An Icelandic Linguistic Benchmark for Large Language Models

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies, Iceland
bjarki.armannsson@arnastofnun.is
fai@hi.is
einar.freyr.sigurdsson@arnastofnun.is

Abstract

This paper introduces a linguistic benchmark for Icelandic-language LLMs, the first of its kind manually constructed by native speakers. We report on the scores obtained by current state-of-the-art models, which indicate room for improvement, and discuss the theoretical problems involved in creating such a benchmark and scoring a model's performance.

1 Introduction

Large Language Models (LLMs) have in the last few years become near ubiquitous in the field of Language Technology (LT) and in their wake follows a growing need to test their capabilities on all kinds of tasks, such as language understanding and generation, mathematics, programming etc. As English is the dominant language in the field and the biggest source of training data for these models, it is only natural that the principal benchmarks for the models (translations aside) also focus on English. However, it is vital to also evaluate the capabilities of the models for lower-resource languages.

We introduce a standard benchmarking dataset (Ármannsson et al., 2024) to evaluate LLMs' grammatical 'knowledge' and linguistic accuracy for Icelandic, a lower-resource language. Such benchmarks can help LLM developers to improve their models' Icelandic proficiency in a measurable way and provide researchers with further insight into these models' output patterns, limitations and unexpected 'behaviour'. As far as the authors are aware, this is the first benchmark of its kind specifically constructed for Icelandic by native speakers and experts in linguistics and LT (see Section 2).

Although the models' capabilities in Icelandic are under scrutiny, we use English for all of our prompts in order to facilitate future cross-linguistic research. As one reviewer points out, it might be interesting to contrast these results with the same prompts in Icelandic, but we leave that for future work. We do not test for proficiency in standard vs. non-standard Icelandic, for instance the widespread use of dative instead of the standard accusative as the subject case of various psych verbs, like *langa* 'want' or *vanta* 'lack, need', i.e. *mér* [dat.] *langar* instead of *mig* [acc.] *langar* 'I want'. We rather aim to focus on features which should be unanimously agreed to be grammatical or ungrammatical by native speakers of Icelandic.[1]

The published benchmark set contains 1160 hand-written items over 19 subcategories of syntax, morphology and semantics, tested with five different methods (see Table 1). We also include a small set of 102 translation tasks to test a model's language understanding and grammatical capabilities in producing Icelandic text.

2 Related Work

In constructing our dataset, we partly look to similar linguistic benchmarks for LLMs that have been constructed for English. Warstadt et al. (2020)'s Benchmark of Linguistic Minimal Pairs for English (BLiMP) is perhaps the most commonly cited example. It is based around 67,000 minimal pairs, where one example is considered grammatical and the other ungrammatical, and models are tasked with 'judging' the grammatical acceptability of each sentence. (As this was before instruction-tuned models like ChatGPT-3 and the tendency towards closed black-box models,

[1] A comparison study of native human speakers, in order to confirm or challenge some of the assumptions made in the construction of this set, is currently a work in progress. Initial results, focusing only on gender agreement, indicate effectively unanimous native speaker preference for the correct answers in this benchmark and rejection of the incorrect ones.

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
March 3-4, 2025

Method	Category	No. of items
Sentence grammaticality check (yes/no)*	Simple bad/good sentences	40
	Attributive agreement	88
	Predicate agreement	28
	Word order	28
	Verb agreement	28
	Subject case	28
	Island effect sentences	80
	wh-movement	20
	Topicalization	32
	Gapping	120
	Reflexivization	40
Well-formedness check of compound nouns (yes/no)*	Word formation	280
Fill-in-the-blank	Anaphoric reference	20
	Coreference resolution	44
	Wug test (past tense of verbs)	20
Fragment answering	Fragment answers	40
Question answering	Coreference resolution	44
	Attributive agreement	30
	Word sense disambiguation	150
Total		**1160**

Table 1: The breakdown of items in our main benchmark set. All items were created manually. For the top two method types, marked with an asterisk, we also double the number of items in order to ask the inverse question, i.e. "Is this sentence grammatically **incorrect** (vs. **correct)**?". For the word sense disambiguation task, we consider pairs of sentences that contain the same lexical form and we double the number of items to ask the same question with the order of the sentence pairs reversed.

the authors simply compared the log probabilities a model assigned to sentences, i.e. making it easy to contrast how likely input sentence A was compared to input sentence B for a given model.) This general blueprint for constructing linguistic benchmarks for LLMs has been widely followed, for instance by the makers of the Zorro test suite (Huebner et al., 2021) and the ScaLa linguistic acceptability dataset for Scandinavian languages (including Icelandic) (Nielsen, 2023).

These test sets all use automatically constructed examples, which makes it possible for the BLiMP dataset, for example, to have 1,000 sentence pairs for each of the 67 grammatical tasks tested. In terms of size, our benchmark certainly pales in comparison. On the other hand, it is possible for a human to have an overview of it, whereas BLiMP is simply too large and lower-quality pairs get lost in the masses (see Vázquez Martínez et al. (2023) for more detailed criticism). In this case, we find our approach preferable, but we are also aware of its drawbacks (see Bowman and Dahl (2021) for arguments that "expert authorship" can be counterproductive, when researchers have direct, fine-grained control over the data, as it may intentionally or unintentionally lead to data "that is oriented toward linguistic phenomena that are widely studied and widely known to be important to the task at hand").

As far as interesting theoretical work on the linguistic capabilities and limitations of LLMs is concerned, there has been an ongoing and interesting debate between researchers that have used two different approaches to evaluate models in this regard. One group is represented by Dentella et al. (2023), who use acceptability judgments, widely used in traditional linguistic research, that are elicited with prompts. The other group is represented by Hu and Levy (2023), who argue that prompting is not a substitute for probability measurements in LLMs and that such metalinguistic judgments of acceptability presuppose a model's understanding of grammatical acceptability. Their approach is to compare the log probabilities of a model's output on the grounds that this gives a better idea of that model's "linguistic generalization". As much as we would have liked to imitate this approach, it was simply not possible in our one-size-fits-all setup, as closed models such as the ones provided by OpenAI and Anthropic offer limited or no access to their log probabilities.[2]

We take some inspiration from Weissweiler et al. (2023), who test the morphological capabilities of ChatGPT via a 'Wug test', where a model is tasked with forming words from nonsense root forms. We also build on the work of Sigurðsson and Nowenstein (2023), who test

[2]At testing time, OpenAI only provided the option of retrieving the top 5 'logprobs' for a models's output, i.e. the top 5 most likely tokens, which we tested in a follow-up work to this benchmark along with input log probabilities for models where those probabilities were available (work in progress).

the capabilities of GPT-4 in Icelandic, partly using methods we include in our benchmark set. Lastly, the Icelandic LT company Miðeind maintains an LLM leaderboard on HuggingFace, where a selection of LLMs are evaluated across six tasks for Icelandic: a reduced Icelandic version of Winogrande, grammatical error detection, inflection, Belebele (multiple-choice reading comprehension), machine-translated ARC-Challenge (multiple-choice question answering) and an Icelandic WikiQA dataset.[3]

3 Benchmark Composition

The benchmark was created by the authors of this paper, who have an academic background in the study of Icelandic, theoretical linguistics and LT. As already mentioned in Section 2, the point of departure were similar linguistic benchmarks for English, but we were also inspired by previous work and studies on Icelandic grammar; we point out some references below, where applicable. Some of the tasks can be applicable in a multitude of languages (such as the sentence grammaticality check), whereas others are more specific to Icelandic and languages that have more complex morphology and a richer inflectional system than, for instance, English (word formation, fill-in-the-blank and fragment-answering). See Appendix A for examples of each task.

3.1 Sentence Grammaticality Tasks

We test for acceptability of different syntactic violations, many of which are tested in similar benchmarks for English. We do this by using grammaticality judgments and prompts of the form: "Is the following Icelandic sentence grammatically correct in Icelandic? <*Example sentence in Icelandic.*> Answer only with one word, yes or no." Others, such as violations of gender agreement, are more tailored towards Icelandic grammar. To try to control for possible yes/no biases, we ask the inverse question ("[...] **incorrect** [...]") for each item. Grammaticality judgments have frequently been used in Icelandic syntax research – see, e.g., Þráinsson et al. (2013).

3.2 Word-Formation Tasks

Similar to the sentence grammaticality tasks described in Section 3.1, we ask about the well-formedness of compounds in which the first noun has one of three suffixes, *-un*, *-ing* or *-uð*, all of which are used in the genitive when they are part of the first noun in a compound: "Is the following compound word in Icelandic well-formed? <*compound*> Answer only with one word, yes or no." As with the task in Section 3.1, we ask an inverse question, trying to control for yes/no biases. For further reading on compounding in Icelandic, see, e.g., Jónsson (1984), Rögnvaldsson (1990) Bjarnadóttir (2005) and Harðarson (2016).

3.3 Fill-in-the-Blank Tasks

We include three different fill-in-the-blank tasks. One tests an LLM's ability in anaphoric reference: "Fill in the blank in the following Icelandic sentence with the correct pronoun: <*Sentence with a blank.*> Answer only with one pronoun in Icelandic." Another task looks at coreference resolution in which the context names two individuals. The continuation of each sentence contains a blank that refers to one of these individuals. The third task tests the past-tense inflection of made-up weak verbs in a Wug test (cf. the classic study by Berko 1958) – for recent studies using Wug tests with native speakers of Icelandic, see Björnsdóttir (2023) and Nowenstein (2023).

3.4 Fragment-Answering Tasks

The question *Who took my car?* does not require a whole sentence as a reply as we could answer it with, e.g. a single name, such as *Ann*. This is a fragment answer. The benchmark contains 40 wh-questions with context where the task is to give a single-word answer: "Here is an Icelandic sentence, followed by a question: <*Context sentence.*> <*Question that refers to the context.*> Answer the question with only one word in Icelandic." This task partly builds on previous work on fragment-answering in native speakers of Icelandic (e.g. Sigurðsson and Stefánsdóttir 2014, Sigurjónsdóttir and Nowenstein 2016 and Örnólfsdóttir 2017).

3.5 Question-Answering Tasks

The question-answering part includes direct questions on coreference resolution ("Which name does the pronoun <*pronoun*> refer to in the following Icelandic sentence [...]"), attributive agreement ("Which of the slash-separated options in the following question forms part of a sentence

[3] https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard

Provider	Model	Score (%)
Anthropic	claude-3-5-sonnet-20240620	**77.24**
Anthropic	claude-3-opus-20240229	71.90
OpenAI	gpt-4o-2024-08-06	72.59
OpenAI	gpt-4-turbo	62.33
OpenAI	gpt-4-0613	63.28
OpenAI	gpt-4o-mini-2024-07-18	66.21
Meta	Meta-Llama-3.1-70B-Instruct	61.21
Meta	Meta-Llama-3.1-405B-Instruct	66.47
Google	gemma-2-27b-it	59.57
Mistral AI	Mixtral-8x22B-Instruct-v0.1	48.71
Qwen	Qwen2-72B-Instruct	55.34
AI-Sweden	gpt-sw3-20b-instruct	46.12
AI-Sweden	gpt-sw3-20b-instruct-4bit-gptq	43.02

Table 2: Models tested and their overall scores.

that is grammatical in Icelandic [...]") and word-sense disambiguation ("Does the word tagged with in the following two Icelandic sentences have the same meaning [...]").

3.6 Translation Tasks

In addition to our main benchmarking set, we also include a set of 102 translation-based tasks, which contains both Icelandic sentences that should be translated into English and vice versa. These tasks are based on the assumptions that: a) Both current and future state-of-the-art models for Icelandic will be primarily trained on English text; and b) A fair way to test understanding of some feature of natural language is to ask the party in question to rephrase it in another language in which they are fluent. Our translation tasks are, as far as we are aware, a novel method of assessing the linguistic capabilities of LLMs (although similar to linguistically-oriented test suites for benchmarking machine translation systems, see e.g. Macketanz et al. 2022).

For the translation from Icelandic to English, we use new garden path sentences, which can be used to check whether a model has successfully parsed the sentence or not. For example, for the sentence *Birta Líf og Heimir niðurstöðurnar í næstu viku?* ('Will Líf and Heimir publish the results next week?'), the word *birta* needs to be read as a verb meaning 'publish' and not as the woman's name *Birta* in order for a reader to comprehend the sentence. If the name *Birta* appears in the English translation, we argue the model has not successfully parsed the sentence.

For translation from English to Icelandic, we include sentences that test: a) Gender agreement in the target output (e.g. for the source sentence *María is a good driver*, the translation of *good* should agree with the masculine *bílstjóri* ('driver'), rather than the feminine *María* in order to form a grammatical sentence), and b) Anaphoric reference in the target output (e.g. for the source sentence *The child poured the milk into the cup and checked to see whether it had gone sour*, the pronoun *it* should be translated in the feminine, *hún*, to refer to the milk rather than the cup, *bolli*, which is a masculine noun in Icelandic). As far as the gender agreement is concerned, all sentences have the same structure as the example above (i.e. <*name*> *is a* <*adjective*> <*noun*>) and we try throughout to select adjectives and nouns that should ideally only have one straightforward translation.

We emphasize that these tasks are not meant as machine translation test sets but can serve as an indicator of a model's NLU performance and grammatical capabilities in producing Icelandic text. The output needs to be manually examined, as we do not include scripts for automatic evaluation, which is why we keep these two tasks separate from the other tasks in our main benchmark. We show the results of an automatic evaluation in Section 4.2.

4 Current Model Performance

4.1 Main Benchmark Set

We show the results on our benchmark set for thirteen currently available LLMs to give an idea of the state of the art for Icelandic.[4] The models we tested are shown along with overall scores in Table 2; we show a further breakdown of scores across individual tasks in Appendix B. Anthropic and OpenAI models were accessed through their respective APIs; the Meta, Google, Mistral and Qwen models were all accessed through Together AI's API. We ran the quantized version of AI-Sweden's GPT-SW3 model locally and the non-quantized variant through a dedicated HuggingFace endpoint.[5] For the API requests, we used default settings with two exceptions, setting the temperature to 0 and restricting maximum output tokens to 5 to try and keep the models' output deterministic and brief.

[4]The models were chosen based on their standing according to the Icelandic LLM Leaderboard hosted by Miðeind and with the aim of including models from different providers.

[5]All tests were run on the 10th and 11th of October 2024, except the two models from AI-Sweden which were tested on the 10th and 13th of January 2025.

Category	Claude-3-5-Sonnet	GPT 4-o
Garden path	51.7	56.7
Agreement	68.2	63.6
Anaphora	100.0	95.0
Total score	64.7	65.7

Table 3: The scores on our set of translation tasks.

The top three scorers overall, and the only models that record over 70% accuracy, are Claude-3-5-Sonnet, GPT 4-o and Claude-3-Opus. Other models record between 43.02% and 66.47% accuracy, indicating considerable room for improvement for Icelandic-language LLMs. The scores vary considerably, however, between different tasks, as seen in Appendix B.

When scoring the outputs, we directly compare the answers obtained from the models with our reference answer but remove additional periods, spaces and the like from correct answers. It could therefore be argued that the scores we present show the models' performance in too favourable a light (see discussion in Section 5). On the other hand, for some tasks it could have been possible to mark a greater variety of answers as correct than we presently do. This is the case for coreference resolution via the 'Question-answering' method, where the models are prompted to name the noun to which a particular pronoun refers. Accounting for the complexities provided by the Icelandic case system, we consider both the particular morphological form used in the example sentence *and* the nominative form of the word (in those cases where those two forms are different) to be correct.

4.2 Translation Task Subset

As previously stated, our set of translation tasks calls for manual evaluation of a model's output. We therefore decide to score and show the results for only two models. We choose the two highest-scoring models according to our results in 4.1 (which gives us one model from each of the two best-performing 'families' of models, Anthropic's Claude and OpenAI's GPT). As seen in Table 3, the models achieve very similar scores overall, 65.7% for GPT 4-o and 64.7% for Claude-3-5-Sonnet. Both the garden path sentences and gender agreement tasks seem to present a challenge for these models but the anaphora resolution tasks are near-maximum for both.

5 Limitations

The limitations are a few. Firstly, we tried to find a suitable base prompt for each task that would be understood in the same way by different models. Even though we feel that the uniformness of the resulting answers reflects that we succeeded in this respect, we cannot be sure that some "fine-tuning" of the prompts would not have yielded better results.

Secondly, although we tried to include clear instructions in English in the prompts on what the output should be, such as "answer only with yes or no", there were some deviations in the answers. These include correct answers in Icelandic, correct answers with an additional tail (e.g. "Yes. The correct sentence"), answers that include a full stop or other additional punctuation etc. To reduce these deviations, we cleaned the model answers for scoring. A correct answer in Icelandic, for instance, was therefore considered correct, as well as answers with a "tail" etc. Even though, as one reviewer points out, post-processing methods are fairly common practice and often used by LLM evaluation frameworks such as Gao et al. (2024), for human-alignment comparisons in LLMs, such lenience has been criticized (Leivada et al., 2024), on the grounds that a human would hardly respond in such a way. We acknowledge this, but would again like to stress that this could perhaps have been avoided with more precise prompts.

It remains an open question how best to score output. In our setup, a model's answer that matches our reference answer gets one point. An answer that does not, gets none. It could be argued that this method does not highlight the differences in performance between different models sufficiently, as two models both get the same score for a wrong answer if one outputs a single pronoun in Icelandic as prompted and the other outputs gibberish. In this regard, our benchmark perhaps is better suited to measure the differences of better-performing models than capturing the differences between lesser models. We would like to encourage the further development of open-source models, which may require an evaluation that can provide information on when one of two wrong answers is more promising than another. Focusing on open-source models would also allow one to compare model output with input log probabilities of the test examples, following the work of Hu and Levy (2023). On the other hand, a more for-

giving scoring metric, based on e.g. Levenshtein distance, would simply not be applicable for our benchmark as the difference between a correct answer and an ungrammatical one is often only one or two letters.

6 Conclusions

We have presented a standard benchmarking dataset for evaluating the linguistic capabilities of LLMs for Icelandic, the first of its kind. We publish the dataset openly and describe its construction in order to hopefully aid further work in this respect for both Icelandic and Nordic NLP in a wider sense. In order to show the current state of the art for Icelandic, we show the results on our set for a variety of currently available models, which indicate considerable room for improvement for some of the tested phenomena. We also discuss some of the still-open questions regarding the best methods for testing the language capabilities of LLMs.

Acknowledgments

We would like to thank the NoDaLiDa/Baltic-HLT 2025 organizers for the assistance and communication while working on this submission and three anonymous reviewers for helpful feedback. We would also like to thank Jennifer Hu, Iris Edda Nowenstein and Þórdís Úlfarsdóttir for helpful input and feedback during the construction of our benchmark set. This work was financed by the Icelandic Ministry of Culture and Business Affairs as part of the Language Technology Programme for Icelandic.

References

Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2024. Icelandic Linguistic Benchmark for LLMs 24.10. CLARIN-IS.

Jean Berko. 1958. The child's learning of English morphology. *WORD*, 14:150–177.

Kristín Bjarnadóttir. 2005. *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum*. Orðabók Háskólans, Reykjavík.

Sigríður Mjöll Björnsdóttir. 2023. Predicting ineffability: Grammatical gender and noun pluralization in Icelandic. *Glossa: a journal of general linguistics*, 8(1).

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Gísli Rúnar Harðarson. 2016. Peeling away the layers of the onion: on layers, inflection and domains in Icelandic compounds. *Journal of Comparative Germanic Linguistics*, 19:1–47.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Baldur Jónsson. 1984. Samsett nafnorð með samsetta liði. Fáeinar athuganir. In Bernt Fossestøl, Kjell Ivar Vannebo, Kjell Venås, and Finn-Erik Vinje, editors, *Festskrift til Einar Lundeby 3. oktober 1984*, pages 158–174. Novus, Oslo.

Evelina Leivada, Vittoria Dentella, and Fritz Günther. 2024. Evaluating the language abilities of Large Language Models vs. humans: Three caveats. *Biolinguistics*, 18:Article e14391.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

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.

Iris Edda Nowenstein. 2023. *Building yourself a variable case system: The acquisition of Icelandic datives*. Doctoral dissertation, University of Iceland.

Eiríkur Rögnvaldsson. 1990. *Íslensk orðhlutafræði*, fourth edition. Málvísindastofnun Háskóla Íslands, Reykjavík.

Einar Freyr Sigurðsson and Brynhildur Stefánsdóttir. 2014. 'By'-phrases in the Icelandic New Impersonal Passive. *University of Pennsylvania Working Papers in Linguistics*, 20(1):311–320.

Sigríður Sigurjónsdóttir and Iris Nowenstein. 2016. Passives and the "New Impersonal" construction in Icelandic language acquisition. In *Proceedings of the 6th Conference on Generative Approaches to Language Acquisition North America (GALANA 2015)*, pages 110–121, Somerville, MA. Cascadilla Proceedings Project.

Einar Freyr Sigurðsson and Iris Edda Nowenstein. 2023. Nýjasta tækni og málvísindi. *Málfregnir*, 32:28–37.

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

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.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Höskuldur Þráinsson, Ásgrímur Angantýsson, and Einar Freyr Sigurðsson, editors. 2013. *Tilbrigði í íslenskri setningagerð I. Markmið, aðferðir og efniviður*. Málvísindastofnun Háskóla Íslands, Reykjavík.

Þórgunnur Anna Örnólfsdóttir. 2017. Hverjum þolmyndin glymur. Umfjöllun um *af*-liði í nýju setningagerðinni og hefðbundinni þolmynd án nafnliðarfærslu. B.A. thesis, University of Iceland, Reykjavík.

A Main Benchmark Set Task Examples

A.1 Sentence Grammaticality Tasks

All prompts in this section are of the form: "Is the following Icelandic sentence grammatically correct in Icelandic? <*Example sentence in Icelandic.*> Answer only with one word, yes or no." Below we show examples for each category in the sentence grammaticality tasks accompanied by English glosses.

A.1.1 Simple Unambiguously Grammatical/Ungrammatical Sentences

(1) **A simple ungrammatical sentence**
Blístrum þið of mjög?
whisper.1PL you.2PL too very

(2) **A simple grammatical sentence**
Sólin skín.
sun-the shines

A.1.2 Attributive Agreement

(3) **Violation**
María er góð
María(female-name) is good.FEM
bílstjóri.
driver.MASC

(4) **Correct version**
María er góður
María(female-name) is good.MASC
bílstjóri.
driver.MASC

A.1.3 Predicate Agreement

(5) **Violation**
Þessar kvikmyndir eru mjög
these.FEM.PL films.FEM.PL are very
skemmtileg.
fun.FEM.SG/NEUT.PL

(6) **Correct version**
Þessar kvikmyndir eru mjög
these.FEM.PL films.FEM.PL are very
skemmtilegar.
fun.FEM.PL

A.1.4 Word Order

(7) **Violation**
Við ekki sáum þau í garðinum.
we not saw them in garden-the

(8) **Correct version**
Við sáum þau ekki í garðinum.
we saw them not in garden-the

A.1.5 Verb Agreement

(9) **Violation**
Af hverju fór þú ekki heim?
for what went.1SG/3SG you.2SG not home?

(10) **Correct version**
Af hverju fórst þú ekki heim?
for what went.2SG you.2SG not home?

A.1.6 Subject Case

(11) **Violation**
Alexanders Daníels langar oft í bíó um helgar.
Alexander.GEN Daníel.GEN wants often in cinema on weekends

(12) **Correct version**
Alexander
Alexander.NOM/ACC/DAT
Daníel langar oft í bíó um helgar.
Daníel.NOM/ACC/DAT wants often in cinema on weekends

A.1.7 Islands

(13) **Violation**
Hvaða próf gefur kennarinn Evu góða einkunn ef hún tekur?
what exam gives teacher-the Eva good grade if she takes

(14) **Correct version**
Hvaða próf óttast kennarinn að Eva taki ekki?
what exam fears teacher-the that Eva takes not

A.1.8 Wh-movement

(15) **Violation**
Hvern taldir þú rétt að gefa hærri laun?
who.ACC thought you right to give higher salary

(16) **Correct version**
Hverjum taldir þú rétt að gefa hærri laun?
who.DAT thought you right to give higher salary

A.1.9 Topicalization

(17) **Violation**
Þessari bók gætir þú lesið.
this.DAT book could you read

(18) **Correct version**
Þessa bók gætir þú lesið.
this.ACC book could you read

A.1.10 Gapping

(19) **Violation**
Þú borðaðir kökuna og ég kleinuhringurinn
you ate cake-the.ACC and I donut-the.NOM

(20) **Correct version**
Þú borðaðir kökuna og ég kleinuhringinn.
you ate cake-the.ACC and I donut-the.ACC

A.1.11 Reflexivization

(21) **Violation**
Hún vonar að ég flýti sér.
she hopes that I hurry REFL.DAT

(22) **Correct version**
Ég vona að hún flýti sér.
I hope that she hurries REFL.DAT

A.2 Word-Formation Tasks

All prompts in this section are of the form: "Is the following compound word in Icelandic well-formed? <*compound*.> Answer only with one word, yes or no." The first part of each compound is a noun ending in *-un*, *-ing* or *-uð*, all of which are used in the genitive when they are part of the first noun in a compound.

(23) **Violation**
Sýkingþreyta.
infection.NOM-fatigue

(24) **Correct version**
Sýkingarþreyta.
infection.GEN-fatigue

A.3 Fill-in-the-blank Tasks

The prompts for the anaphoric reference task in this section are of the form "Fill in the blank in the following Icelandic sentence with the correct pronoun: <*Example sentence containing a blank*.> Answer only with one pronoun in Icelandic." The same prompt is used for the coreference resolution task, except the models are prompted for a name or noun instead of a pronoun. The prompts used for the Wug tests were as follows: "Fill in the blank in the following Icelandic sentence with the correct past tense of the verb tagged with : <*Example text showing a verb in the infinitive, tagged as stated, and then a blank to be filled with the past tense of the verb.*> Answer only with one word."

A.3.1 Anaphoric Reference

(25) Hún ætlaði að telja fuglana í
she meant to count birds-the.MASC in
tjörnunum en _ voru á flugi.
ponds-the.FEM but _ were in flight

Incorrect answer

Þær.
they.FEM

Correct answer

Þeir.
they.MASC

A.3.2 Coreference Resolution

(26) Lína ætlaði að sópa kjallarann með
Lína meant to sweep basement-the with
kústi en _ var ekki á sínum stað.
broom-the but _ was not in its place

Incorrect answer

Kjallarinn.
basement-the

Correct answer

Kústurinn.
broom-the

A.3.3 Wug Verbs

(27) Okkur langaði að krata fiskinn
we wanted to krata fish-the
örlítið, þannig að við _ hann áður en
little so that we _ it before than
hann fór í ofninn.
it went in oven-the

Correct answer

Krötuðum.
'krated'.1PL

A.4 Fragment-Answering Tasks

All prompts in this section are of the form: "Here is an Icelandic sentence, followed by a question: <*Context sentence.*> <*Question that refers to the context.*> Answer the question with only one word in Icelandic."

(28) Hún bað mig um að hjálpa sér
she.NOM asked me to help REFL.DAT
og ég gerði það. Hverjum hjálpaði ég?
and I did that who.DAT helped I

Correct answer

Henni.
her.DAT

A.5 Question-Answering Tasks

The prompts for the coreference resolution task use the same example sentences as in the fill-in-the-blank tasks. The prompts are on the form: "Which noun does the pronoun <*pronoun*> refer to in the following Icelandic sentence: <*Example sentence in Icelandic.*> Answer only with one noun." The prompts for the attributive agreement task are on the form: "Which of the slash-separated options in the following question forms part of a sentence that is grammatical in Icelandic? <*Example sentence in Icelandic with the word 'one' displayed in all three genders.*> Answer only with one word." Note the attributive agreement task does not use the same sentences as when the same feature is tested via grammaticality judgments. The prompts for the word sense disambiguation task are on the form: "Does the word tagged with in the following two Icelandic sentences have the same meaning? <*Two example sentences in Icelandic containing the same word form.*> Answer only with one word: True or False."

A.5.1 Coreference Resolution

(29) Lína ætlaði að sópa kjallarann
Lína meant to sweep basement-the.MASC
með kústi en hann var ekki
with broom-the.MASC but it.MASC was not
á sínum stað.
in its place

Incorrect answer

Kjallarinn.
basement-the

Correct answer

Kústurinn.
broom-the

A.5.2 Attributive Agreement

(30) Einn/Ein/Eitt
one.MASC/one.FEM/one.NEUT
húðflúranna var af stórum dreka.
tattoos-the.GEN.NEUT was of big dragon

Correct answer

Eitt.
one.NEUT

A.5.3 Word Sense Disambiguation

(31) **Words used in the same sense**

a. Hún nam lögfræði við
she studied law at
Háskólann.
university-the

b. Hún nam grísku við
she studied Greek at
Háskólann.
university-the

(32) **Words used in a different sense**

a. Gosið var kraftlítið.
eruption-the was weak

b. Gosið var sykurlaust.
soda-the was sugar-free

B Model Scores by Task

We break down the overall scores for each model by task included in our main benchmark set (see final page). Note that we use truncated model names due to space limitations, see Table 2 for full names.

Grammaticality checks

Model	Simple	AA	PA	WO	VA	SC	Islands	wh	Top.	Gapp.	Refl.
Claude-3-5-Sonnet	90.00	55.68	100.0	92.86	71.43	78.57	87.50	45.00	59.38	82.50	77.50
Claude-3-Opus	95.00	39.77	100.0	82.14	71.43	64.29	83.75	45.00	68.75	81.67	85.00
GPT-4o	100.0	39.77	96.43	78.57	85.71	75.00	93.75	40.00	59.38	73.33	80.00
GPT-4-Turbo	95.00	36.36	75.00	82.14	71.43	64.29	78.75	40.00	75.00	81.67	57.50
GPT-4	100.0	38.64	67.86	78.57	57.14	53.57	82.50	60.00	62.50	69.17	75.00
GPT-4o-Mini	90.00	53.41	85.71	85.71	71.43	46.43	86.25	60.00	59.38	75.00	80.00
Llama-3.1-70B	95.00	39.77	60.71	64.29	67.86	60.71	62.50	50.00	50.00	83.33	42.50
Llama-3.1-405B	85.00	30.68	64.29	60.71	60.71	57.14	85.00	50.00	50.00	72.50	80.00
Gemma-2-27B	95.00	37.50	64.29	53.57	64.29	50.00	82.50	30.00	53.13	70.83	77.50
Mixtral-8x22B	90.00	39.77	53.57	64.29	60.71	46.43	80.00	40.00	53.13	68.33	47.50
Qwen2-72B	85.00	45.45	57.14	57.14	57.14	53.57	42.50	60.00	71.88	75.83	62.50
GPT-SW3-20B	58.00	48.86	50.00	46.43	50.00	50.00	50.00	50.00	50.00	50.00	52.50
GPT-SW3-20B-4bit	55.00	51.14	50.00	39.29	50.00	50.00	65.00	40.00	46.88	46.67	45.00

Table 4: A breakdown of the overall scores for the sentence grammaticality tasks: Simple, unambigously grammatical or ungrammatical sentences (Simple), attributive agreement (AA), predicate agreement (PA), word order (WO), verb agreement (VA), subject case (SC), island effect sentences (Islands), wh-movement (wh), topicalization (Top.), gapping (Gapp.) and reflexivization (Reflex.).

Well-formedness check

Model	Word formation
Claude-3-5-Sonnet	74.29
Claude-3-Opus	67.14
GPT-4o	62.86
GPT-4-Turbo	38.57
GPT-4	59.29
GPT-4o-Mini	68.21
Llama-3.1-70B	57.14
Llama-3.1-405B	65.00
Gemma-2-27B	65.36
Mixtral-8x22B	42.86
Qwen2-72B	60.00
GPT-SW3-20B	50.00
GPT-SW3-20B-4bit	50.36

Table 5: A breakdown of the overall scores for the well-formedness check of compound nouns.

Fragment answering

Model	Fragment answers
Claude-3-5-Sonnet	100.0
Claude-3-Opus	100.0
GPT-4o	77.50
GPT-4-Turbo	72.50
GPT-4	82.50
GPT-4o-Mini	62.50
Llama-3.1-70B	72.50
Llama-3.1-405B	97.50
Gemma-2-27B	45.00
Mixtral-8x22B	25.00
Qwen2-72B	25.00
GPT-SW3-20B	0.00
GPT-SW3-20B-4bit	2.50

Table 7: A breakdown of the overall scores for the fragment answering tasks.

Fill-in-the-blank

Model	Anaphor.	Coref.	Wug
Claude-3-5-Sonnet	100.0	61.36	40.00
Claude-3-Opus	90.00	45.45	10.00
GPT-4o	80.00	52.27	40.00
GPT-4-Turbo	85.00	50.00	20.00
GPT-4	75.00	50.00	20.00
GPT-4o-Mini	45.00	27.27	20.00
Llama-3.1-70B	30.00	40.91	20.00
Llama-3.1-405B	65.00	59.09	20.00
Gemma-2-27B	50.00	13.64	10.00
Mixtral-8x22B	0.00	11.36	0.00
Qwen2-72B	25.00	18.18	0.00
GPT-SW3-20B	10.00	20.45	0.00
GPT-SW3-20B-4bit	0.00	20.45	0.00

Table 6: A breakdown of the overall scores for the fill-in-the-blank tasks: Anaphoric reference (Anaphor.), coreference resolution (Coref.) and wug tests (Wug).

Question-answering

Model	Coref.	AA	WSD
Claude-3-5-Sonnet	81.82	73.33	84.00
Claude-3-Opus	63.64	80.00	81.33
GPT-4o	86.36	80.00	90.00
GPT-4-Turbo	68.18	63.33	84.00
GPT-4	77.27	60.00	56.67
GPT-4o-Mini	59.09	50.00	66.67
Llama-3.1-70B	65.91	46.67	75.33
Llama-3.1-405B	63.64	83.33	74.67
Gemma-2-27B	59.09	36.67	62.67
Mixtral-8x22B	43.18	3.33	57.33
Qwen2-72B	47.73	33.33	65.33
GPT-SW3-20B	25.00	66.67	56.67
GPT-SW3-20B-4bit	29.55	43.33	35.33

Table 8: A breakdown of the overall scores for the question-answering tasks: Coreference resolution (Coref.), attributive agreement (AA) and word sense disambiguation (WSD).