BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation

Michelle Suijkerbuijk[1*], Zoë Prins[2*], Marianne de Heer Kloots[2], Willem Zuidema[2], and Stefan L. Frank[1**]

[1]Centre for Language Studies, Radboud University
michelle.suijkerbuijk@ru.nl, stefan.frank@ru.nl
[2]Institute for Logic, Language and Computation, University of Amsterdam
zoe.prins@hotmail.com, m.l.s.deheerkloots@uva.nl, W.H.Zuidema@uva.nl

We present a corpus of 8,400 Dutch sentence pairs, intended primarily for the grammatical evaluation of language models. Each pair consists of a grammatical sentence and a minimally different ungrammatical sentence. The corpus covers 84 paradigms, classified into 22 syntactic phenomena. Ten sentence pairs of each paradigm were created by hand, while the remaining 90 were generated semi-automatically and manually validated afterwards. Nine of the 10 hand-crafted sentences of each paradigm are rated for acceptability by at least 30 participants each, and for the same 9 sentences reading times are recorded per word, through self-paced reading. Here, we report on the construction of the dataset, the measured acceptability ratings and reading times, as well as the extent to which a variety of language models can be used to predict both the ground-truth grammaticality and human acceptability ratings.

1. Introduction

With the increasing success of neural language models (LMs) at generating human-like texts, the question arises whether these generation capabilities are driven by human-like knowledge of grammatical structures. Given their design as linear sequence processors and training on raw text data, a particular interest among researchers concerns whether such models might learn to capture the abstract hierarchical generalizations described by syntactic theory (Linzen and Baroni 2021; Kulmizev and Nivre 2022).

Whereas some studies have investigated to what extent syntactic categories and relationships can be decoded from LMs' hidden states (e.g., Tenney, Das, and Pavlick 2019; Hewitt and Manning 2019), a large body of work has approached the question through *targeted syntactic evaluations* of the models' output probabilities. These

[*] Zoë Prins and Michelle Suijkerbuijk contributed equally to the work presented in this article.
[**]Corresponding author.

Action Editor: Nianwen Xue. Submission received: 28 June 2024; revised version received: 16 January 2025; accepted for publication: 24 February 2025.

https://doi.org/10.1162/COLI_a_00559

evaluations essentially target the models' ability to assign higher probabilities to correct over incorrect word forms in specific syntactic environments of interest. For example, a model's capacity for subject-verb agreement would be reflected in a higher probability assigned to the singular verb form *is* vs. the plural verb form *are* in the context *The key to the cabinets ___ on the table* (Linzen, Dupoux, and Goldberg 2016). Note that this is related to, but differs from, the way that (binary) grammatical acceptability judgments are used as an important source of empirical data to test linguistic theories (Chomsky 1965; Schütze 2016).

Extending such initial studies on isolated syntactic constructions, researchers have worked towards more general evaluation suites that aim to offer a broader picture of model performance across larger varieties of syntactic phenomena. One line of research studies to what extent human-like grammaticality signals can be extracted from LMs by comparing against sentence acceptability corpora collected empirically or from textbook examples. For example, the Corpus of Linguistic Acceptability (CoLA) comprises a large selection of grammatical and ungrammatical sentences from the linguistics literature, on which LMs are evaluated by finetuning a binary classification head to predict grammaticality (Warstadt, Singh, and Bowman 2019). Other studies have also examined how well LMs can estimate acceptability from single sentences, by computing syntactic log odds ratio scores for direct comparison against human graded acceptability judgments (Lau, Clark, and Lappin 2017; Lu et al. 2024).

A second line of evaluation suites makes use of *minimal pairs* to evaluate LMs' grammatical capabilities. These datasets consist of pairs of minimally differing word sequences, of which only one forms a grammatical sentence. Models are then evaluated on a pairwise zero-shot forced choice task, by measuring whether they assign a higher probability to the grammatical vs. ungrammatical sequence. Warstadt et al. (2020) created the **Benchmark of Linguistic Minimal Pairs** (BLiMP) to evaluate LMs' knowledge of English grammar on a large scale. BLiMP consists of 67,000 of such pairs, semi-automatically generated using a collection of scripts and grammar rules, with samples of their output checked by human experts.

BLiMP has been extraordinarily impactful in the evaluation of English LMs, and has become a standard and major component of many evaluation benchmarks. The availability of a large corpus of grammatical minimal pairs has not only facilitated quantifying LMs' grammatical abilities, but also diving into the internal representation of grammatical knowledge in these models (Gauthier et al. 2020). In the years prior to the publication of BLiMP, some studies had already investigated LMs' performance on minimal pairs, but only with a small set of grammatical phenomena and never in a uniform way (i.e., using the same metrics and testing the same grammatical phenomena [Linzen, Dupoux, and Goldberg 2016; Marvin and Linzen 2018]). Evaluating English LMs on a large benchmark like BLiMP is thus a valuable and informative addition to the field and enables researchers to make possible "big-picture conclusions" (Warstadt et al. 2020, p. 1).

While this has been a valuable development for evaluating English LMs, there is still a much poorer understanding of the grammatical abilities of LMs in other languages. Moreover, as we will discuss below, BLiMP resulted from some specific methodological choices that are debatable. It is, important therefore, to develop similar corpora for other languages, using different (and improved) methodologies. Such sister corpora can complement BLiMP and together contribute to a better understanding of grammatical abilities of LMs across a range of languages.

In the current article, we develop a corpus of minimal pairs and acceptability judgments for Dutch (BLiMP-NL) to evaluate language models on their ability to distinguish

Dutch sentences on grammaticality, as well as to mimic human judgments. We improve on BLiMP and other existing benchmarks in (1) the creation of the minimal pairs by having native speakers create and check them and (2) the validation procedure of these minimal pairs by using more, different, and improved human measures (i.e., reading times and acceptability ratings).

2. Other Benchmarks of Syntactic Minimal Pairs

In the SyntaxGym benchmark (Hu et al. 2020), 34 syntactic phenomena are evaluated using two to four manually designed test conditions each, and model probabilities are measured at a critical region within each sentence. The BLiMP (Warstadt et al. 2020) is a larger-scale dataset comprising 67 evaluation paradigms with 1,000 sentence pairs each. These minimal pair paradigms (e.g., anaphor gender agreement, anaphor number agreement) are subcategories of 12 core phenomena covering morphology (e.g., anaphor agreement, determiner-noun agreement, irregular forms), syntax (e.g., argument structure, filler-gap dependencies, island constraints), and semantics (e.g., negative polarity item [NPI] licensing).

To systematically create evaluation pairs at this scale, Warstadt et al. automatically generated sentences from grammar templates and a vocabulary of more than 3,000 items. These grammar templates specify the general order of the sentences' phrases (e.g., subject noun phrase and verb phrase). Specific words from the vocabulary are then inserted into the templates' phrase slots to create sentences with different content. The authors verified their contrasts with human validators in a forced-choice task; five minimal pairs of each paradigm were validated by 20 native English speakers by choosing which sentence of the two was more acceptable.

Model accuracy on the BLiMP dataset is computed as the proportion of minimal pairs in which the grammatical sentence receives a higher probability than the ungrammatical sentence. Of the models evaluated by Warstadt et al., GPT-2 performed significantly better than Transformer-XL and an LSTM model, but still below human accuracy. Across phenomena, models performed the best on the morphological tasks (e.g., anaphor agreement), while tasks involving syntactic (island constraints, argument structure) and semantic (e.g., NPI licensing and quantifiers) dependencies proved to be the most challenging. The Transformer-XL and LSTM model evaluated by Warstadt et al. were trained on comparable amounts of data, while the GPT-2 model was trained on a larger dataset.

Hence, the results of Warstadt et al. remain somewhat inconclusive about the comparative advantages of different model architectures and training data quantities. However, more systematic manipulations of architecture and training data found that performance on SyntaxGym varies more across architectures than across training dataset sizes (Hu et al. 2020), in line with other work showing that increases in training corpus size yield only minimal benefits on targeted syntactic evaluation tasks beyond certain thresholds (van Schijndel, Mueller, and Linzen 2019).

There is thus a rich tradition of evaluating language models on minimal-pair benchmarks like SyntaxGym and BLiMP for English. Before turning to the evaluation of language models in other languages, it is worth pointing out that the minimal-pair methodology also has its limitations. In psycholinguistic studies, minimal pair paradigms are frequently used to elicit sentence acceptability in the binary forced choice task, but empirical data pervasively show gradience in human sentence acceptability (Sorace and Keller 2005; Aarts 2007). Such gradience in judgments may perhaps be explained by so-called performance factors, but many linguists have also argued that there is genuine variability in acceptability (Bresnan et al. 2005; Hu et al. 2024). Moreover, substantial variance in English graded acceptability judgments may be predictable from probabilistic and neural language models (Lau, Clark, and Lappin 2017; Lu et al. 2024), but such predictions require additional steps beyond simply comparing output probabilities (and methods to do so are not systematically compared in existing work).

Following the English BLiMP dataset, large-scale minimal-pair benchmarks have been developed for a range of other languages, including Mandarin Chinese (CLiMP; Xiang et al. 2021; Sling; Song et al. 2022, and ZhoBLiMP; Liu et al. 2024), Japanese (JBLiMP; Someya and Oseki 2023), Norwegian (NoCoLA; Jentoft and Samuel 2023), Swedish (DaLAJ; Volodina, Mohammed, and Klezl 2021), Russian (RuBLiMP; Taktasheva et al. 2024), Indonesian, and Tamil (Leong et al. 2023; Volodina, Mohammed, and Klezl 2021).

A recurring challenge across these efforts is the creation of a large enough sample of minimal pairs to be able to detect systematic variation in language model performance across a range of grammatical phenomena. While some benchmarks rely on expert-written minimal pairs based on grammaticality constraints described in the literature (LINDSEA, JBLiMP), the amount of work involved in such careful curation lead these datasets to remain relatively limited in size, comprising on average only up to ten minimal pairs per phenomenon (compared with 1,000 in English BLiMP). To alleviate the labor needed to create larger-scale datasets, other benchmarks make use of a variety of automation techniques. For example, DaLAJ and NoCoLA source naturally occurring ungrammatical sentences and corrected counterparts from a corpus of texts produced by second-language (L2) learners. Other benchmarks follow the approach of Warstadt et al. (2020) in generating sentences using grammar templates. CLiMP was created by sampling from a translated version of the English BLiMP vocabulary and sourcing grammar templates from a grammar book, with newly added items specific to some Chinese phenomena. For Sling and ZhoBLiMP, grammar templates were instead expert-written, based on minimal pairs sourced from textbooks and journal articles for ZhoBLiMP, and naturally occurring sentences in a treebank corpus for Sling. Finally, data for RuBLiMP was sourced by searching a parsed dependency corpus for specific lexical units and linguistic structures, and editing them using expert-written perturbation rules to create minimal pairs.

Automatic data generation procedures for minimal-pair benchmarks have clearly evolved since the original BLiMP, and any of the above procedures may be preferred for different reasons. Using grammar templates without any human interference can be problematic, as it has been shown that this approach can create unnatural and nonsensical sentences (e.g., by Warstadt et al. 2020 themselves for BLiMP, and by Song et al. 2022 for CLiMP). Sourcing ungrammatical sentences from L2 learner corpora, as in DaLAJ and NoCoLA, ensures that the benchmark includes natural occurring ungrammatical forms, but limits the types of grammatical phenomena covered to those that are challenging to L2 learners. Searching parsed syntactic corpora can be an efficient way to sample a varied range of syntactic constructions, but is similarly limited to the available data. Sourcing evaluation samples from existing corpora generally also risks contaminating the evaluation benchmark with data that large-scale pre-trained models may have seen in training. Overall, most recent benchmarks opt to include a human validation procedure in the data generation process, which seems a sensible strategy to ensure generated sentences follow the intended grammatical patterns without being unnatural.

Models evaluated using the above benchmarks span a range of architectures, including mono- and multilingual variants of common GPT- and BERT-style architectures.

Some studies comparing model performance to human validation data (e.g., CLiMP, Sling, JBLiMP) report that the gap between models and humans remains large (and larger than reported in the findings of English BLiMP), especially for phenomena involving long-distance dependencies. Effects of training data size are not widely observed, though Song et al. (2022) interestingly report that smaller variants performed better for some model architectures (though not for all).

These variations of BLiMP show that it is valuable to develop BLiMP-like datasets across languages and to improve methodologies for doing so. Moreover, in creating these datasets and evaluating LMs on them, researchers can get a better overview of the performance of LMs across languages and phenomena. The current article will discuss BLiMP-NL, a minimal-pair benchmark created for Dutch. Before we go into detail about how this dataset was developed, we will first point out what we improved on the previous benchmarks discussed and why.

3. Creating a Corpus of Minimal Pairs in Dutch

3.1 Improving Data Generation and Validation

In developing BLiMP-NL, we aimed to improve previous methodologies for minimal pair generation and validation. As discussed above, previous datasets have used different strategies for creating their minimal pairs. Although it is not entirely clear if one of these procedures is preferable, we agree with the authors of other benchmarks that an accurate balance between sentence naturalness and a wide coverage of theoretically motivated grammatical phenomena is desirable. While several other benchmarks make use of grammar templates to efficiently generate large amounts of minimal pairs, this procedure also increases workload in the careful creation of the templates and subsequent validation procedure needed to avoid non-sensical evaluation data.

For BLiMP-NL, we thus did not make use of grammar templates but manually created 10 minimal pairs per paradigm to make sure that the sentence pairs tested the intended contrast and that the sentences actually made sense. These 10 sentence pairs were then used as input for ChatGPT/GPT-3.5 Turbo to make 90 additional minimal pairs, which were all checked manually and corrected when needed, to once again make sure that they tested the intended contrast and that they made sense. BLiMP-NL thus diverges from the previous datasets in the data generation procedure.

The current study also differs from the previous datasets in its data validation procedure. In English BLiMP, as well as several subsequent benchmarks for other languages, the grammaticality contrast in the minimal pairs is validated by human native speakers in a forced-choice task. They were presented with both sentences in the minimal pair and had to choose which one they considered most acceptable. Their accuracy was computed as the proportion of minimal pairs in which the grammatical sentence was chosen over the ungrammatical sentence, similarly to that of the models described in Section 2. While a forced-choice task is easy and fast to perform by the validators, resembled the task that the LMs performed, and made it easy for the authors to validate the contrasts in the datasets, it has been shown that acceptability judgments are gradient and not binary in nature (Lau, Clark, and Lappin 2017). By binarizing, the set-up of previous BLiMP datasets thus throws away degrees of acceptability that might be relevant in judging alignment between humans and LMs. In the current study, Dutch native speakers were therefore presented with a 7-point scale for their judgment. Moreover, in the evaluation of LMs, it is important to test whether they can show sensitivity to this gradient acceptability instead of simply choosing between two sentences.

3.2 Linguistic Phenomena

BLiMP-NL consists of 84 Dutch minimal pair paradigms, which are grouped into 22 phenomena. An example minimal pair for each phenomenon can be found in Table 1. An example for each paradigm appears in Table 5 of the Appendix.

All phenomena except Crossing Dependencies were picked from a comprehensive description of Dutch syntax, namely, the 7 volumes of *The Syntax of Dutch* (Broekhuis

Table 1
Example minimal pairs for all 22 phenomena, with English glosses. *N* is the number of minimal pair paradigms within each phenomenon. The **critical word or region** is printed in bold. How the grammatical (before slash) and ungrammatical (after slash) sentences differ is noted between square brackets.

Phenomenon	*N*	Grammatical/Ungrammatical Sentence
1 Anaphor Agreement	2	[Ik bekijk/Wij bekijken] de foto van **mezelf** in de kamer. *[I watch/We watch] the photograph of **myself** in the room.*
2 Binding	2	[Ik/Mijn moeder] haatte **mezelf** op de middelbare school. *[I/My mother] hated **myself** at highschool.*
3 Argument Structure	8	Het vuur [heeft/is] een lange tijd **gebrand** voor de tent. *The fire [is/has] a long time **burnt** in front of the tent.*
4 Complementive	5	De pers geeft [in de ochtend/de aanwezige mensen] het nieuws **vrij** over de gebeurtenis. *The press gives [in the morning/the people present] the news **free** about the event.*
5 Passive	4	[Er/Yara] wordt veel **gelachen** door de vriendinnen. *[There/Yara] is a lot **laughed** by the friends.*
6 Infinitival Argument Clause	9	Hij laat [∅/dat] de technische medewerker **het account** [activeren/activeert]. *He lets [∅/that] the technical employee **the account** [activate/activates].*
7 Finite Argument Clause	6	Teun zegt [dat/∅] zijn schoonbroer ziek **is** van het eten. *Teun says [that/∅] his brother-in-law ill **is** from the food.*
8 Auxiliaries	5	De vrouw is vanmorgen [**gaan zwemmen/zwemmen gaan**] in de woeste zee. *The woman is this morning [**gone swimming/swimming gone**] in the wild sea.*
9 Adverbial modification	2	De gids woont [**er waarschijnlijk/waarschijnlijk er**] al jaren. *The guide lives [**there probably/probably there**] for years.*
10 Verb Second	2	Vorig jaar [**heeft Roos/Roos heeft**] veel ruimtes gedecoreerd voor grote bruiloften. *Last year [**has Roos/Roos has**] many rooms decorated for big weddings.*
11 Wh-Movement	6	De agent ziet [waar/wat] het kind **omheen** fietst op de weg. *The officer sees [where/what] the child **around** bikes on the road.*
12 Wh-Movement Restrictions	6	Wat [zegt/roept] de docent dat de leerling **moet** leren? *What [says/calls] the teacher that the student **must** learn?*
13 Relativization	3	De jongen [**naar wie je/wie je naar**] kijkt is mijn broer. *The boy [**at whom you/whom you at**] look is my brother.*
14 Topicalization	4	Fenna onthulde [welke/die] prijs **haar broer** had gewonnen. *Fenna revealed [what/that] prize **her brother** had won.*
15 Parasitic gaps	4	Welke boeken heeft Joris zonder [ze/∅] echt goed te **bekijken** opgeborgen? *Which books has Joris without [them/∅] really good to **inspect** stored.*
16 R-Words	2	De monteur weet dat je daar de auto [**niet mee/mee niet**] kan repareren. *The mechanic knows that you there the car [**not with/with not**] can repare.*
17 Nominalization	2	Merel hoorde van haar zoon dat [het/een] dagelijks **zeilen** erg leuk is. *Merel heard from her son that [the/a] daily **sailing** very fun is.*
18 Determiners	4	Er stond [**gisteren geen melk/geen melk gisteren**] in de koelkast. *There stood [**yesterday no milk/no milk yesterday**] in the fridge.*
19 Quantifiers	2	[Iedere student moet/alle studenten moeten] zijn **opdracht** afmaken. *[Every student has to/All students have to] his **assignment** finish.*
20 Adpositional Phrases	2	Ik weet dat Gabriël [**de jacht op zwijnen/op zwijnen de jacht**] verafschuwt. *I know that Gabriël [**the hunt for boars/for boars the hunt**] despises.*
21 Crossing Dependencies	1	Oscar heeft de athleet de marathon [**zien lopen/lopen zien**] afgelopen weekend. *Oscar has the athlete the marathon [**seen walk/walk seen**] last weekend.*
22 Extraposition	3	Jij zag dat Evi [**tevreden wegliep/wegliep tevreden**] na de bijeenkomst. *You saw that Evi [**satisfied walked away/walked away satisfied**] after the meeting.*
Total	84	

2013; Broekhuis, Corver, and Vos 2015; Broekhuis and Keizer 2012). After going through all of its 7 volumes, we came up with an almost exhaustive list of grammatical phenomena in Dutch and their paradigms; this list included over 25 phenomena and 200 paradigms. For each paradigm, we identified the critical region in the minimal pair, which is the point at which the sentence in the ungrammatical condition actually turns ungrammatical. For example, in Table 1, the ungrammatical sentence for the phenomenon Anaphor Agreement becomes truly ungrammatical from the word *mezelf* ('myself') onwards. Only the minimal pair paradigms with a clear critical region were kept. The critical region (printed in **bold** in Table 1) could consist of 1 to 4 words.

We aimed to include only those paradigms in which the sentences of a pair had only minimal differences. Therefore, we set up three conditions a minimal pair paradigm had to meet:

1. The words of the critical region must be the same for the sentences of the minimal pair.
2. At least 1 word directly preceding the critical region must be identical in the sentences of the minimal pair.
3. The number of words before the critical region can differ by a maximum of 1 word between the sentences of the minimal pair.

Mostly due to condition 1, the design of our minimal pair sentences differs from the English BLiMP: In our minimal pairs, the words of the critical region are identical (although their order can differ) between the grammatical and ungrammatical sentences. We enforce this to prevent confounds in the evaluation of both the language models and the human evaluators. For the language models, this is because it eliminates the potential effect of word frequency. For the human evaluators, we will collect acceptability ratings and reading times. Applying condition 1 (but also condition 2) is crucial for an optimal comparison of the reading times between the grammatical and ungrammatical sentences, as any differences in reading times on the critical region can then not be explained by the different characteristics of the critical region, or by spillover from the directly preceding word.

After filtering out the paradigms that did not meet all the conditions, we ended up with our current list of 84 paradigms, grouped into 22 phenomena. While 6 of these phenomena are similar to phenomena used in BLiMP (i.e., phenomena 1, 2, 3, 11, 12, and 19 in Table 1), the paradigms within these phenomena are specific to Dutch. The other 16 phenomena are specific to Dutch as well. The 22 phenomena in BLiMP-NL can be described as follows:[1]

1. **Anaphor Agreement** (*Nouns and Noun Phrases*, ch. 5): This covers the requirement that reflexive pronouns such as *mezelf* ('myself') agree with their antecedents in person and number.
2. **Binding** (*Nouns and Noun Phrases*, ch. 5): This covers the structural relationship between the reflexive pronoun and its antecedent.

[1] The volume (e.g., *Nouns and Noun Phrases*, Broekhuis and Keizer 2012; *Verbs and Verb Phrases*, Broekhuis, Corver, and Vos 2015; and *Adpositions and Adpositional Phrases*, Broekhuis 2013) and chapter of *The Syntax of Dutch* that the phenomenon was taken from is indicated after each phenomenon name.

3. **Argument Structure** (*Verbs and Verb Phrases*, ch. 2): This covers the different verb types and their characteristics, such as the number of arguments (in-/di-)transitive verbs take and the specific auxiliary (a)telic unaccusative and NOM-DAT verbs select.
4. **Complementive** (*Verbs and Verb Phrases*, ch. 2): This covers the possibility of having secondary predication on (in-/di)transitive verbs and the position of that predication.
5. **Passive** (*Verbs and Verb Phrases*, ch. 3): This covers the formation of the impersonal and regular passive construction.
6. **Infinitival Argument Clause** (*Verbs and Verb Phrases*, ch. 4/5): This covers the argument clause that is infinitival, and specifically the verbs that select this clause and the differences between the infinitival markers *te* and *om te*.
7. **Finite Argument Clause** (*Verbs and Verb Phrases*, ch. 4/5): This covers the argument clause that is finite, and specifically the obligatory complementizer, the position of the clause, and the verbs that select this clause.
8. **Auxiliaries** (*Verbs and Verb Phrases*, ch. 6/7): This covers the different types of auxiliary verbs and their behavior.
9. **Adverbial Modification** (*Verbs and Verb Phrases*, ch. 8): This covers the position of adverbs in the sentence.
10. **Verb Second** (*Verbs and Verb Phrases*, ch. 10): This covers the different word order restrictions in main and embedded clauses.
11. **Wh-Movement** (*Verbs and Verb Phrases*, ch. 11): This covers the requirements for wh-movement and the related phenomenon stranding.
12. **Wh-Movement Restrictions** (*Verbs and Verb Phrases*, ch. 11): This covers the restrictions that exist on wh-movement, such as island and superiority constraints.
13. **Relativization** (*Verbs and Verb Phrases*, ch. 11): This covers the characteristics of relativization and the restrictions thereon.
14. **Topicalization** (*Verbs and Verb Phrases*, ch. 11): This covers the characteristics of topicalization and the restrictions thereon.
15. **Parasitic Gaps** (*Verbs and Verb Phrases*, ch. 11): This covers the characteristics of parasitic gap formation.
16. **R-Words** (*Adpositions and Adpositional Phrases*, ch. 4): This covers the formation and extraction of R-words (e.g., *daar* and *er*).
17. **Nominalization** (*Nouns and Noun Phrases*, ch. 1): This covers the ways in which words from different categories can be turned into nouns.
18. **Determiners** (*Nouns and Noun Phrases*, ch. 5): This covers the special determiner *geen* 'no' and its characteristics.
19. **Quantifiers** (*Nouns and Noun Phrases*, ch. 6): This covers the behavior of quantifiers, specifically their agreement with nouns and verbs.

20. **Adpositional Phrases** (*Adpositions and Adpositional Phrases*, ch. 4): This covers the characteristics of different types of adpositional phrases, such as the PP-complement of a noun phrase or containing an R-word.
21. **Crossing Dependencies**: This covers the specific feature that verbs and arguments are ordered cross-serially.
22. **Extraposition** (*Verbs and Verb Phrases*, ch. 12): This covers the possibility of extraposing nouns and adverbs.

For all minimal pair paradigms, critical region and cue word (wherever possible) are annotated in the dataset. The cue word is the part of the sentence that causes the sentence to become ungrammatical at the critical region. This is therefore usually the part that differs between the grammatical and ungrammatical versions of a sentence. For example, in Table 1, the ungrammatical sentence for the phenomenon Anaphor Agreement is ungrammatical at the critical word *mezelf* ('myself') because of the cue word *wij* ('we') at the start of the sentence. Not all phenomena have a cue word because the difference between the sentences is sometimes due to an omission of a word or a scrambling of words, for example. In these cases, the cue word column in the dataset remains empty.

3.3 Generating Minimal Pairs

We first handcrafted 10 sentence pairs for each paradigm, resulting in a small dataset referred to as BLiMP-NL small. Subsequently, based on BLiMP-NL small, we created a larger dataset with an additional 90 sentence pairs per paradigm, referred to as BLiMP-NL large. These additional 90 sentence pairs were generated semi-synthetically using ChatGPT by giving the original 10 sentence pairs to GPT-3.5 Turbo and asking it to generate more examples. The generated examples are all manually checked and corrected where needed to ensure they correctly represent the intended paradigm, do not contain any additional errors, and were semantically plausible. It is important to note that we asked ChatGPT to recognize patterns in sentence pairs that contained a very clear pattern, including the explicit labels "correct" and "incorrect." Hence, we did not ask it to generate examples of (violations of) certain grammatical phenomena. Its ability to generate additional examples thus does not indicate the model has any understanding or knowledge of these phenomena. More information about the data generation with GPT-3.5 Turbo can be found in Appendix 8.3.

4. Collecting Native Speaker Data

4.1 Experimental Design

To test human native speakers of Dutch, the 84 paradigms were evenly divided over 7 experiments, with each experiment thus testing 12 different paradigms. Each experiment was further divided into 3 versions. Each of these versions included 3 ungrammatical sentences and 6 grammatical sentences of each paradigm (different ungrammatical sentences across the 3 versions and never both sentences of the minimal pair in the same version). Only 9 of the 10 hand-crafted minimal pairs of each paradigm were thus used in order to maintain the balance between the 3 versions of each experiment. The participants were presented with either the grammatical or the ungrammatical sentence

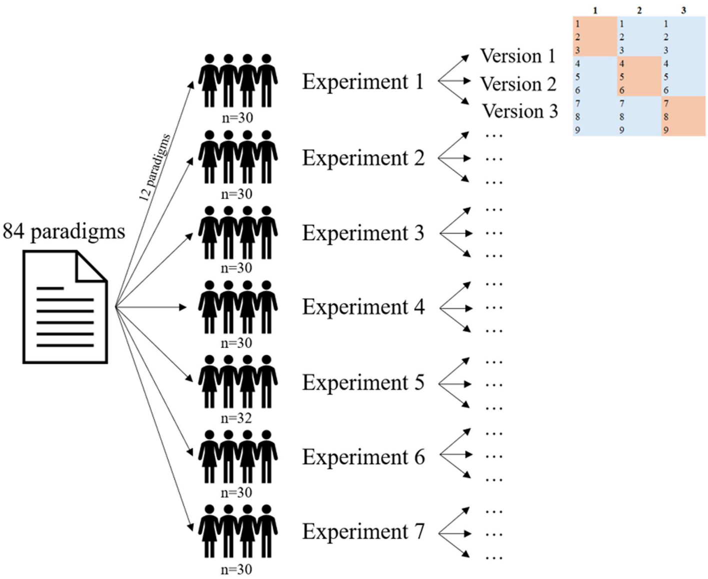


Figure 1
An illustration of the design of the experiments. *n* indicates the number of participants that completed the experiments. The columns on the right side of the figure represent the three versions, in which the orange color indicates the ungrammatical items and the blue color the grammatical items.

of a minimal pair, so each participant saw 12 × 9 = 108 sentences. The presentation order of these sentences was randomized such that sentences of the same paradigm never directly followed each other. The design process of the experiments and versions is illustrated in Figure 1.

At least 30 participants (see Figure 1 for the specific numbers), recruited via *Prolific*,[2] completed each of the 7 online experiments in *Gorilla*.[3] After being assigned to one version of an experiment, participants were instructed to read the sentences at their own pace, imagine that they were spoken by a native Dutch speaker they know well, and subsequently rate how good the sentence sounds on a 7-point scale.[4] Sentences appeared word-by-word in the middle of the screen. Each word was replaced by the next when the participant pressed the space bar. We consider the time between two keystrokes as the reading time of a word. The median completion time was just under 16 minutes and participants were paid GBP 3 for their effort. In total, 756 minimal pairs were tested (9 minimal pairs × 84 paradigms). For each of these pairs, the grammatical version received approximately 20 judgments and the ungrammatical version approximately 10 judgments.

[2] `https://www.prolific.com/`.

[3] `https://gorilla.sc/`.

[4] The exact instructions were: *Probeert u zich voor te stellen dat deze zin uitgesproken wordt door een moedertaalspreker van het Nederlands die u goed kent (bijvoorbeeld een goede vriend of vriendin). Geef vervolgens op de schaal van 1 ('Erg slecht') tot 7 ('Erg goed') aan hoe goed u de zin vond klinken door op de bijbehorende toets op uw toetsenbord te klikken. Ga uit van uw eerste intuïtie; er zijn geen goede of foute antwoorden.* (English translation: ' Try to imagine the sentence being uttered by a native Dutch speaker that you know well (for example, a close friend). Subsequently, indicate on a scale from 1 ('Very bad') to 7 ('Very good') how good the sentence sounded by clicking on the corresponding key on your keyboard. Trust your first intuition; there are no right or wrong answers.')

After data collection, we discovered spelling errors in 9 minimal pairs. Therefore, we set up an eighth experiment retesting all sentences of the 8 paradigms with the corrected sentence pairs. Each of 30 participants saw 9 sentences of these 8 paradigms, and thus 72 sentences in total. This experiment was otherwise set up in exactly the same way as the other 7 experiments. Thirty participants, recruited via *Prolific*, completed this experiment. The data for the erroneous sentence pairs have been deleted.

The reading times and acceptability ratings are available at `https://doi.org/10.34973/tj4p-y007`.

4.2 Acceptability Judgments

The average acceptability rating per phenomenon and grammaticality condition is shown in Figure 2. In the current analyses, no participants were excluded. In the data repository, we included for each participant the outcome of a Mann-Whitney U test for the difference between ratings of the grammatical and ungrammatical items. The participants that did not show a significant difference are marked to indicate that their data should be analyzed with caution.

For each paradigm (for each of the three versions of the experiment), the intraclass correlation coefficient (ICC) was calculated to test the inter-rater reliability, specifically a two-way random effect model based on the mean of multiple raters and absolute agreement. For paradigms with an ICC below 0.5, representing poor agreement between raters (Perinetti 2018), the distribution of the participant ratings was checked. These distributions were (nearly) unimodal, indicating that the disagreement between raters was more quantitative than qualitative in nature. Hence, the mean rating per condition is a meaningful measure, even in absence of strong agreement between participants. A detailed report of this check can be found in Appendix 8.2.

For the statistical analysis, the raw acceptability judgment scores were converted to z-scores per participant using all items to correct for individual differences in scale

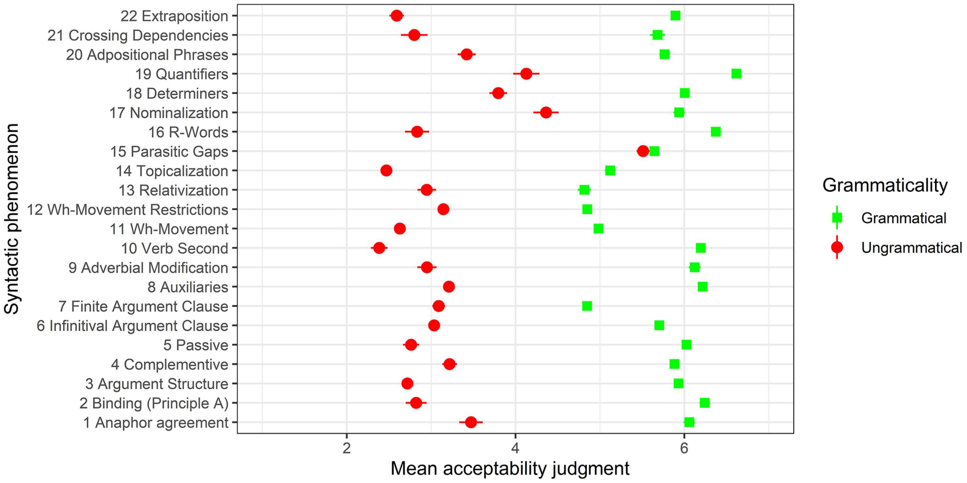


Figure 2
Mean unstandardized acceptability ratings for the ungrammatical and grammatical sentences of each phenomenon (thus averaging over different paradigms within each phenomenon). Error bars represent standard errors.

use. A linear mixed-effects (LME) model was fitted to the z-scores for each phenomenon with GRAMMATICALITY (grammatical vs. ungrammatical) as fixed effect and included a random intercept only for participants and items. After applying the multiple comparison correction of Benjamini and Hochberg (1995) to the p-values from the LME model, the results showed that all differences between the ungrammatical and grammatical conditions were significant (Parasitic Gaps: $p = .041$; other phenomena: $p < .001$). While this thus confirms the obvious pattern in Figure 2 that the ungrammatical items are judged as less acceptable than the grammatical items in the minimal pairs, Figure 2 also shows the gradience often found in acceptability ratings (Lau, Clark, and Lappin 2017). The figure shows that the (un)grammatical conditions are not equally (un)acceptable for all phenomena; the mean acceptability of the ungrammatical conditions ranges between 2 and 6 and for the grammatical conditions between 4 and 7. Moreover, for parasitic gaps, the difference between the ungrammatical and the grammatical condition is minimal. Such graded differences between conditions could not have been found by the benchmarks discussed in Section 2 as they all used a binary scale in acceptability.

4.3 Reading Times

We did not perform outlier removal in the current analysis, but we strongly advise to do so when using the reading-time data.

The log-transformed reading times were analyzed in up to three regions: (1) on the critical region only, (2) on the critical region plus the word directly following the critical region, and (3) on the critical region plus the two words directly following the critical region. Regions (2) or (3) were only included if at least two or three words, respectively, followed the critical region.

An LME model was fitted to the log-transformed reading times for each phenomenon in each region with GRAMMATICALITY (grammatical vs. ungrammatical) as fixed effect and included a random intercept only for participants and items. The coefficients for each phenomenon, accompanied by an indication of the p-value, are presented in Figure 3 for each region. The general trend in these figures, a positive difference

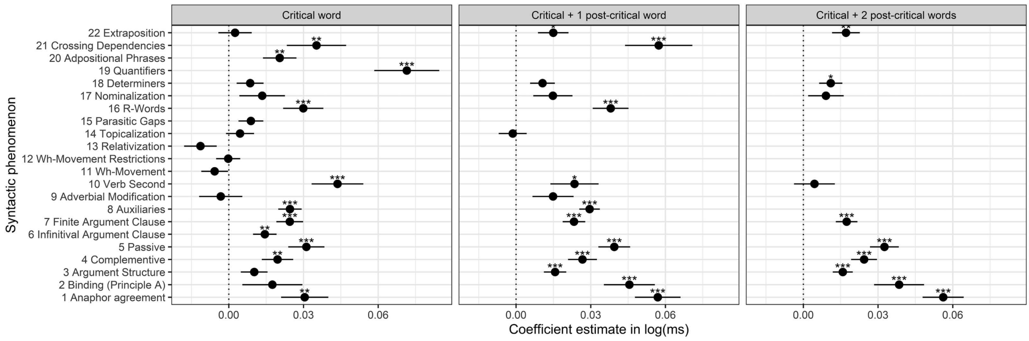


Figure 3
The coefficients for GRAMMATICALITY of the LME model for each phenomenon on the critical region (left), the critical region plus the word immediately following the critical region (middle), and on the critical region plus the two words directly following the critical region (right). In the middle and right figure, a data point for a phenomenon is only included if the corresponding post-critical word exists for that phenomenon and is not the sentence-final word. Significance stars are added for each phenomenon in each region: * represents $p < 0.025$, ** represents $p < 0.010$, and *** represents $p < 0.001$. Error bars represent standard errors.

between ungrammatical and grammatical conditions, confirms that the grammatical items are generally read faster than the ungrammatical items.

After applying the multiple comparison correction of Benjamini and Hochberg (1995) to the *p*-values from the LME model, reading times were not found to be significantly different between GRAMMATICALITY conditions for all phenomena. For the phenomena that did show a significant difference, it was in the expected direction (i.e., the grammatical condition was read faster than the ungrammatical condition).

5. Evaluating Language Models on Minimal Pairs and Acceptability Ratings

We use the dataset we created to evaluate language models on their knowledge of the grammatical phenomena. In addition to a baseline *n*-gram model, we evaluate the best-known monolingual Dutch language models and multilingual models that performed best on another Dutch model benchmark (de Vries, Wieling, and Nissim 2023). The results of these models can then be used as a benchmark to evaluate new Dutch language models.

We also use different model evaluation methods. There is not one correct way to evaluate the language models, but we will discuss the pros and cons of each evaluation method and when each method is most appropriate.

5.1 Models

We include both Transformer-encoder and Transformer-decoder (Vaswani et al. 2017) variants for our evaluation. The encoder model variants we evaluate are based on BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019). The decoder model variants we consider are based on GPT-2 (Radford et al. 2019), LLaMA (Touvron et al. 2023), and Mistral (Jiang et al. 2023). All models we evaluate and their numbers of parameters and other basic information can be found in Table 2.

We evaluate the BERT-type model BERTje (de Vries et al. 2019), a monolingual Dutch language model. This model is pre-trained with the Sentence Order Prediction (SOP) and Masked Language Modeling (MLM) objectives.

The remaining Transformer-encoder variants we evaluate are all RoBERTa-type models. We include the multilingual XLM-RoBERTa (Conneau et al. 2020) and two

Table 2
The evaluated models and their number of parameters, training set size, architecture (Encoder (Enc.) or Decoder (Dec.)), and training objective (Masked Language Modeling (MLM), Sentence Order Prediction (SOP), Standard Language Modeling (SLM), or Direct Preference Optimization (DPO)). Training set size is given in number of samples for LLaMA chat and GEITje chat, in number of bytes for GPT-2 small, both RobBERT variants, and XLM-RoBERTa, and in number of tokens for all other models.

	GPT-2 small GroNLP	GPT-2 large	LLaMA	LLaMA chat	GEITje	GEITje chat	GEITje ultra SFT	GEITje ultra DPO	BERTje	RoBERTa	RoBERTa 2023 large	XLM-RoBERTa
Parameters	124M	774M	13B	13B	7B	7B	7B	7B	109M	117M	355M	279M
Training set size	13GB	33B	33B	168K	10B	20K	240M	56M	2.4B	39GB	7GB	2.5TB
Architecture	Dec.	Dec.	Dec.	Dec.	Dec.	Dec.	Dec.	Dec.	Enc.	Enc.	Enc.	Enc.
Training objective	SLM	SLM	SLM	SLM	SLM	SLM	SLM	DPO	MLM+SOP	MLM	MLM	MLM

versions of the monolingual RobBERT (Delobelle, Winters, and Berendt 2020; Delobelle and Remy 2024). These models are only pre-trained with the MLM objective, as Liu et al. (2019) found that the SOP was not beneficial for model performance.

Furthermore, we evaluate a selection of decoder models, which have become more widely used in the last two years. All these models have a Standard Language Modeling (SLM) objective, where the model is trained to predict the next word. One of the models is trained with a Direct Preference Optimization (DPO) objective. First of all, we evaluate two GPT-2-type models. The first is an adaptation of the English GPT-2 small with retrained lexical embeddings without tuning of the Transformer layers (de Vries and Nissim 2021). The second GPT-2-based model is a GPT-2-large model trained from scratch on Dutch.[5]

We also evaluate a LLaMA-based model that is finetuned to perform better on Dutch (Vanroy 2023). This model with 13 billion parameters is a finetuned version of LLaMA 2, which already has some proficiency in Dutch albeit very limited. We evaluate both the finetuned base model, a model version that was subsequently finetuned on a collection of synthetic instruction, and chat datasets.

Furthermore, we test a collection of Mistral-based models. GEITje[6] is a model with 7B parameters, based on Mistral 7B. It is trained further with a full-parameter finetune on 10 billion tokens of Dutch text. GEITje also comes as a base model and a dedicated chat model which was finetuned on dialogue, which we both tested. Furthermore, we evaluate GEITje-ultra,[7] which is trained on even more chat data and optimized for dialogues with DPO (Rafailov et al. 2024).

Lastly, we evaluate a simple *n*-gram model on our dataset as a baseline, to investigate to what extent the shallow metric of word-string probability already suffices to assign higher probability to the correct sentence of the minimal pairs. The model we use for this is a Kneser-Ney smoothed 5-gram model, which is identical to that in Frank and Aumeistere (2024) except that it was retrained after correcting some small mistakes in the tokenization of the training data. There were only two unknown words in the sentence after tokenization.

5.2 Evaluation Methods

There are different ways to evaluate language models on BLiMP-NL. We identify several dimensions on which different evaluation choices can be made, include experiments for two evaluation methods in the main text, and describe a few more in Appendix 8.4. We use BLiMP-NL large when evaluating LMs, except when comparing to human acceptability ratings.

5.2.1 Linguistic Theory or Human Ratings. The first decision to make is what we regard as the ground truth. We see three options for this and explore two in the current article. First, we can consider the syntactic theory as the ground truth and evaluate the LMs on whether they know what linguistic phenomena are technically grammatically correct or incorrect. Second, we can regard the human acceptability ratings as the ground truth and evaluate the language models on how close they are to those ratings. Third, we could consider the reading times as the ground truth but we leave this for future work.

[5] https://huggingface.co/yhavinga/gpt2-large-dutch.

[6] https://github.com/Rijgersberg/GEITje.

[7] https://huggingface.co/BramVanroy/GEITje-7B-ultra.

The first option is most widely used in model evaluation in general. This would lead us to regard the "Grammatical" and "Ungrammatical" sentences in Table 1 as examples indicating the ground truth for grammaticality. Language models are then evaluated on whether they behave in accordance with this syntactic theory, regardless of how much actual language users agree on the sentences' grammaticality. This type of evaluation, which allows models to have "superhuman" performance, can be found in Section 6.1.

The second option evaluates language models on how well they correlate with human acceptability ratings. In this case, we take the human ratings' distribution as the ground truth and compare it to the LM scores' distribution. This also allows us to factor in that acceptability is not binary. This type of evaluation can be found in Section 6.2.

5.2.2 Language Model Probabilities or Prompting. Another principal distinction we can make is whether we evaluate the language model based on the probabilities it assigns (i.e., intrinsically) or based on its prompt responses (i.e., extrinsically). The English BLiMP was evaluated using language model probabilities but since large causal language models have gained popularity, many people also want to evaluate these models using prompting.

Intrinsic evaluation allows us to directly compare the probabilities the language model assigns to correct and incorrect sentences. An advantage of this method is that we can unquestionably ascertain whether there is a difference in the sentences' probabilities given the language model. Although this method may appear to preclude direct comparison between model and human behavior, it is the most reliable way of getting a glimpse of the LM's grammatical knowledge, so these are the only types of experiments in Section 6.

It is straightforward to obtain these probabilities for causal language models by just applying the chain rule and summing the log-likelihood values for each successive token. In practice, we can easily use the loss of the model on a sentence to get the sentence perplexity. It is not as easy for masked language models to do this, but we can estimate a pseudo-log-likelihood (PLL) score (Salazar et al. 2020; Kauf and Ivanova 2023). Salazar et al. (2020) first introduced the PLL scores, which they computed by masking each token in the sentence successively, retrieving its score with the rest of the sentence as context, and then summing the resulting values. Kauf and Ivanova (2023) expanded on this method by masking not only the target token but also all within-word tokens to its right. We use the PLL-word-l2r metric from Kauf and Ivanova (2023) since that seemed to be the most theoretically sound, by satisfying theoretical desiderata and better correlating with the scores of causal models.

It is most common to just use a sentence probability normalized by sentence length when we use language model scoring. However, it also makes sense to normalize by word frequency as this can have a significant impact on probabilities while it should not affect a sentence's grammaticality or, arguably, its acceptability. One measure that does this is the syntactic log-odds ratio (SLOR; Pauls and Klein 2012):

$$\mathrm{SLOR} = \frac{\log p_m(\zeta) - \log p_t(\zeta)}{|\zeta|} \tag{1}$$

where ζ is the sentence, $p_m(\zeta)$ is the probability of the sentence given by the model, and $p_t(\zeta)$ is the product of the token probabilities in the sentence. A token's probability is its relative frequency according to the relevant tokenizer. SLOR thus additionally normalizes for the token frequencies, making it more valid as a measure of grammaticality/acceptability. Lau, Clark, and Lappin (2017) found it was the best linking-hypothesis between LM probabilities and human grammaticality ratings in their work, albeit for non-Transformer language models. Thus, we also evaluate on the SLOR measure. The model-assigned scores for each sentence are available at `https://doi.org/10.34973/tj4p-y007`.

In contrast to intrinsic evaluation, the use of prompting is informative about the model's language generation behavior. Additionally, it allows us to evaluate large models without needing to load the model in memory. Nevertheless, prompting results can be altered considerably through "prompt engineering" and the results are very erratic in general (e.g., Holtzman et al. 2021; Zhao et al. 2021; Hu and Levy 2023). Even slight modifications in a prompt may affect the results of the model, making it very difficult to assess its performance and to compare between different models. If a small change in the wording of the question changes the model's performance, it is impossible to conclude anything about its syntactic knowledge. We do, however, include some prompting experiments in Appendix 8.4.4 to show that our dataset could be used in this way although we do not necessarily recommend it.

5.2.3 Entire Sentence or Critical Word. When we do evaluate based on model probabilities, this can also be done in different ways. Usually, these types of evaluations are performed using sentences probabilities, since the entire ungrammatical sentence should be less likely than the entire grammatical one. We evaluate LMs this way in Section 6.1.

Alternatively, one could evaluate on the probability assigned to the critical word(s). Because this is the exact place where the sentence becomes ungrammatical, we would expect this word in particular to receive a lower probability in the ungrammatical than in the grammatical sentences. In BLiMP-NL, the critical words are the same for the two sentences of a minimal pair, so only the preceding context differs and the critical words' frequencies (or other properties) cannot affect the outcome. Warstadt et al. (2020) call this method of evaluation the **two-prefix method**, based on the work of Wilcox et al. (2019). Warstadt et al. only use single critical words, while our dataset also contains critical regions with several words. We leave this evaluation for Appendix 8.4.3 since the performance difference between the two methods is only very small.

5.2.4 Minimal Pair or Single Sentence. Lastly, while the dataset consists of minimal pairs and is intended to be evaluated pair-wise, one could also attempt to judge grammaticality of individual sentences. If language models were to have actual knowledge of grammar, they should be able to judge a single sentence's grammaticality, just like humans (often) can. When evaluating with syntactic theory as ground truth, this task is still quite difficult for LMs. A dataset like CoLA (Warstadt, Singh, and Bowman 2019), which is intended for finetuning models to judge grammaticality, would be more suited for this task if we want to use syntactic theory as ground truth. When using human ratings as the ground truth, however, it does make sense to compare to single sentence scores since this was also how people were tested. We include both a pair-wise and single-sentence evaluation strategy in Section 6.2.

If one evaluates the model using prompting, it is fairly straightforward to test its ability to judge grammaticality or rate acceptability of a single sentence. One can, in theory, ask the model the exact same question that we gave people to obtain their ratings and see how it fares. In practice, this approach might be less suited to evaluate models than to evaluate people, at the current state of the technology. We do already know that asking if a sentence is grammatical with a yes/no question showcases the strong yes bias of language models in some cases and is thus not very insightful (Dentella, Günther, and Leivada 2023). Based on this kind of behavior, we might expect language models to give high ratings to all sentences. This indeed seemed to be the case, as can be seen in Appendix 8.4.4.

6. Experiments and Results

6.1 Default Analysis

First, we evaluated the language models as is common for these kinds of minimal pair datasets. That is, the overall accuracy of the LM is the proportion of minimal pairs for which the model assigns a higher probability to the grammatical than the ungrammatical sentence.

The results are in Table 3. We can see that performance differs quite noticeably between different phenomena. Specifically, performance is far lower on the Parasitic Gaps phenomenon than on all other phenomena. Notably, as we discussed in Section 4, human raters also rated ungrammatical versions of these sentences almost as high as grammatical versions. If there actually is no unacceptable variant, the ungrammatical structure will regularly occur in the training data, resulting in high probability according to the models. Moreover, the phenomena that cover Wh-Movement have a relatively low performance for all models. This is in agreement with earlier findings that longer-distance relationships are more difficult for models (Xiang et al. 2021; Song et al. 2022; Someya and Oseki 2023).

Table 3
The accuracy of evaluated language models on BLiMP-NL large using the proportion of pairs where the grammatical sentence received a higher probability than the ungrammatical sentences. Random guessing would result in an accuracy of 0.5. The Model Average gives the average over all paradigms, and the Phenomenon Average is the average over models excluding the 5-gram model. Confidence intervals for the Model Average are based on the Wilson Score Interval (Wilson 1927). The colors indicate performance, ranging from dark red for the random baseline of 0.5 to dark blue for the perfect score of 1.

Phenomenon	GPT-2 small GroNLP	GPT-2 large	LLaMA 13B	LLaMA 13B chat	GEITje 7B	GEITje 7B chat	GEITje ultra DPO	BERTje	RobBERT	RobBERT 2023 large	XLM-RoBERTa	Phenomenon Average	5-gram
1. Anaphor agreement	0.88	0.92	0.77	0.39	0.86	0.84	0.84	0.92	0.93	**0.98**	0.86	0.83	0.71
2. Binding (Principle A)	0.97	0.96	0.88	0.50	0.88	0.94	0.81	0.97	0.95	**1.00**	0.94	0.89	0.89
3. Argument Structure	0.84	0.92	0.88	0.82	0.93	0.93	0.93	**0.94**	0.91	0.89	0.85	0.89	0.59
4. Complementive	0.80	**0.94**	0.82	0.76	0.89	0.89	0.93	0.92	0.93	0.91	0.81	0.87	0.78
5. Passive	0.71	0.91	0.81	0.70	0.82	0.83	0.84	**0.93**	0.90	0.91	0.89	0.84	0.85
6. Infinitival Argument Clause	0.83	0.93	0.84	0.70	0.92	0.92	0.93	**0.95**	0.92	0.85	0.74	0.87	0.69
7. Finite Argument Clause	**0.90**	0.80	0.83	0.74	0.87	0.84	0.82	0.83	0.77	0.83	0.69	0.81	0.88
8. Auxiliaries	0.95	**1.00**	0.96	0.94	**1.00**	**1.00**	**1.00**	0.97	0.93	0.87	0.86	0.95	0.77
9. Adverbial Modification	0.96	1.00	0.96	0.92	0.99	**1.00**	0.99	0.99	0.99	0.99	0.97	0.98	0.86
10. Verb Second	0.94	1.00	1.00	0.91	**1.00**	**1.00**	**1.00**	0.99	**1.00**	0.99	0.89	0.95	0.59
11. Wh-Movement	0.62	0.88	0.84	0.74	0.88	0.88	**0.90**	0.84	0.88	0.80	0.70	0.81	0.55
12. Wh-Movement Restrictions	0.72	0.87	0.80	0.66	0.87	0.87	**0.90**	0.84	**0.90**	0.88	0.85	0.83	0.70
13. Relativization	0.61	0.82	0.93	0.91	**0.96**	**0.96**	0.95	0.94	0.95	0.93	0.92	0.90	0.61
14. Topicalization	0.86	0.92	0.91	0.82	0.87	0.84	0.87	0.96	**0.99**	0.94	0.91	0.90	0.67
15. Parasitic Gaps	0.59	0.80	0.64	0.53	0.66	0.72	0.68	**0.87**	0.76	0.79	0.71	0.70	0.49
16. R-Words	0.99	**1.00**	0.98	0.97	**1.00**	1.00	**1.00**	0.98	0.91	0.96	0.97	0.98	0.97
17. Nominalization	0.89	0.96	0.95	0.94	0.94	0.95	**0.97**	0.91	0.94	0.96	0.87	0.93	0.95
18. Determiners	0.87	0.98	0.85	0.74	0.98	**1.00**	0.95	0.96	0.93	0.88	0.82	0.90	0.69
19. Quantifiers	0.83	0.95	0.66	0.76	**0.97**	0.70	0.92	0.86	0.92	0.89	0.82	0.84	0.77
20. Adpositional Phrases	**1.00**	1.00	0.98	0.96	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99	0.95
21. Crossing Dependencies	0.93	**1.00**	**1.00**	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.92	0.98	0.42
22. Extraposition	0.96	0.99	0.98	0.87	**0.99**	**0.99**	**0.99**	0.98	0.97	0.93	0.89	0.96	0.66
Model Average	0.82	**0.92**	0.86	0.77	0.91	0.90	0.91	**0.92**	0.90	0.89	0.83	0.87	0.71
95% CI (lower)	0.81	0.91	0.85	0.76	0.90	0.90	0.90	0.92	0.90	0.88	0.82	0.87	0.70
95% CI (upper)	0.83	0.92	0.87	0.77	0.91	0.91	0.92	0.93	0.91	0.90	0.83	0.87	0.72

Results of the 5-gram model can follow only from word order probability as opposed to more abstract syntactic patterns. Nevertheless, it does well on some of the phenomena. This is most apparent for the paradigms R-words, Nominalization, and Adpositional Phrases. It is also interesting to see that neural language models sometimes perform worse than the 5-gram model. This is for instance the case for LLaMA 13B chat on the paradigm Binding (Principle A).

Overall, BERTje and GPT-2 large have the highest performance, followed closely by GEITje and GEITje ultra. The masked language models perform quite well in general. Interestingly, despite their far larger model sizes, the LLaMA and GEITje models do not perform better than a Dutch version of GPT-2 large.

In addition to using likelihoods as the acceptability measure for language models, we also investigate if the accuracy changes when we use SLOR (Pauls and Klein 2012, Equation (1)). The results when using this measure can be found in Table 6 in the Appendix. Using SLOR instead of probabilities slightly lowers performance overall, indicating that the higher probabilities of grammatical sentences might to some extent be caused by word frequency differences.

6.2 Correlation with Human Ratings

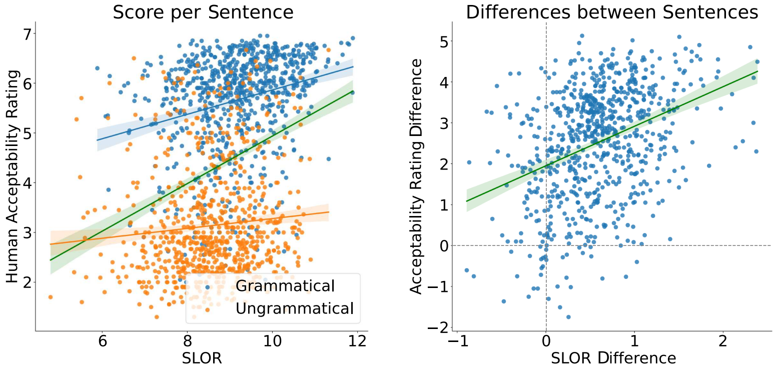
Since BLiMP-NL includes human acceptability ratings and reading times, it is also possible to evaluate model performance against human performance. We present a comparison with acceptability ratings in this article and leave the comparison with reading times for future work.

We perform our comparison with the GPT-2 large model because this was one of the best-performing models overall and estimating sentence probabilities is more suited for causal language models since they do not require pseudo-probabilities, as explained in Section 5.2.2.

There are multiple ways to then compare human ratings with model scores. One such way was introduced by Hu et al. (2024), who compared acceptability ratings with the surprisal difference, defined as the sentence surprisal for a presented sentence minus the surprisal of its minimal pair counterpart. We find it odd to compare individual sentence ratings to the difference score between sentences of a minimal pair because it implies that the human rating depends on the sentence variant that was not presented. We also saw that comparing models and humans in this way resulted in a very high correlation between surprisal differences and ratings (see Figure 8 of the Appendix).

We decide to compare SLOR scores and acceptability ratings of individual sentences. Also, we zoom in on only grammatical or only ungrammatical sentences, and on the differences between sentences of each minimal pair. The results of these comparisons can be found in Figure 4. In all comparisons, SLOR scores from GPT-2 show a positive correlation with human ratings. When looking exclusively at ungrammatical sentences, the correlation is very weak but still significant.

Comparing SLOR differences to rating difference is theoretically most sound because it factors out any effect not related to the difference in grammaticality (e.g., the arbitrary choice of words). Hence, we include such a comparison at the phenomenon level, as shown in Figure 5. Here, for each phenomenon, we take the average rating differences between grammatical and ungrammatical sentences and the average SLOR difference on BLiMP-NL large. We see that the correlation is quite high, indicating that the LM and the humans agree to a certain degree on how the acceptability difference between sentences of a minimal pair differ across phenomena.



(a) For all sentences: $\rho = 0.31, p < .0001$
For Grammatical Sentences: $\rho = 0.25, p < .0001$
For Ungrammatical Sentences: $\rho = 0.09$, $p = .012$

(b) $\rho = 0.37, p < .0001$

Figure 4
Comparison between SLOR scores and acceptability ratings calculated in four ways. Each point represents a single sentence (left) or a minimal pair (right). ρ gives the Pearson correlation between SLOR and ratings, and p its corresponding p-value. A linear regression line is calculated for the data points; the shaded area represents the 95% confidence interval.

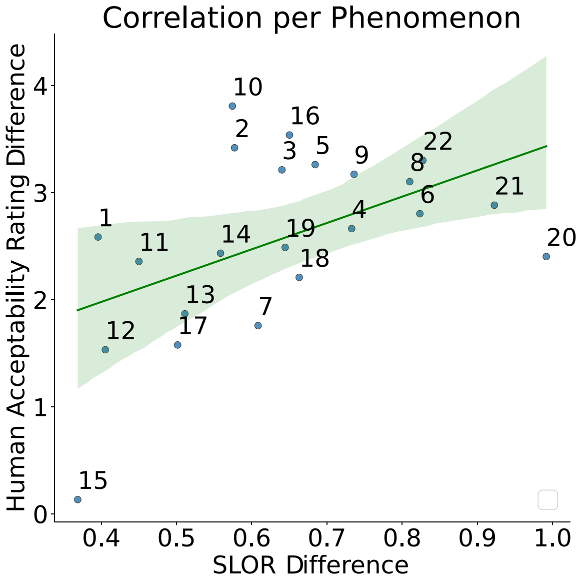


Figure 5
Comparison between SLOR scores and acceptability ratings per phenomenon (ρ = 0.49, $p = 0.019$). Each point represents the average over a phenomenon. The numbers refer to the phenomena in Table 1. A linear regression line is calculated for the data points; the shaded area represents the 95% confidence interval.

7. Discussion and Conclusions

7.1 Summary

We have introduced BLiMP-NL, a benchmark of linguistic minimal pairs for Dutch to evaluate Dutch LMs on their knowledge of grammar and the human-likeness of their grammatical abilities. BLiMP-NL consists of a small and a large dataset, used for experiments with native speakers and LMs, respectively. Both of these datasets have 84 minimal pair paradigms, selected from a comprehensive description of Dutch syntax, which are grouped into 22 phenomena. BLiMP-NL small has 10 carefully handcrafted sentence pairs per paradigm, while BLiMP-NL large supplemented these 10 with 90 semi-synthetically created sentence pairs. Both datasets and the human data can be accessed at https://doi.org/10.34973/tj4p-y007.

The creation of BLiMP-NL was inspired by BLiMP for English (Warstadt et al. 2020), but also in reaction to large-scale minimal-pair benchmarks for a range of other languages, to continue developing similar benchmarks to get a better understanding of the performance of LMs on a large scale across different languages. Moreover, in creating more of these datasets, we can keep developing the methodology. BLiMP-NL mainly improved on the other datasets in (1) having native speakers create and check the minimal pairs and (2) using more and different evaluation measures, such as reading times and acceptability ratings, for the validation procedure.

Moreover, unlike previous studies, we introduced a range of methods for LM evaluation on a minimal pair benchmark, both through theoretical grammaticality and alignment with human acceptability ratings. We presented results from two of these methods in the main article. The typical method is to take the fraction of sentence pairs for which the grammatical sentence is assigned a higher probability than the ungrammatical sentence, either through a single overall score or by looking at specific linguistic phenomena. However, we decided to also use SLOR scores as these are normalized for word frequency and sentence length.

The most notable alternative LM evaluation method using BLiMP-NL is to compare with human performance, which our dataset allows for more thoroughly than the other BLiMP variants because it includes graded acceptability ratings. We have shown that sentence probabilities from the GPT-2 large model trained on Dutch correlate significantly with these ratings.

Overall, we did not find LMs with larger training sizes to perform better when using the standard evaluation strategy. Moreover, MLMs performed similarly to larger causal language models, although these models could not be tested in the exact same way. This is quite surprising, as larger models often perform better at many tasks than their smaller counterparts. For instance, Warstadt et al. (2020) also found that training size for English LMs had a large effect on performance on BLiMP. Our findings fall more in line with the studies that also did not find a positive effect of training size, or even found a negative effect (e.g., Song et al. 2022; Xiang et al. 2021; Someya and Oseki 2023). Also in line with our findings, Oh and Schuler (2023) found that surprisal from larger transformer-based LMs provides a poorer fit to human reading times, indicating that larger models do not necessarily perform more human-like.

7.2 Conclusions and Recommendations for Using BLiMP-NL

We find that the scores and relative rankings of Dutch LMs vary considerably depending on the grammatical phenomenon and the choice of evaluation metric, although

Table 4
Ranking of causal LMs on BLiMP-NL when using different evaluation methods. The accuracy of each model is given between brackets. The first three columns result from evaluations on different kinds of probabilities and the last column results from our prompting experiment in Appendix 8.4.4.

Perplexity	SLOR	Critical Word	Prompting
GPT-2 large (0.92)	GPT-2 large (0.91)	GPT-2 large (0.92)	GEITje 7B chat (0.68)
GEITje 7B/ultra DPO (0.91)	GEITje 7B chat/ultra DPO (0.87)	GEITje 7B/chat/ultra DPO (0.91)	LLaMA 13B chat (0.61)
GEITje 7B chat (0.90)	GEITje 7B (0.86)	LLaMA 13B (0.88)	GEITje ultra DPO (0.58)

the difference in ranking between evaluation metrics is not as great, as can be seen in Table 4. Which evaluation setup to choose depends on the particular use case. The most common use, perhaps, will be to identify which LM can distinguish between “textbook” grammatical and ungrammatical sentences. For this use case, our results suggest that we should:

1. Focus on paradigms where LMs differ from each other, i.e., exclude those that almost all models handle perfectly. These are the paradigms from the phenomena Adpositional Phrases (20) and Crossing Dependencies (21).
2. Avoid evaluating on phenomena that can be handled very well by an *n*-gram model and therefore do not require abstract syntactic knowledge (see Section 6.1). These include the phenomena R-Words (16) and Nominalization (17).
3. Do not include the phenomenon Parasitic Gaps (15) because native speakers tend to accept the “ungrammatical” constructions, suggesting that they may not be truly ungrammatical (see Section 6.1).
4. Avoid relying on prompting, as prompting results don’t align well with those based on model probabilities and many models don’t handle prompt-based evaluation well (i.e., discussion about whether a model “understands” the prompt and whether its output reflects its grammatical knowledge is best avoided). Moreover, prompting excludes the use of smaller, non-chat models (see Section 5.2.2 and Appendix 8.4.4).
5. Use SLOR to correct for irrelevant effects of word frequency (see Sections 5.2.2 and 6.2).

In short, when evaluating LMs in the typical manner, phenomena 15, 16, 17, 20, and 21 can be excluded and we advise to compare LMs using SLOR scores.

Although such evaluation of language models is the most obvious use case for our dataset, it is important to stress that there are many other uses. In particular, in the ongoing debate about how much language models rely on memorization rather than productive combination (e.g., McCoy et al. 2024), the grammatical phenomena included in BLiMP-NL and other BLiMP-style datasets provide an interesting test case. Our finding that model and human ratings only correlate very weakly for ungrammatical sentences (see Figure 4a) suggests that language models have not learned the underlying linguistic generalization in all its details. This is contrary to popular conviction about the power of LLMs, perhaps, but in line with results in the literature that show LLMs are surprisingly weak at detecting grammatical errors (Kruijsbergen et al. 2024). These issues warrant further research; the availability of minimal pairs in our dataset, and other BLiMP-style datasets, provide a perfect fit with "causal intervention" methods for model analysis that rely on such minimal contrast (Arora, Jurafsky, and Potts 2024), making such interesting new research avenues possible.

7.3 Conclusions and Recommendations for Creating a Similar Dataset

Here, we list some recommendations for creating BLiMP-like datasets for other languages. First, in creating the dataset, we recommend having as much human control as possible. For BLiMP-NL, we created 10 minimal pairs per paradigm by hand to make sure they reflect the contrast that the paradigm was supposed to test. We then used GPT-3.5 Turbo to make 90 additional minimal pairs, but checked each sentence pair to again make sure it tested the intended contrast. In this way, all pairs in BLiMP-NL test the intended contrast and BliMP-NL not include any nonsensical sentences (in contrast to earlier work using grammar templates, e.g., Warstadt et al. 2020, Xiang et al. 2021).

Second, in creating the minimal pairs, we suggest making the critical word identical between sentences of a pair, in contrast to, for instance, BLiMP (Warstadt et al. 2020). In this way, an optimal comparison is created by eliminating the effect of specific word characteristics (e.g., word frequency). This allows for a more thorough evaluation of language models and reading times. Moreover, we recommend that not only the critical word be annotated, but also the cue word(s), because this can be useful when using the sentences in a human eye-tracking study (see Section 7.4).

Lastly, it is important to realize that sentence acceptability is graded rather than binary in nature. To validate the grammaticality contrast in the minimal pairs, we thus recommend collecting graded ratings with a similar method as described in Section 4 instead of using a binary forced-choice task as in previous minimal pair benchmarks.

7.4 Limitations and Future Work

In the course of creating and evaluating BLiMP-NL, we identified several potential limitations and weaknesses that could inspire future work. First, we used ChatGPT (GPT-3.5 Turbo) for the creation of BLiMP-NL large and subsequently evaluated other LMs on these sentences. Evaluating models on model-generated sentences might artificially inflate model performance. While we did check each example manually, there might still be some patterns in the sentences that give an advantage to LMs with similar training data or training strategy to GPT-3.5 Turbo. As can be seen in Appendix 8.3, the models we evaluated did not show great differences in performance between hand-crafted and model-generated sentences, but models more closely resembling GPT-3.5 Turbo might.

Second, we have no information on the occurrence frequencies of the BLiMP-NL phenomena (and paradigms within them). More frequent phenomena will no doubt be better learned by the models, and this might be why some of the phenomena could be considered too easy for the LMs. There are a few phenomena on which most models perform almost perfectly, making it difficult to distinguish between models' performance on these phenomena. In fact, abstract syntactic knowledge is not always required to identify the grammatical sentence of a pair. This is the case when the 5-gram model scores highly, indicating that models can rely on the frequencies of contiguous sequences of words for these paradigms. One could, of course, decide to remove the "easy" phenomena from the dataset, although it will still be telling if a model fails to score perfectly, or possibly even worse than the 5-gram model, on these cases. Overall, accuracies are still low enough on average to distinguish between models.

Third, within each BLiMP-NL paradigm, sentences are quite uniform in length and structure. BLiMP-NL small contained little variation to start with, so GPT-3.5 replicated this lack of variation when generating the additional sentence pairs. This lack of variation might reduce confidence in the generalizability of results.

Moreover, our self-paced reading method could be further improved in future studies. While we saw in the acceptability judgment study that the ungrammatical condition was rated significantly less acceptable than the grammatical condition for all phenomena, this was not the case for the reading times: The grammatical sentences were generally read faster than the ungrammatical sentences, but the reading times did not differ significantly between these conditions for all phenomena. This could be due to two reasons. First, although previous studies have argued that there is no qualitative difference between the results of Web-based and lab-based experiments (Hilbig 2015; Keller et al. 2009; Semmelmann and Weigelt 2016), we should consider the possibility that running our self-paced reading experiment on the Web introduced more measurement noise than in a controlled (e.g., lab) environment, reducing statistical reliability. Second, Laurinavichyute and Von der Malsburg (2024) found that an acceptability rating task increases participants' expectation that they will encounter ungrammatical sentences. Consequently, for the current study, participants will be less surprised to read an ungrammatical sentence, reducing the reading time effects. Both of these points should be carefully considered in future studies using the current methods.

Lastly, presenting the sentences word-by-word resembles an auditory acceptability judgment task in that the reader cannot return to earlier words, and this might have affected the participants' judgments. Several studies have found that acceptability judgments are affected by their modality, for example, participants are less accurate and slower in an auditory task (Murphy 1997), and their ratings are more favorable as processing the spoken word is more demanding (Yi and Park 2024). Consequently, presenting the sentences word-by-word instead of as full sentences might have influenced the participants' judgments, and this is an important point to consider in future studies.

Other lines of future work the BLiMP-NL corpus could be used for is to compare the difference in human ratings between paradigms or phenomena to investigate if models' correlation with human performance differs across linguistic phenomena. Additionally, comparing LM word surprisal scores with the reading times we collected might be very insightful. In this case we would mainly focus on comparing at the critical word. Moreover, we have recently collected eye-tracking data from native speakers reading BLiMP-NL sentences, specifically from paradigms which also have cue words. In future work, we plan to test the hypothesis that readers look back at the cue word when they arrive at the critical word. In the same analysis, it would also be interesting to compare the eye-tracking data with LMs by using attribution methods (e.g., Covert, Lundberg, and Lee 2021; Jumelet and Zuidema 2023) to see whether the cue word is more important in the prediction of the critical word.

8. Appendix

8.1 All Paradigms

Table 5 lists all paradigms and shows an example sentence pair for each.

Table 5
Example sentence pairs for all 84 paradigms. Critical words are printed in **bold**. The first column gives the phenomenon number each paradigm falls under.

Ph.	Paradigm	Grammatical/Ungrammatical Sentence
1	1. number	[Ik bekijk/Wij bekijken] de foto van **mezelf** in de kamer. *[I watch/We watch] the photograph of **myself** in the room.*
	2. person	[Jij/Hij] bekijkt de foto van **jezelf** in de kamer. *[You/He] watches the photograph of **yourself** in the room.*
2	3. c_command	[Ik/Mijn moeder] haatte **mezelf** op de middelbare school. *[I/My mother] hated **myself** at highschool.*
	4. monomorpemic	[Zij/Jij] legde **zich** neer bij de beslissingen tijdens de pandemie. *[She/You] resigned **herself** with the decisions during the pandemic.*
3	5. argument_number_(in)transitive	Adam [bezoekt/loopt] graag af en toe **een museum** tijdens de vakantie. *Adam [visits/walks] happily once in a while **a museum** during the vacation.*
	6. argument_number_ditransitive	Mila [ontnam/beroofde] Lars **de motor** in de echtscheiding. *Mila [deprived/stole] Lars **the motorcycle** in the divorce.*
	7. intransitive_unaccusative_1	De toerist [is/heeft] laat in de avond **gearriveerd** bij het verblijf. *The tourist [is/has] late in the evening **arrived** at the accommodation.*
	8. intransitive_unaccusative_2	Amy [heeft/is] tot haar vijftiende **gedanst** op professioneel niveau. *Amy [has/is] until her fifteenth **danced** on a professional level.*
	9. intransitive_unaccusative_3	Het vuur [heeft/is] een lange tijd **gebrand** voor de tent. *The fire [is/has] a long time **burnt** in front of the tent.*
	10. ditransitive_nomdat_1	De behandeling [is/heeft] Jasmijn heel erg **meegevallen** vorige week zaterdag. *The treatment [is/has] Jasmijn very much **appreciated** last week Saturday.*
	11. ditransitive_nomdat_2	De vriendengroep [heeft/is] Ryan eens goed **aangesproken** op zijn gedrag. *The friendgroup [has/is] Ryan once properly **addressed** on his behaviour.*
	12. ditransitive_nomdat_3	De grootouders [hebben/zijn] hem veel geld **nagelaten** in hun testament. *The grandparents [have/are] him much money **inherited** in their will.*
4	13. position_verb	Dirk zei dat de man de minnaar meteen [**dood sloeg/sloeg dood**] in het café. *Dirk said that the man the lover immediately [**dead beat/beat dead**] in the café.*
	14. position_adverb	Dirk zei dat de man de minnaar [**meteen dood/dood meteen**] sloeg in het café. *Dirk said that the man the lover [**immediately dead/dead immediately**] beat in the café.*
	15. intransitive	Rick huilde [zijn ogen/afgelopen vrijdag] echt helemaal **rood** na dit nieuws. *Rick cried [his eyes/last Friday] completely **red** after this interview.*
	16. transitive	De man [sloeg/mishandelde] de minnaar **dood** in het café. *The man [beat/abused] the lover **dead** in the café.*
	17. ditransitive	De pers geeft [in de ochtend/de aanwezige mensen] het nieuws **vrij** over de gebeurtenis. *The press gives [in the morning/the people present] the news **free** about the event.*
5	18. impersonal	[Er/Yara] wordt veel **gelachen** door de vriendinnen. *[There/Yara] is a lot **laughed** by the friends.*
	19. ditransitive_1	Het boek [wordt/krijgt] de nieuwe studenten **toegestuurd** voor de cursus. *The book [is/gets] the new students **sent** for the course.*
	20. ditransitive_2	De studenten [krijgen/worden] het nieuwe boek **toegestuurd** voor de cursus. *The students [get/are] the new book **sent** for the course.*
	21. AcI	[Zij heeft/Er wordt] de man **laten** dansen tijdens de bruiloft. *[She has/There is] the man **let** dance during the wedding.*
6	22. verb_type	Hij laat [∅/dat] de technische medewerker **het account** [activeren/activeert]. *He lets [∅/that] the technical employee **the account** [activate/activates].*
	23. bare_verb_cluster	Rosie denkt dat de gespierde man [**het pakket kan/kan het pakket**] dragen. *Rosie thinks that the muscular man [**the package can/can the package**] carry.*
	24. bare_verb_type_1	Mara weet dat de professor [haar/trots] de diersoort zag **onderzoeken** in het lab. *Mara knows that the professor [her/proudly] the animal species saw **investigate** in the lab.*
	25. bare_verb_type_2	Deze tekst [laat/ziet] zichzelf **moeilijk** interpreteren. *This text [lets/sees] itself **difficultly** interpretate.*
	26. bare_verb_type_3	De politieagent [zag/keek] de inbreker **ontsnappen** uit het raam. *The police offices [saw/looked] the burglar **escape** from the window.*
	27. te_transparant_split	Jullie weten dat Gwen [**weer zwanger schijnt/schijnt weer zwanger**] te zijn. *You know that Gwen [**again pregnant seems/seems again pregnant**] to be.*
	28. om+te	Oscar zag dat jouw neef [**besloot om te stretchen/om te stretchen besloot**] voor het sporten. *Oscar saw that your cousin [**decided to stretch/to stretch decided**] before the exercise.*
	29. te_om+te_difference_1	De vriend [beweert/belooft] het tijdschrift **al** te lezen. *The friend [claims/promises] the magazine **already** to read.*
	30. te_om+te_difference_2	De vriend [beweert/belooft] het tijdschrift te **willen** te lezen. *The friend [claims/promises] the magazine to **want** to read.*
7	31. complementizer	Teun zegt [dat/∅] zijn schoonbroer ziek **is** van het eten. *Teun says [that/∅] his brother-in-law ill **is** from the food.*
	32. perception_dat	De dokter [keek/zag] voorzichtig **dat** het oor dichtzat. *The doctor [looked/saw] carefully **that** the ear was blocked.*
	33. perception_of	De dokter [zag/keek] voorzichtig **of** het oor dichtzat. *The doctor [saw/looked] carefully **if** the ear was blocked.*
	34. position	Sem heeft zijn moeder [verteld/gisteren] dat **hij** de toets heeft gehaald [∅/verteld]. Sem has his mother [told/yesterday] that **he** the test has passed [∅/told].
	35. sluicing_1	Het meisje heeft iets gekocht en jij [weet/beweert] volgens mij **wat** door ons gesprek. *The girl has something bought and you [know/claim] I think **what** by our conversation.*
	36. sluicing_2	Het meisje heeft iets gekocht en ik denk dat jij [**weet wat/wat weet**] over een paar dagen. *The girl has something bought and I think that you [**know what/what know**] in a few days.*
8	37. perfect	De vrouw is vanmorgen [**gaan zwemmen/zwemmen gaan**] in de woeste zee. *The woman is this morning [**gone swimming/swimming gone**] in the wild sea.*
	38. semi_aspectual_1	Jij ziet dat de bokser [**klappen staat/staat klappe**n] te incasseren. *You see that the boxer [**punches stands/stands punches**] to take.*
	39. semi_aspectual_2	Willem heeft de hele dag [**zitten kletsen/kletsen zitten**] op het werk. *Willem has the entire day [**sit chatting/chatting sit**] at work.*
	40. order_1	Ik denk dat jij ziek [**gaat worden/worden gaat**] op de boot. *I think that you sick [**will become/become will**] on the boat.*
	41. order_2	Jij ziet dat de bokser klappen [**staat te incasseren/te incasseren staat**] tijdens het gevecht. You see that the boxer punches [**stands to take/to take stands**] during the fight.
9	42. position_type	Emir dacht dat de man [**waarschijnlijk langzaam/langzaam waarschijnlijk**] zou zwemmen. *Emir thought that the man [**probably slowly/slowly probably**] would swim.*
	43. position_proform	De gids woont [**er waarschijnlijk/waarschijnlijk er**] al jaren. *The guide lives [**there probably/probably there**] for years.*
10	44. order_main	Vorig jaar [**heeft Roos/Roos heeft**] veel ruimtes gedecoreerd voor grote bruiloften. *Last year [**has Roos/Roos has**] many rooms decorated for big weddings.*
	45. order_embedded	Stef zegt dat de politieagent [**de inbreker heeft/heeft de inbreker**] geïdentificeerd in het publiek. *Stef says that the police officer [**the burglar has/has the burglar**] identified in the audience.*

Table 5
Continued.

11	46. question_formation	Wie [**heeft wat/wat heeft**] gezegd tijdens de bijeenkomst? *Who [**has what/what has**] said during the meeting?*
	47. filler_effect_gap	Ik weet [wat/dat] jij denkt dat de bakker **maakt** in de bakkerij. *I know [what/that] you think that the baker **makes** in the bakery.*
	48. filler_effect_no_gap	Ik weet [dat/wat] jij denkt dat de bakker **koekjes** maakt in de bakkerij. *I know [that/what] you think that the baker **cookies** makes in the bakery.*
	49. hierarchy	Het feit dat mijn broer zei wie zijn vriend [∅ /mijn oom] had **verrast** op het feest verraste mijn dochter. *The fact that my brother said who his friend [∅ /my uncle] had **surprised** at the party surprised my daughter.*
	50. stranding_1	Na de oefeningen vraagt de coach [waar/wat] zijn team **voor** traint. After the exercises the coach asks [where/what] his team **for** trains.
	51. stranding_2	De agent ziet [waar/wat] het kind **omheen** fietst op de weg. *The officer sees [where/what] the child **around** bikes on the road.*
12	52. bridge_verb_1	Wat [wil/weet] de docent dat de leerling **leert**? *What [wants/knows] the teacher that the student **learns**?*
	53. bridge_verb_2	Wat [zegt/roept] de docent dat de leerling }textbfmoet leren? *What [says/calls] the teacher that the student **must** learn?*
	54. island_1	Wat [denk jij dat/vraag jij of] de bakker **maakt** in de bakkerij? *What [think you that/ask you if] the baker **makes** in the bakery?*
	55. island_2	Wat is Sophie teleurgesteld [dat/als] haar vader **heeft** gekocht? *What is Sophie disappointed [that/if] her father **has** bought?*
	56. resumptive_prolepsis	[Van welke zaden/Wat] vraag jij je af of de tuinman ze **heeft** gezaaid in de tuin? *[Of which seeds/What] do you wonder if the gardener them **has** planted in the garden?*
	57. superiority	De klant vraagt [**wie wat/wat wie**] gebakken heeft in de bakkerij. *The customer asks [**who what/what who**] baked has in the bakery.*
13	58. pied_piping	De jongen [**naar wie je/wie je naar**] kijkt is mijn broer. *The boy [**at whom you/whom you at**] look is my brother.*
	59. island	Dit zijn de koekjes die jij [begreep dat/vroeg of] de bakker **had** gemaakt in de bakkerij. *These are the cookies that you [understood that/asked if] the baker **had** made in the bakery.*
	60. resumptive_prolepsis	Dit is het cadeau [waarvan/dat] Sophie teleurgesteld is als haar vader **het** heeft gekocht. *This is the gift [of which/that] Sophie disappointed is if her father **it** has bought.*
14	61. question_similarity_1	Fenna onthulde [welke/die] prijs **haar broer** had gewonnen. *Fenna revealed [what/that] prize **her brother** had won.*
	62. question_similarity_2	Deze gerechten [**heeft de kok/de kok heeft**] gemaakt voor het menu. *These dishes [**has the cook/the cook has**] made for the menu.*
	63. island	Deze planten [begreep jij dat/vroeg jij of] de tuinman **had** gezaaid in de tuin *These plants [understood you that/asked you if] the gardener **had** sown in the garden.*
	64. resumptive_prolepsis	[Van/∅] deze planten [begreep jij dat/vroeg jij of] de tuinman **ze** had gezaaid in de tuin. *[Of/∅] these plants [understood you that/asked you if] the gardener **them** had sown in the garden.*
15	65. structure_type_1	Welke boeken heeft Joris zonder [ze/∅] echt goed te **bekijken** opgeborgen? *Which books has Joris without [them/∅] really good to **inspect** stored?*
	66. structure_type_2	Deze boeken heeft Joris zonder [ze/∅] echt goed te **bekijken** opgeborgen. *These books has Joris without [them/∅] really good to **inspect** stored.*
	67. structure_type_3	De boeken die Joris zonder [ze/∅] echt goed te **bekijken** heeft opgeborgen blijken toch nuttig te zijn. *The books that Joris without [them/∅] really good to **inspect** has stored appear after all useful to be.*
	68. scrambling	Joris heeft die boeken zonder [ze/∅] echt goed te **bekijken** opgeborgen. *Joris has those books without [them/∅] really good to **inspect** stored.*
16	69. adverbial	De monteur weet dat je daar de auto [**niet mee/mee niet**] kan repareren. *The mechanic knows that you there the car [**not with/with not**] can repare.*
	70. weak_proform	[Daar/Er] heeft **Gijs** lange tijd gewerkt. *[There/There] has **Gijs** a long time worked.*
17	71. type_inf_1	Merel hoorde van haar zoon dat [het/een] dagelijks **zeilen** erg leuk is. *Merel heard from her son that [the/a] daily **sailing** very fun is.*
	72. type_inf_2	Tegenwoordig is het [**jagen op dieren/op dieren jagen**] niet meer zo populair These days is the [**hunting on animals/on animals hunting**] not anymore so popular.
18	73. negative_polarity	Ik zou [geen/een] moment **ook maar ergens** met hem willen praten. *I would [no/a] moment **anywhere** with him want to talk.*
	74. geen_expletive	Jouw moeder dacht dat [er/∅] gisteren **geen melk** in de koelkast stond. *Your mother thought that [there/∅] yesterday **no milk** in the refrigerator stood.*
	75. geen_scrambling_1	Er stond [**gisteren geen melk/geen melk gisteren**] in de koelkast. *There stood [**yesterday no milk/no milk yesterday**] in the fridge.*
	76. geen_scrambling_2	Isabelle beheerst [**waarschijnlijk geen danspassen/geen danspassen waarschijnlijk**] tijdens de dansles. *Isabelle masters [**probably no dance steps/no dance steps probably**] during the dancing lesson.*
19	77. univeral_difference_agreement_singular	[Iedere student moet/alle studenten moeten] zijn **opdracht** afmaken. *[Every student has to/All students have to] his **assignment** finish.*
	78. univeral_difference_agreement_plural	[Alle studenten moeten/Iedere student moet] hun **opdracht** afmaken. *[All students have to/Every student has to] their **assignment** finish.*
20	79. argument_scrambling	Ik weet dat Gabriël [**de jacht op zwijnen/op zwijnen de jacht**] verafschuwt. *I know that Gabriël [**the hunt for boars/for boars the hunt**] despises.*
	80. argument_R-extraction	De student vindt [**waarschijnlijk ergens/ergens waarschijnlijk**] anders een leuke plek. *The student finds [**probably somewhere/somewhere probably**] else a fun spot.*
21	81. cross_dependency	Oscar heeft de athleet de marathon [**zien lopen/lopen zien**] afgelopen weekend. *Oscar has the athlete the marathon [**seen walk/walk seen**] last weekend.*
22	82. argument_nominal	De bewaker hoorde dat er [iemand/daarstraks] om hulp **riep** [0/iemand] in het gebouw. *The guard heard that there [someone/earlier] for help **called** [0/someone] in the building.*
	83. adjectival_adverbial	Jack wist dat de vriendin de handleiding [**erg zorgvuldig las/las erg zorgvuldig**] tijdens het klussen. *Jack knew that the friend the manual [**very carefully read/read very carefully**] during the DIYing.*
	84. adjectival_supplementive	Jij zag dat Evi [**tevreden wegliep/wegliep tevreden**] na de bijeenkomst. *You saw that Evi [**satisfied walked away/walked away satisfied**] after the meeting.*

8.2 Intraclass Correlation Coefficient (ICC) Analysis

For paradigms 35, 36, 55, 59, 63, 65, 66, 67, and 68 the ICC was below 0.5 in one or more experiment versions, indicating poor reliability (Perinetti 2018). These paradigms belong to the phenomena Finite Argument Clause, Wh-Movement Restrictions,

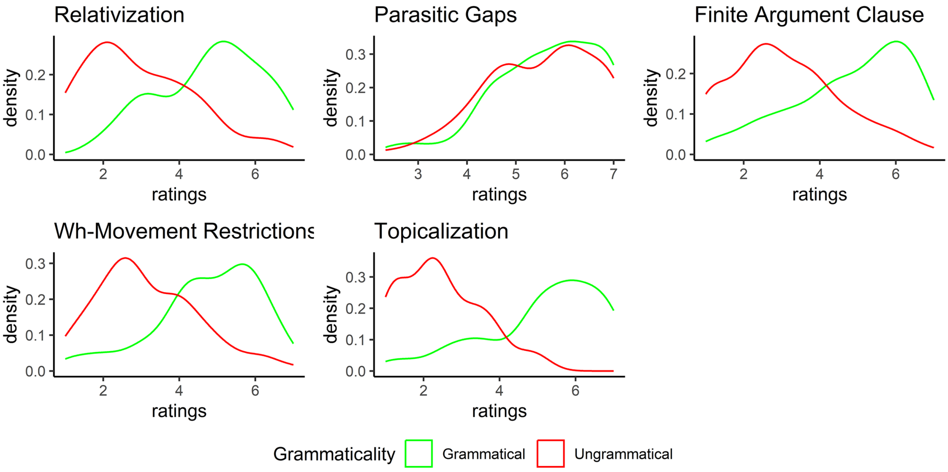


Figure 6
Distribution of acceptability ratings for all participants for the phenomena Finite Argument Clause, Wh-Movement Restrictions, Relativization, Topicalization, and Parasitic Gaps.

Relativization, Topicalization, and Parasitic Gaps. Subsequently, we examined the distribution of the participant ratings for these phenomena, which are illustrated in Figure 6. The distributions' (near) unimodality indicates that the mean rating per paradigm and per condition is a meaningful measure, in spite of the relatively low ICC.

8.3 Generating Data with ChatGPT

8.3.1 Generation Procedure. As discussed in the main text, we used GPT-3.5 Turbo to generate the minimal pairs in BLiMP-NL large, but not in BLiMP-NL small. All these minimal pairs were checked manually by one person each, to ascertain that they correctly followed the relevant paradigm and did not introduce any errors apart from the paradigm to test. If the minimal pairs did contain additional mistakes, these were usually corrected if the minimal pair followed the desired paradigm and was unique enough otherwise. The quality assurance process for accepting the generated examples consisted of checking for grammaticality, spelling, uniqueness, and whether the sentence adhered to the paradigm. We checked our dataset for duplicates and used a spell checker to improve the quality, but some (human or ChatGPT) errors could still be present. In total, a group of eight people worked on this, although most minimal pairs were checked by only one.

To generate the minimal pairs, we used the Web interface of ChatGPT, accessing the GPT-3.5 Turbo model, with all exportation of data turned off. We used variations of prompts, but most of them were in the following form:

> "Kan je het patroon in de volgende zinnen herkennen en nog 30 van dit soort voorbeelden bedenken?" / *"Can you recognize the pattern in the following sentence en think of 30 more of these types of examples?"*

The prompts were then followed by the paradigm's (hand-crafted) BLiMP-NL small sentences, preceded by the word "Correct:", or "Incorrect:", indicating the sentence's grammaticality. Variations of the prompt could include asking specifically to alternate correct and incorrect examples or the number of examples being asked. We ended up asking for 30 examples at a time since more than that often lead to repetition of examples. After receiving the examples, we would use a follow-up prompt in the following form:

> "Dankjewel! Kan je nog 30 voorbeelden bedenken?" / *"Thank you! Can you think of 30 more examples?"*

These follow-up questions sometimes included specific requests already, such as using more variation or using less of a specific word if one was repeated a lot. Sometimes, we also asked GPT-3.5 Turbo to make the examples on a specific subject if we thought the examples did not vary enough or if many examples were repeated. In a few cases, we also provided a list of words to use in sentences if there was too little variation. For example, in paradigm 6, which concerned ditransitive verbs, the same verbs that were already in BLiMP-NL small were often used so we provided a list of common Dutch ditransitive verbs to use in the sentences. This was only necessary for the few paradigms that concerned these kinds of very specific word types.

This process was repeated until we collected 90 additional examples per paradigm. All generated examples followed the exact same grammatical pattern that was present in the handwritten examples.

8.3.2 Difference in Model Performance on Hand-crafted and ChatGPT-generated Pairs. The fact that 90 sentence pairs in each paradigm were created with help of GPT-3.5 Turbo raises the question whether it is fair to evaluate similar LMs on these same pairs. Figure 7 shows the difference in accuracy on hand-crafted versus ChatGPT-generated pairs, across all models and paradigms. It is clear that these differences cluster around 0, that is, ChatGPT-generated items and hand-crafted items were equally difficult. Over

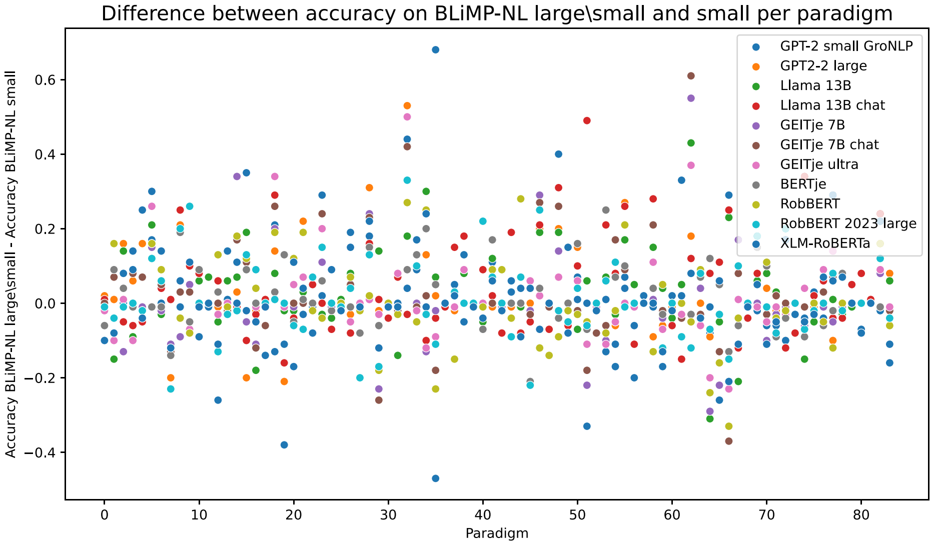


Figure 7
Comparison between model accuracies on hand-crafted and ChatGPT-generated pairs, for all models tested in the main paper and all paradigms. BLiMP-NL large\small refers to ChatGPT-generated pairs, i.e., the BLiMP-NL large dataset without sentences also present in BLiMP-NL small.

all minimal pairs, however, all models were slightly more accurate on the ChatGPT-generated than hand-crafted sentences. This difference was the highest for LLaMA-chat with 3.18 percentage points and was below 2 percentage points for all other models, but a z-test for two proportions revealed that this difference was far from significant for all the models (all $z < 0.6$, all $p > 0.5$).

8.4 Alternative Evaluations

8.4.1 Evaluation using SLOR. Table 6 shows the performance of the evaluated causal models when using the SLOR score to identify the grammatical sentence of a pair.

8.4.2 Comparison to Human Ratings Using the Method of Hu et al. (2024). Figure 8 shows the correlation between GPT-2 SLOR scores and human acceptability ratings following the comparison method from Hu et al. (2024). We indeed see that this method yields a far higher correlation than any of the other methods from Section 6.2. We believe the correlation is artificially inflated by comparing a SLOR score difference to absolute ratings.

8.4.3 Evaluation on Critical Word(s). We perform a similar kind of experiment as in Section 6.1 but looking specifically at the probability of the critical word(s), where we

Table 6
The accuracy of 8 causal language models on BLiMP-NL large using the fraction of grammatical sentences that received a higher SLOR score than their ungrammatical counterparts. Random guessing would result in an accuracy of 0.5.

Phenomenon	GPT-2 small GroNLP	GPT-2 large	LLaMA 13B	LLaMA 13B chat	GEITje 7B	GEITje 7B chat	GEITje ultra DPO	Phenomenon Average
Anaphor agreement	0.87	**0.93**	0.61	0.34	0.77	0.84	0.73	0.72
Binding (Principle A)	**0.97**	**0.97**	0.68	0.48	0.70	0.90	0.67	0.77
Argument Structure	0.81	**0.91**	0.76	0.81	0.87	0.87	0.89	0.84
Complementive	0.81	**0.94**	0.78	0.74	0.88	0.85	0.88	0.84
Passive	0.73	**0.89**	0.75	0.64	0.78	0.79	0.81	0.77
Infinitival Argument Clause	0.84	**0.93**	0.78	0.66	0.91	0.90	0.91	0.85
Finite Argument Clause	**0.90**	0.81	0.76	0.71	0.85	0.84	0.81	0.81
Auxiliaries	0.95	**1.00**	0.96	0.94	0.99	**1.00**	**1.00**	0.98
Adverbial Modification	0.96	1.00	0.96	0.92	0.99	**1.00**	0.98	0.97
Verb Second	0.94	1.00	1.00	0.91	**1.00**	**1.00**	**1.00**	0.98
Wh-Movement	0.61	**0.87**	0.79	0.70	0.84	0.85	**0.87**	0.79
Wh-Movement Restrictions	0.68	0.80	0.74	0.62	0.82	0.84	**0.87**	0.77
Relativization	0.56	0.74	0.79	0.82	0.80	0.83	**0.85**	0.77
Topicalization	0.77	**0.91**	0.81	0.72	0.73	0.70	0.75	0.77
Parasitic Gaps	0.79	**0.86**	0.61	0.51	0.60	0.65	0.61	0.66
R-Words	0.99	1.00	0.84	0.90	**1.00**	**1.00**	0.99	0.96
Nominalization	0.89	0.96	0.96	0.94	0.93	0.95	**0.97**	0.94
Determiners	0.85	0.98	0.80	0.69	0.99	**1.00**	0.96	0.89
Quantifiers	0.78	**0.83**	0.52	0.80	0.76	0.70	0.71	0.73
Adpositional Phrases	**1.00**	1.00	0.98	0.96	1.00	**1.00**	**1.00**	0.99
Crossing Dependencies	0.93	**1.00**	**1.00**	0.98	**1.00**	**1.00**	**1.00**	0.99
Extraposition	0.96	**0.99**	0.98	0.87	**0.99**	**0.99**	**0.99**	0.97
Model Average	0.82	**0.91**	0.80	0.73	0.86	0.87	0.87	0.84

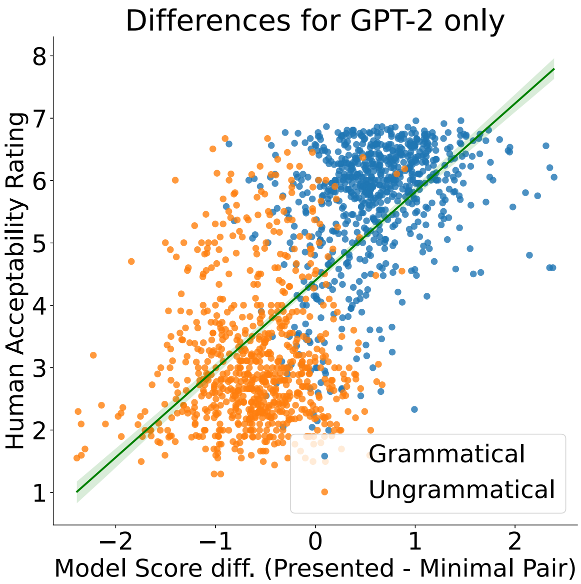


Figure 8
Comparison between model scores and human ratings using method of Hu et al. (2024). The x-axis represents the difference in SLOR between the sentence presented to humans versus its counterpart in the minimal pair. Each point represents a single sentence. The Pearson correlation between human rating and model scores is $\rho = 0.68$, with $p < .0001$. A linear regression line is calculated for the data points; the shaded area represents the 95% confidence interval.

expect the probability to differ between the sentences of a pair since that is where the sentence becomes ungrammatical.

When we evaluate models on BLiMP-NL (see Table 7), we do see a small difference: For most models, evaluating on the critical word yields higher scores than evaluating on the entire sentence. Overall the differences are slight, but this varies over phenomena. For example, models receive much higher scores on average for the Anaphor Agreement phenomenon.

8.4.4 Prompting. We perform two prompting experiments; one where the models have to decide which of two sentences is grammatical and another where the models have to rate individual sentences' acceptability, similarly to the human raters' task. Only the causal language models can be evaluated with prompting and thus we do not include the masked language models in these results. Furthermore, some of the causal language models have a dedicated chat model. We test these models using their chat function, while for the other causal language models we tried using the generating function of the model. This did not work, however, since the language models did not seem to understand the task in this setting and thus did not give a possible answer. We will therefore only show the results for models with a dedicated chat template.

For the first experiment where the language model has to choose which of two sentences is grammatical, the language model gets a prompt in the following format:

> "Hieronder zie je twee zinnen die op elkaar lijken waarvan 1 grammaticaal correct is en de ander niet. Welke van de twee zinnen is grammaticaal correct? Zin 1: Ik zie mezelf in de spiegel. Zin 2: Wij zien mezelf in de spiegel."

Table 7
The accuracy of 8 causal language models on BLiMP-NL large using the fraction of grammatical sentences for which the probability of critical word(s) is higher than for the ungrammatical counterpart. A random guessing baseline would receive an accuracy of 0.5.

Phenomenon	GPT-2 small GroNLP	GPT-2 large	LLaMA 13B	LLaMA 13B chat	GEITje 7B	GEITje 7B chat	GEITje ultra DPO	Phenomenon Average
Anaphor agreement	**1.00**	1.00	1.00	0.81	1.00	**1.00**	**1.00**	0.97
Binding (Principle A)	**1.00**	0.97	**1.00**	0.67	1.00	0.98	1.00	0.94
Argument Structure	0.86	0.95	0.90	0.84	**0.96**	**0.96**	0.94	0.92
Complementive	0.80	**0.96**	0.91	0.88	0.95	**0.96**	**0.96**	0.92
Passive	0.68	0.76	0.78	0.72	0.79	0.78	**0.80**	0.76
Infinitival Argument Clause	0.88	0.95	0.87	0.78	0.94	**0.96**	0.92	0.90
Finite Argument Clause	0.90	**0.93**	0.90	0.82	0.91	0.89	0.86	0.89
Auxiliaries	0.96	0.98	0.97	0.94	**0.99**	**0.99**	**0.99**	0.98
Adverbial Modification	0.95	**0.98**	0.95	0.89	0.95	0.94	0.92	0.94
Verb Second	0.98	1.00	**1.00**	0.97	**1.00**	**1.00**	**1.00**	0.99
Wh-Movement	0.70	0.89	0.83	0.73	0.87	0.88	**0.90**	0.83
Wh-Movement Restrictions	0.61	**0.81**	0.78	0.73	0.77	0.78	0.79	0.75
Relativization	0.70	**0.81**	0.74	0.75	**0.81**	0.81	0.80	0.77
Topicalization	0.65	0.86	0.71	0.70	**0.90**	0.81	0.83	0.78
Parasitic Gaps	0.66	**0.92**	0.86	0.88	0.88	0.91	0.87	0.85
R-Words	0.98	**1.00**	**1.00**	**1.00**	1.00	**1.00**	1.00	0.99
Nominalization	0.70	0.91	0.86	0.90	0.91	0.88	**0.93**	0.87
Determiners	0.94	**0.98**	0.94	0.81	0.96	**0.98**	0.93	0.94
Quantifiers	0.82	0.78	0.74	**0.88**	0.76	0.79	0.76	0.79
Adpositional Phrases	**1.00**	**1.00**	0.99	0.97	**1.00**	**1.00**	**1.00**	0.99
Crossing Dependencies	0.92	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99
Extraposition	0.96	**0.99**	0.97	0.87	0.98	0.98	**0.99**	0.96
Model Average	0.82	**0.92**	0.88	0.82	0.91	0.91	0.91	0.88

> *"Below you can see two similar sentences of which one is gramatically correct and the other is not. Which of the two sentences is grammatically correct? Sentence 1: I see myself in the mirror. Sentence 2: We see myself in the mirror."*

The language model then had to respond either "Zin 1" ("Sentence 1") or "Zin 2" ("Sentence 2"). The order of the acceptable and unacceptable sentences was shuffled between examples to ensure we were not measuring a bias for Sentence 1 or 2.

The results in Table 8 show that the basic GEITje chat model performed best, while still performing very poorly. It is in line with other research that this Mistral-based model performs better than the LLaMA-based model. It is surprising, however, that GEITje ultra with DPO does not perform better and falls so far behind. This can, of course, be due to the specific prompt used, as slight changes in prompts can drastically alter results.

In the second experiment, where the language model has to rate the acceptability of individual sentences, it receives a prompt in the following format:

> "Welk cijfer zou je deze zin geven op een schaal van 1 ('Erg slecht') tot 7 ('Erg goed') als je de grammaticaliteit van de zin moet beoordelen: Wij zien mezelf in de spiegel."

Table 8
Accuracies when asking LLMs which of two sentences is grammatically correct. Highest accuracy printed in **bold**.

Model	Accuracy
LLaMA 13B chat	0.6071
GEITje 7B chat	**0.6821**
GEITje ultra SFT	0.4810
GEITje ultra DPO	0.5321

> *"What score would you give the following sentence on a scale from 1 ('Very bad') to 7 ('Very good') if you have to judge the grammaticality of the sentence: We see myself in the mirror."*

However, the models were unable to perform this task and mostly gave the same score to both sentences of the pair. They mostly assigned sentences a score of 4 or 7.

Acknowledgments

The research presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium. We thank Anna Langedijk, Bram Fresen, Jan Athmer, Sabijn Perdijk, Sven Terpstra, and Venja Beck for help in generating and validating sentence pairs.

References

Aarts, Bas. 2007. *Syntactic Gradience: The Nature of Grammatical Indeterminacy*. Oxford University Press. https://doi.org/10.1093/oso/9780199219261.001.0001

Arora, Aryaman, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663. https://doi.org/10.18653/v1/2024.acl-long.785

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2005. Predicting the dative alternation. In *Cognitive Foundations of Interpretation*. Royal Netherlands Academy of Science, pages 69–94.

Broekhuis, Hans. 2013. *Syntax of Dutch. Adpositions and Adpositional Phrases*. Amsterdam University Press. https://doi.org/10.1017/9789048522255

Broekhuis, Hans, Norbert Corver, and Riet Vos. 2015. *Syntax of Dutch. Verbs and Verb Phrases*. Amsterdam University Press. https://doi.org/10.1515/9789048524839

Broekhuis, Hans and Evelien Keizer. 2012. *Syntax of Dutch. Nouns and Noun Phrases*. Amsterdam University Press. https://doi.org/10.1017/9789048517602

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press. https://doi.org/10.21236/AD0616323

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Covert, Ian, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

Delobelle, P. and F. Remy. 2024. RobBERT-2023: Keeping Dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 14:193–203.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt. 2020. RobBERT: A Dutch RoBERTa-based language model. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

Dentella, Vittoria, 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. https://doi.org/10.1073/pnas.2309583120, PubMed: 38091290

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

de Vries, Wietse and Malvina Nissim. 2021. As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846. https://doi.org/10.18653/v1/2021.findings-acl.74

de Vries, Wietse, Martijn Wieling, and Malvina Nissim. 2023. DUMB: A Dutch model benchmark. In the *2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.447

Frank, Stefan L. and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58:641–657. https://doi.org/10.1007/s10579-023-09684-x

Gauthier, Jon, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76. https://doi.org/10.18653/v1/2020.acl-demos.10

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hilbig, Benjamin E. 2015. Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, 48:1718–1724. https://doi.org/10.3758/s13428-015-0678-9, PubMed: 26542972

Holtzman, Ari, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051. https://doi.org/10.18653/v1/2021.emnlp-main.564

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. https://doi.org/10.18653/v1/2020.acl-main.158

Hu, Jennifer 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. https://doi.org/10.18653/v1/2023.emnlp-main.306

Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121. https://doi.org/10.1073/pnas.2400917121, PubMed: 39186652

Jentoft, Matias and David Samuel. 2023. NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jumelet, Jaap and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational*

Linguistics: ACL 2023, pages 8697–8712. https://doi.org/10.18653/v1/2023.findings-acl.554

Kauf, Carina and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935. https://doi.org/10.18653/v1/2023.acl-short.80

Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. 2009. Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):1–12. https://doi.org/10.3758/BRM.41.1.12, PubMed: 19182118

Kruijsbergen, Joni, Serafina Van Geertruyen, Véronique Hoste, and Orphée De Clercq. 2024. Exploring LLMs' capabilities for error detection in Dutch L1 and L2 writing products. *Computational Linguistics in the Netherlands Journal*, 13:173–191.

Kulmizev, Artur and Joakim Nivre. 2022. Schrödinger's tree—On syntax and neural language models. *Frontiers in Artificial Intelligence*, 5:796788. https://doi.org/10.3389/frai.2022.796788, PubMed: 36325030

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241. https://doi.org/10.1111/cogs.12414, PubMed: 27732744

Laurinavichyute, Anna and Titus Von der Malsburg. 2024. Agreement attraction in grammatical sentences and the role of the task. *Journal of Memory and Language*, 137:1–16. https://doi.org/10.1016/j.jml.2024.104525

Leong, Wei Qi, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models. *arXiv preprint: arXiv:2309.06085*.

Linzen, Tal and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Yikang, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. ZhoBLiMP: A systematic assessment of language models with linguistic minimal pairs in Chinese. *arXiv preprint: arXiv:2411.06096*.

Lu, Jiayi, Jonathan Merchan, Lian Wang, and Judith Degen. 2024. Can syntactic log-odds ratio predict acceptability and satiation? In *Proceedings of the Society for Computation in Linguistics 2024*, pages 10–19.

Marvin, Rebecca and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. https://doi.org/10.18653/v1/D18-1151

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121. https://doi.org/10.1073/pnas.2322420121, PubMed: 39365822

Murphy, Victoria A. 1997. The effect of modality on a grammaticality judgement task. *Second Language Research*, 13:34–65. https://doi.org/10.1191/026765897671676818

Oh, Byung Doh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350. https://doi.org/10.1162/tacl_a_00548

Pauls, Adam and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 959–968.

Perinetti, Giuseppe. 2018. StaTips Part IV: Selection, interpretation and reporting of the intraclass correlation coefficient. *South European Journal of Orthodontics and Dentofacial Research*, 5:3–5. https://doi.org/10.5937/sejodr5-17434

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. https://doi.org/10.18653/v1/2020.acl-main.240

Schütze, Carson T. 2016. *The Empirical Base of Linguistics*. Language Science Press.

Semmelmann, Kilian and Sarah Weigelt. 2016. Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, 49:1241–1260. https://doi.org/10.3758/s13428-016-0783-4, PubMed: 27496171

Someya, Taiga and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594. https://doi.org/10.18653/v1/2023.findings-eacl.117

Song, Yixiao, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. Sling: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634. https://doi.org/10.18653/v1/2022.emnlp-main.305

Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(11):1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002

Taktasheva, Ekaterina, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299. https://doi.org/10.18653/v1/2024.emnlp-main.522

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vanroy, Bram. 2023. Language resources for Dutch large language modelling. *arXiv preprint arXiv:2312.12852*.

van Schijndel, Marten, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837. https://doi.org/10.18653/v1/D19-1592

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Volodina, Elena, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – A dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37.

Warstadt, Alex, 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. https://doi.org/10.1162/tacl_a_00321

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641. https://doi.org/10.1162/tacl_a_00290

Wilcox, Ethan, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3302–3312. https://doi.org/10.18653/v1/N19-1334

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212. https://doi

.org/10.1080/01621459.1927.10502953

Xiang, Beilei, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790. https://doi.org/10.1175/JCLI-D-20-0301.1

Yi, Eunkyung and Sang-Hee Park. 2024. Spoken acceptability judgment, reaction time and a comparison with written judgment. *Journal of Cognitive Science*, 24:437–464.

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706.