represent formal, standardized language rather than contemporary usage - a distinction particularly relevant for Faroese. Similarly, our error corpus, derived from corrected school essays, reflects institutionally sanctioned norms of linguistic correctness rather than naturalistic language variation. Since the 19th century, intensive language planning has produced deliberate standardization and purification from Danish influence, creating a

diglossia between the formal written standard captured in dictionaries and error corpora and everyday spoken varieties (Jacobsen, 2021). Benchmarks constructed from dictionary resources and error corpora therefore evaluate whether models understand this official, standardized register rather than colloquial modern Faroese. This creates a potential mismatch: while LLMs are predominantly trained on recent web-based materials reflecting contemporary language use, our evaluation data draws from older, prescriptively curated sources. The value of our benchmarks lies precisely in their native, expert-curated authenticity and their representation of the codified standard that remains culturally and institutionally significant. However, this also means they may not fully capture the linguistic reality of modern Faroese speakers or align closely with the types of language data models encounter during training. This gap warrants consideration when interpreting model performance and highlights broader questions about what constitutes appropriate evaluation targets for languages undergoing active standardization.

5.4. Future Work

Several key extensions would strengthen these benchmarks. First, annotating the grammatical correctness dataset with specific error categories would allow fine-grained analysis of which grammatical phenomena models find most challenging, potentially revealing systematic patterns in their understanding of Faroese-specific linguistic structures. Second, incorporating sentential context from corpora into the semantic relation tasks would enable evaluation under more naturalistic conditions, revealing whether models' semantic understanding depends critically on contextual cues. Third, including the literal meaning of metaphorically used words would expand the extent to which we are actually evaluating metaphor understanding. Fourth, the benchmark could be extended to cover colloquial or spoken language by leveraging transcripts from recordings. Finally, if the benchmark were extended to cover evaluation with other model architectures than LLMs, a training/validation split would be created.

6. Conclusion

This study introduces the first native Faroese benchmarks for evaluating grammatical competence and semantic understanding in language models by systematically repurposing lexicographic resources and error corpora. Our evaluation across three tasks - grammatical correctness, semantic relations, and metaphorical comprehension, - reveals consistent performance patterns: proprietary models achieve strong results across all benchmarks, while smaller open models struggle to exceed baseline performance. Beyond benchmark results, this work establishes a replicable methodology for low-resource language evaluation and contributes machine-readable Faroese dictionary resources for future computational applications. However, our reliance on dictionary data means these benchmarks evaluate formal, standardized Faroese shaped by prescriptive language planning rather than contemporary colloquial usage, raising important questions about linguistic representativeness and the alignment between evaluation targets and modern training data.

7. Limitations

The paper has several limitations to be addressed. First, the relatively small size of our datasets limits the statistical power of our evaluation and may not capture the full range of linguistic phenomena within each category.

Second, the benchmarks evaluate understanding of standardized Faroese rather than colloquial usage, potentially creating a mismatch with the contemporary web-based language on which LLMs are typically trained.

Third, task design imposes limitations in e.g. lack of contextual information and setting up artificial task constraints. The semantic relation tasks present lexical items without sentential or discourse context, an unnatural condition for language use. While this enables controlled evaluation of lexical knowledge, it may disadvantage models that rely on contextual cues for semantic processing and may not reflect performance in authentic language contexts. The binary and multiple-choice formats impose artificial constraints that do not reflect natural language processing. These forced-choice paradigms, while enabling systematic evaluation, may favor classification over generation capabilities and may not capture the full complexity of semantic and grammatical knowledge in spontaneous language use. Furthermore, the current benchmark uses uniformly sampled outliers for the semantic multi-class task. As Faroese computational resources evolve, future iterations of this benchmark should move toward more adversarial outlier selection and more granular semantic labeling.

Fourth, a limitation of this study is the lack of a separate development split, which introduces a theoretical risk of implicit optimization during prompt selection. While this is less of a concern for the closed-source models evaluated here, we acknowledge that it deviates from standard protocols. For future work involving smaller, tunable models on this benchmark, we recommend using cross-validation to ensure robustness.

8. Ethical considerations

Language vitality and digital inclusion Accessible language technology across diverse languages is essential to sustaining linguistic variety and supporting language vitality in the digital era. Evaluation benchmarks serve as a critical foundation for developing models and tools for small languages: without robust assessment frameworks, progress in language technology remains unmeasurable and unsystematic. By creating evaluation resources for Faroese, this work addresses the technological gap faced by speakers of underresourced languages, an increasingly important societal issue.

Data representativeness As mentioned in the Discussion, diglossia between formal and colloquial Faroese means our benchmarks may not be fully representative of everyday language use. Nevertheless, evaluation resources grounded in native, expert-curated Faroese materials provide more appropriate assessment than relying on translations from other languages or lacking benchmarks entirely.

Licensing and Benchmark Integrity All original resources used in this project are copyrighted and were acquired with explicit consent from the copyright holders. The curated datasets will be released under appropriate licenses that respect the original work. This licensing approach also helps mitigate benchmark contamination concerns, an increasingly important issue because of models being trained on ever-larger portions of publicly available web data, including evaluation benchmarks, which can lead to memorization rather than genuine demonstration of understanding.

9. Bibliographical References

Anthropic. 2025. Claude Sonnet 4. Technical report, Anthropic. Model version: claude-sonnet-4-20250514.

Bjarki Ármannsson, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2025a. Playing by the rules: A benchmark set for standardized Icelandic orthography. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 28–36, Tallinn, Estonia. University of Tartu Library.

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025b. An Icelandic linguistic benchmark for large language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

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.

Iben Nyholm Debess, Alina Karakanta, and Barbara Scalvini. 2025. What's wrong with this translation? Simplifying error annotation for crowd evaluation. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 42–47, Tallinn, Estonia. The University of Tartu Library.

Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2024. Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7814–7824, Torino, Italia. ELRA and ICCL.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL. Model variant: gpt-sw3-6.7b-v2-instruct.

Google DeepMind. 2025a. Gemini 2.0 flash model card. Technical report, Google DeepMind.

Google DeepMind. 2025b. Gemma 3: Advancing open models. Technical report, Google DeepMind. Model variant: gemma-3-4b-it.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with

(genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jógvan í Lon Jacobsen. 2021. *Føroysk Purisma*. Fróðskapur, Faroe University Press.

Uni Johannesen. 2024. Error Corpus and Large Language Models: Can LLMs understand Faroese grammar and correct errors? Master's thesis, University of the Faroe Islands.

Jenny Kunz, Iben Nyholm Debess, and Annika Simonsen. 2025. Family Matters: Language Transfer and Merging for Adapting Small LLMs to Faroese. *arXiv preprint*, arXiv:2510.00810.

Meta AI. 2024. The llama 3 herd of models. Technical report, Meta AI. Model variant: Llama-3.2-3B-Instruct.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Max Müller-Eberstein, Mike Zhang, Elisa Bassignana, Peter Brunsgaard Trolle, and Rob Van Der Goot. 2025. DaKultur: Evaluating the cultural awareness of language models for Danish with native speakers. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 50–58, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

OpenAI. 2024. GPT-4o system card. Technical report, OpenAI.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI.

Bolette Pedersen, Nathalie Sørensen, Sussi Olsen, Sanni Nimb, and Simon Gray. 2024a. 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.

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. 2024b. Evaluering af sprogforståelsen i danske sprogmodeller - med udgangspunkt i semantiske ordbøger. *NyS, Nydanske Sprogstudier*, pages 8–40.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Barbara Scalvini and Iben Nyholm Debess. 2024. Evaluating the potential of language-family-specific generative models for low-resource data augmentation: A Faroese case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6496–6503, Torino, Italia. ELRA and ICCL.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025a. A Survey of Faroese Machine Translation: The Road to Community-Driven Language Technology. Currently undergoing peer-review.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025b. Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Barbara Scalvini, Annika Simonsen, Iben Nyholm Debess, and Hafsteinn Einarsson. 2025c. Prompt engineering enhances Faroese MT, but only humans can tell. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT*

2025), pages 622–633, Tallinn, Estonia. University of Tartu Library.

Annika Simonsen, Dan Saattrup Nielsen, and Hafsteinn Einarsson. 2025. FoQA: A Faroese question-answering dataset. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 48–57, Tallinn, Estonia. University of Tartu Library, Estonia.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00321/1923697/tacl_a_00321

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.

10. Language Resource References

Anton Karl Ingason, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel C. Wallenberg. 2012. Faroese parsed historical corpus (FarPaHC) 0.1. CLARIN-IS.

Johannesen, Uni and Christiansen, Birita Dal. *Faroese Essay Error Corpus*. University of the Faroe Islands. Centre for Language Technology.

Poulsen, Jóhan Hendrik Winther and Simonsen, Marjun and Jacobsen, Jógvan í Lon and Johansen, Anfinnur and Hansen, Zakaris Svabo and Sigrunarson, Ragnar. 1998. *Føroysk orðabók*. Orðabókagrunnurin.

Thomsen, Henning. 2000. *Føroysk samheitaorðabók*. Orðabókagrunnurin.

11. Appendices

11.1. Prompts

```
You are asked to perform a semantic relation task. You are given a
query word,  with a semantically related word and an antonym in random
order.

- The semantically related word has a meaning similar to the query.
- The antonym has an opposite meaning to the query.

Your task: Identify the semantically related word.

Below are some examples:

{fewshot_prompt}

Query word: {word}
1. {options[0]}
2. {options[1]}

Which of the two is the semantically related word of the query word?
Respond ONLY with the number 1 or 2. No explanation. Just the digit.
```

Figure 2: Standard prompt for binary semantic relation task

```
You are asked to perform a semantic relation task. You are given a
query word,  with a semantically related word and an antonym in random
order.

- The semantically related word has a meaning similar to the query.
- The antonym has an opposite meaning to the query.

Your task: Identify the antonym.

Below are some examples:

{fewshot_prompt}

Query word: {word}
1. {options[0]}
2. {options[1]}

Which of the two is the semantically related word of the query word?
Respond ONLY with the number 1 or 2. No explanation. Just the digit.
```

Figure 3: Standard prompt for binary antonym task

```
You are evaluating semantic relationships in Faroese.

Target word: {item.target_word} (part of speech: {item.pos})

For each of the following words, classify its semantic relationship to
the target word as exactly one of:
- "antonym" (opposite meaning)
- "semrel" (semantically related - similar meaning, same category, etc.)
- "outlier" (unrelated/random word)

Words to classify:
"""
        for i, c in enumerate(item.choices, 1):
            prompt += f"{i}. {c['word']}\n"

        prompt += """\nProvide your answer in this exact format:
1. [word]: [relationship]
2. [word]: [relationship]
...

Example format:
1. djarvur: antonym
2. dapurskygdur: semrel
3. skaðaleysur: outlier

Your classifications:
```

Figure 4: Standard English prompt for multi-class semantic task

```
Tú metir um merkingarlig sambond í føroyskum.

Høvuðsorð: {item.target_word} (orðaflokkur: {item.pos})

Fyri hvørt av hesum orðum, flokka merkingaliga samband tess til høvuðsorðið sum júst
eitt av hesum:
- "andheiti" (mótsett merking)
- "merksam" (merkingarligt samband (líknandi merking, sami bólkur, undirheiti, o.s.fr.)
- "tilvild" (einki samband/tilvildarligt orð)

Orð at flokka:
"""
        for i, c in enumerate(item.choices, 1):
            prompt += f"{i}. {c['word']}\n"

        prompt += """\nGev títt svar eftir júst hesum frymli:
1. [orð]: [samband]
2. [orð]: [samband]
...

Dømi um frymil:
1. djarvur: andheiti
2. dapurskygdur: merksam
3. skaðaleysur: tilvild

Tínar flokkingar:
```

Figure 5: Standard Faroese prompt for multi-class semantic task

```
Read the sentence below and choose the correct explanation of
the metaphorical use of the word “{lemma}”. '

'Answer with a single letter: {", ".join(letters)}.\n\n'

'Sentence: {example}\n\n{lettered}'
```

Figure 6: Standard prompt for metafor explanation task