Evaluating the Cross-Lingual Syntactic Capabilities of Language Models

Suchir Salhan

27 October 2025

In this lecture, we focus on how syntactic theory is applied to evaluate Language Models, and how syntactic typology can be leveraged to develop evaluation metrics for language models in a cross-lingual setting. [1]

Grammars, Multilinguality and Interpretability

Human grammars exhibit systematicity and compositionality: speakers can generate and interpret an unbounded number of novel sentences by applying abstract rules to smaller units. Syntax provides a theory of these rules, explaining how hierarchical representations govern phenomena such as agreement, argument structure, and movement.

From a machine learning perspective, syntax provides a principled account of the structure and generalisation behaviours that human language exhibits: systematicity, compositionality, and the ability to generate and interpret novel utterances from finite data. These are precisely the properties we want LLMs to master. Studying syntax equips us with targeted evaluation methods (such as minimal pairs) that isolate whether models have learned causal grammatical rules rather than shallow statistical correlations. It also offers a framework for understanding cross-lingual generalisation, since grammatical universals and typological variation define the space of solutions any language learner—human or machine—must discover. In short, syntax is not just linguistic theory: it is a rigorous tool for diagnosing what models truly know about language, and for guiding the development of systems that learn robust, interpretable, human-like representations.

Grammaticality judgments therefore form a natural basis for *evaluating the linguistic competence of Large Language Models (LLMs)*. Minimal pair evaluation—where two closely matched sentences differ only in one targeted syntactic property—tests whether a model has internalised some linguistic generalisations rather than relying on surface-level memorisation, but do not provide any insight as to how models learn these capabilities.

Crucially, human language varies systematically across the world's languages. Crosslinguistic differences in word order, case mark-

[1] This Handout is associated with MPhil Advanced Computer Science /Part III Computer Science Tripos Module – **L95 Introduction to Natural Language Syntax and Parsing**. These notes are copyright Suchir Salhan, unless otherwise indicated.

ing,agreement systems, and morphological complexity reveal both the diversity and the *universals* of grammatical architecture. Syntactic typology models how languages cluster into families of possible grammars.[2] As massively multilingual LLMs become an increasingly dominant paradigm for developing language technology for most languages beyond English and communities, applying typological knowledge allows us to evaluate whether models merely mimic patterns from training data or discover deeper universal properties of linguistic structure. Minimal pair datasets beyond English enable us to test whether universal constraints transfer across languages while respecting language-specific variation. This is not only theoretically interesting, but also practically important to interpret and understand the limitations of standard cross-lingual transfer or multilingual pretraining techniques.

Mechanistic interpretability strengthens this connection by probing how grammatical knowledge is *represented* inside models. An emerging research question of interest to several NLP researchers is whether independently trained monolingual LLMs converge on shared latent features corresponding to shared linguistic representations or have common linguistically-interpretable subspaces. By aligning these across languages using activation correlations, (mechanistic) interpretability methods can be applied to identify universal syntactic features that reliably emerge regardless of the specific language input.

Despite limitations (e.g., simplified contexts and binary contrasts), minimal pair evaluation remains one of the clearest ways to connect model behaviour to linguistic theory. Combined with mechanistic approaches that reveal internal structure, syntactic knowledge enables us to assess whether LLMs learn the kinds of systematic, compositional, and typologically grounded generalisations characteristic of human grammar.

Evaluating Grammaticality in Language Models

Language Models can be evaluated intrinsically on datasets that assess different linguistic capabilities or extrinsically on real-world tasks or applications. To assess a model's syntactic competence, we require an appropriate *metric* which can be used to meaningfully compare systems. The linguistic capabilities of language models can be intrinsically evaluated using minimal pairs of datasets consisting of pairs of contrasting grammatical and ungrammatical sentences. This is the dominant method for evaluating Natural Language Syn-

[2] In Linguistics, there are two main camps: descriptive typology (describing featural differences using large-scale statistical databases) and **formal generative typology**, which seeks to identify more abstract Parameters of variation. See Ponti et al (2019) for a survey on typological linguistics for NLP: https://aclanthology.org/J19-3005.pdf. For an introduction to Functional Generative Typology (historically associated with the Chomskyan Principles & Parameters approach), you can read Baker (2008): https://sites.rutgers.edu/mark-baker/Functional-Generative-Typology

tax. A common approach to arrive at an overall score of the syntactic capabilities of a Language Model is to macro-average the accuracies across test sets covering various syntactic phenomena. [3] Accuracy calculations differ between **causal/autoregressive language models** (e.g., GPT or LLama), where the chain rule is applied by summing the log-likelihood values for each successive token, and **Masked Language Models (MLM)** (e.g., BERT or RoBERTa).

Sentence pseudo-log-likelihood (PLL) scores are estimated for MLMs by successively masking each sentence token, retrieving its score using the rest of the sentence as context, and summing the resulting values. [4]

$$\mathrm{PLL}(\mathbf{W}) := \sum_{t=1}^{|\mathbf{W}|} \log P_{\mathrm{MLM}}(\mathbf{w}_t \mid \mathbf{W}_{\setminus t}; \Theta).$$

This is based on an interpretation of the MLM objective as a stochastic *maximum pseudolikelihood estimation* (MPLE) on a training set $\mathcal{W}$, which approximates the conventional Maximum Likelihood Estimation (MLE). This is by asymptotically maximising an objective:

$$\mathcal{J}_{\mathrm{PL}}(\Theta; \mathcal{W}) = \frac{1}{|\mathcal{W}|} \sum_{\mathbf{W} \in \mathcal{W}} \mathrm{PLL}(\mathbf{W}; \Theta).$$

, where $\{\mathbf{w}_t\}_{t=1}^{|\mathbf{W}|}$ are random variables in a fully connected graph. In this way, MLMs learn an underlying joint distribution whose conditional distributions $\mathbf{w}_t \mid \mathbf{W}_{\setminus t}$ are modelled by masking at position t. This PLL metric is very popular in LLM work that assesses the effect of training data and model fluency. [5]

However, PLL leads to over-inflated scores for out-of-vocabulary (OOV) tokens that are tokenised into subword tokens, which are predicted using a token's bidirectional context. To address this, a new metric `PLL-word-l2r` places a [MASK] over the current target token (now: s_{w_t}), but also over all future sentence tokens that belong to the same word s_w as the target. As shown in *Figure* , this is an intermediate strategy to compute a PPL score for an OOV token like `souvenir`, which is tokenised as subwords *so ##uven ##ir* instead of whole word masking and the default strategy. Inference is then conditioned on a context that includes all preceding sentence tokens (including those belonging to the current word) and all sentence tokens from future words. [6] The final score of a sentence S is obtained as the sum of the log probabilities of each of the w tokens in each of the S words:

$$\mathrm{PLL}_{\mathrm{l2r}}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}\left(s_{w_t} \mid S \setminus \{s_{w_{t'} \geq t}\}\right) \qquad (1)$$

For both causal and masked language models, probabilities are normalized by sentence length. Recently, an alternative approach has

[3] ***Is Macro-Averaging a meaningful or cognitively plausible way to assess the capabilities of a system?*** Language learners acquire syntactic phenomena concurrently, so macro-averaging may not align with realistic scenarios for evaluating the *development* of linguistic capabilities.

[4] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240

[5] Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.90. URL https://aclanthology.org/2021.acl-long.90

[6] Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada, July 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-short.80. URL https://aclanthology.org/2023.acl-short.80

been to prompt LLMs to **rate item plausibility**, absolutely or on a Likert scale.[7] LLMs perform worse with direct prompting and metalinguistic prompts, which cannot be taken as conclusive evidence that LLM lacks a particular linguistic generalisation [8]. A new evaluation metric called the **Elements of World Knowledge (EWoK)**.[9] uses both traditional plausibility estimates via log probability and two prompt-based strategies called LIKERT and CHOICE. The metric for correctness of a given item is the recovery of the designed item structure such that

$$\text{score}(T_1 \mid C_1) > \text{score}(T_1 \mid C_2)$$

and

$$\text{score}(T_2 \mid C_1) < \text{score}(T_2 \mid C_2),$$

where score reflects P_θ for log probabilities, an integer rating for LIKERT, and the correct context index selection for CHOICE, and T is the target sentence and C is the context of the minimal pair.

Another possibility is evaluating models on the probability that a language model assigns to a **critical word**, which is the word in the sentence where it can become ungrammatical. Dubbed the *"two-prefix method"*, we would expect the language models to give this word in particular a lower probability in the ungrammatical than in the grammatical sentences. As the critical word will be the same for ungrammatical/grammatical sentences and the frequency of the critical word is the same, the only thing that differs is the **preceding context**.

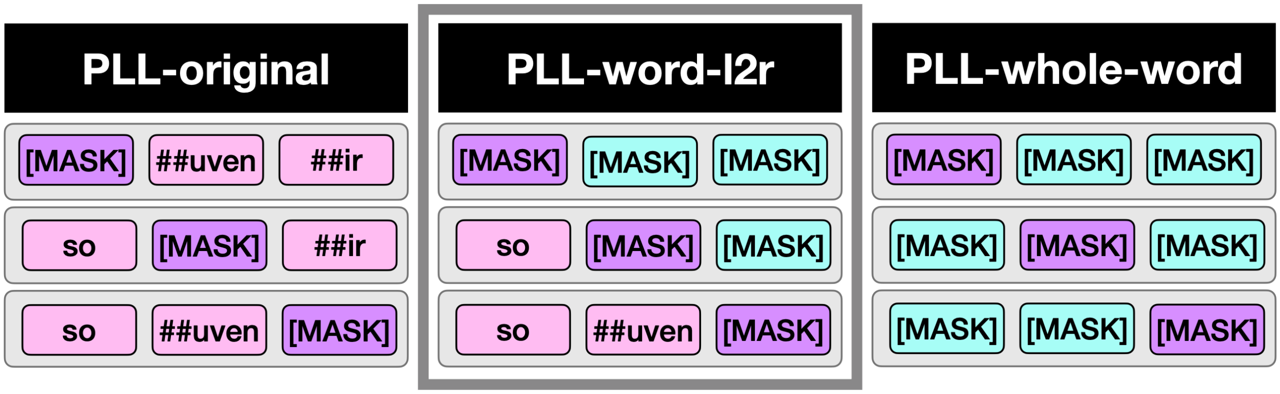


Figure 1: The PLL score of a multi-token OOV items, split into subword tokens, can be computed in different ways. *Purple*: target token, *pink*: within-word tokens that are available during inference, *turquoise*: within-word tokens that are masked during inference. Sentence tokens that do not belong to the current word are always available during inference. *Figure from Kauf & Ivanova (2023)*

Linguistic Limitations:

- Scores can be influenced by superfluous factors, e.g., the number of available synonyms. Therefore, PLL approaches are only useful in highly restricted minimal pair setups

[7] The latter is standard experimental practice in Experimental Syntax. If you are interested about best practices in acceptability experiments, consider reading the introductory chapters in *The Cambridge Handbook of Experimental Syntax*

[8] Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.306. URL `https://aclanthology.org/2023.emnlp-main.306`

[9] Here is some more information about this: `https://ewok-core.github.io/`

- A revised PLL_{l2r} metric may not generalise to agglutinative languages (increased uncertainty due to a high number of tokens per word) and the metric probably is not as important for isolating/analytic languages where word-level items can be represented as single tokens.

BLiMP: Benchmark of Linguistic Minimal Pairs

BLiMP consists of 12 syntactic phenomena in English with unique identifiers (UIDs) of 67 syntactic paradigms. NLP practitioners standardly report BLiMP macro-averages.[10]

- Minimal pairs were **artificially generated** using from abstract grammars that exemplify syntactic phenomena – this easily yields a large number of sentences, which can help control for other possible sources of noise in test materials. Generation scripts use templates to sample lexical items with **selectional restrictions**, which annotate the morphological, syntactic, and semantic features of over 3000 items.
- **Human Evaluation:** Human benchmarking is important in several NLP tasks. It is a useful proxy for the difficulty of different tasks. For BLiMP, the authors used 20 validators who rated five pairs from each of the 67 paradigms for 6,700 judgments.

BLiMP is standardly used for evaluating monolingual English Language Models with other semantic evaluation benchmarks like the **General Language Understanding Evaluation (GLUE) benchmark**. [11] *Table* shows an example of benchmarking in the BabyLM Shared Task, which uses BLiMP alongside GLUE and EWoK.

Model	BLiMP	BLiMP Suppl.	EWoK	GLUE	Av.
BabyLlama	69.8	59.5	50.7	63.3	60.8
LTG-BERT	60.6	60.8	48.9	60.3	57.7

Table 1: Example of Language Model Evaluation from the BabyLM Shared Task 2024

Agreement

BLiMP's Subject-verb agreement dataset has minimal pairs of contrast sentences with correct and incorrect agreement (e.g., *These casseroles* ***disgust/*disgusts*** *Kayla*). Similarly, BLiMP's Determiner-noun agreement dataset consists of minimal pairs about number agreement between demonstrative determiners (e.g. *this/these*) and the associated noun. The determiner-noun agreement and subject-verb agreement phenomena also include paradigms illustrating

[10] BLiMP: The Benchmark of Linguistic Minimal Pairs for English, available at: `https://aclanthology.org/2020.tacl-1.25.pdf`

[11] Here is the GLUE paper: `https://openreview.net/pdf?id=rJ4km2R5t7`

irregular morphology. BLiMP's IRREGULAR FORMS contain irregular English past participles morphology in an adjectival case (*The forgotten newspaper article was bad.* v **The forgot newspaper article was bad.*) and a verbal case (*Edward hid the cats.* v *Edward hidden the cats.*) BLiMP does not evaluate models on non-existent forms like **breaked* because such forms are out of the vocabulary for some LMs.

Filler-Gap Dependencies and Island Effects

FILLER-GAP Dependencies arise from phrasal movement in, e.g., *wh*-questions. BLiMP's dataset contains minimal pairs across *interveners*. These include **subject gaps** (e.g., *Cheryl thought about* ***some dog*** *that upset Sandra* v. **Cheryl thought about who some dog upset Sandra.*) and **object gaps** (e.g., *Joel discovered the vase that Patricia might take.* v **Joel discovered what Patricia might take the vase.*). Filler-gap dependencies can be **long-distance dependencies** with interveners:
Susan won't discover a car that Jane admired that is aggravating a lot of cashiers./Susan won't discover who a car that Jane admired is aggravating a lot of cashiers. **(subject gap)**
*Laurie forgot some alumnus that most organizations that competed have known./ *Laurie forgot who most organizations that competed have known some alumnus.* **(object gap)**

Figure shows that GPT 2 and GPT 3 have a negative wh-effect in the gap condition and a positive wh-effect in thegap condition, which shows that to some degree models are learning the basic filler-gap dependency.

ISLAND EFFECTS characterise restrictions on syntactic environments where the gap in a filler-gap dependency may occur. Descriptively, we can identify several classes of ungrammatical sentences summarised in *Table* , where the strings written in brackets indicate "copies" (or traces, in earlier terminology) of constituents under syntactic theories assuming movement.

- ADJUNCT ISLANDS: Gaps cannot be licensed inside an adjunct clause
- COMPLEX NP ISLANDS: Gaps are not licensed inside S nodes that are dominated by a lexical head noun
- COORDINATION ISLANDS: Gaps cannot occur in only one half of a coordinate structure
- LEFT BRANCH ISLANDS: Modifiers that appear on the "left branch" under an NP cannot be gapped.

Statistical evidence for island effects have been found across Language Models by contrasting the wh-effects in an island condition

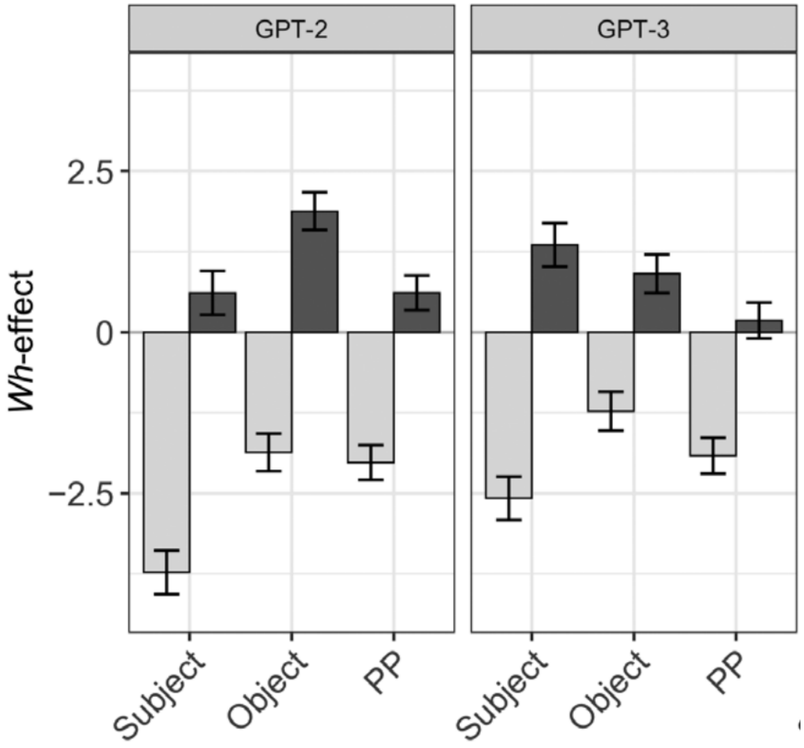
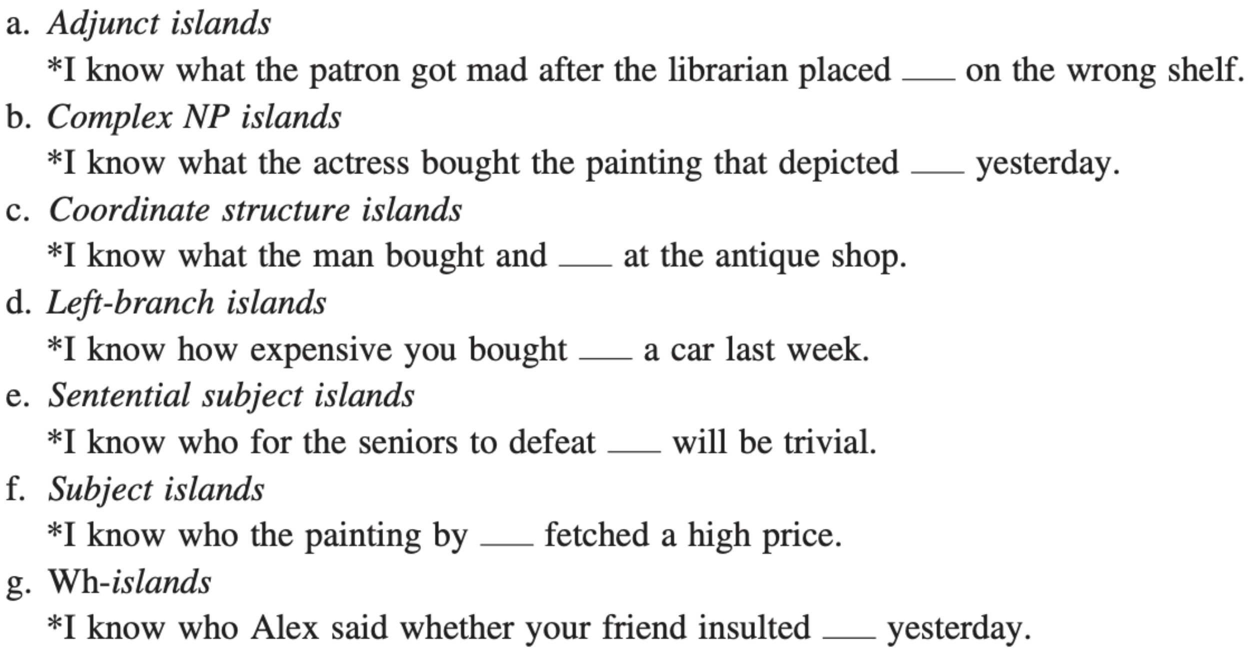


Figure 2: GPT-2 and GPT-3 show sensitivity to island conditions. *Figure from Wilcox, Futrell & Levy (2024) "Using Syntactic Models to Test Syntactic Learnability"*, available from: `https://www.colinphillips.net/wp-content/uploads/2024/05/wilcox2023.pdf`

with the wh-effects in a nonisland minimal-pair counterpart, typically with gaps in object position. GPT-2 was found to be sensitive to all islands. Despite cross-model architectural variation, strongest effects have been found for coordination, adjunct, and complex NP islands. However, language models have weaker effects for left-branch, subject, and sentential subject islands [12].

[12] Ethan Gotlieb Wilcox, Richard Futrell, Roger Levy; Using Computational Models to Test Syntactic Learnability. Linguistic Inquiry 2024; 55 (4): 805–848. `https://www.colinphillips.net/wp-content/uploads/2024/05/wilcox2023.pdf`



a. *Adjunct islands*
*I know what the patron got mad after the librarian placed ___ on the wrong shelf.

b. *Complex NP islands*
*I know what the actress bought the painting that depicted ___ yesterday.

c. *Coordinate structure islands*
*I know what the man bought and ___ at the antique shop.

d. *Left-branch islands*
*I know how expensive you bought ___ a car last week.

e. *Sentential subject islands*
*I know who for the seniors to defeat ___ will be trivial.

f. *Subject islands*
*I know who the painting by ___ fetched a high price.

g. Wh-*islands*
*I know who Alex said whether your friend insulted ___ yesterday.

Figure 3: Islands associated with syntactic constraints, based on Ross (1967) and Huang (1982)

Binding

Syntacticians distinguish **anaphors** (reflexive pronouns like him/her/themselves), **pronouns** (he, her) and **R-expressions**, which are Noun Phrases that get meaning from referring to an entity in the world. BLiMP's Anaphor agreement dataset contains minimal pairs that differ in the grammaticality of anaphors, which are required to agree with their antecedents in person, number, gender and animacy.

BLiMP's Binding dataset contains properties of the structural relationship between a pronoun and its antecedent. All paradigms, summarised in *Table* , illustrate aspects of Principle A, which characterises restrictions on the distribution of anaphors.

Since co-indexation cannot be annotated in BLiMP, Principles B and C, which characterise restrictions on pronouns and R-expressions, are not contained in the minimal pairs dataset. Binding Principle B precludes pronouns from being locally bound in the same manner as anaphors (e.g., *Mary said that Joe liked these pictures of her* v **Mary said that Joe liked these pictures of him*).

Control and Raising

BLiMP's Control/raising constructions highlight syntactic and semantic differences between various types of predicates in non-finite

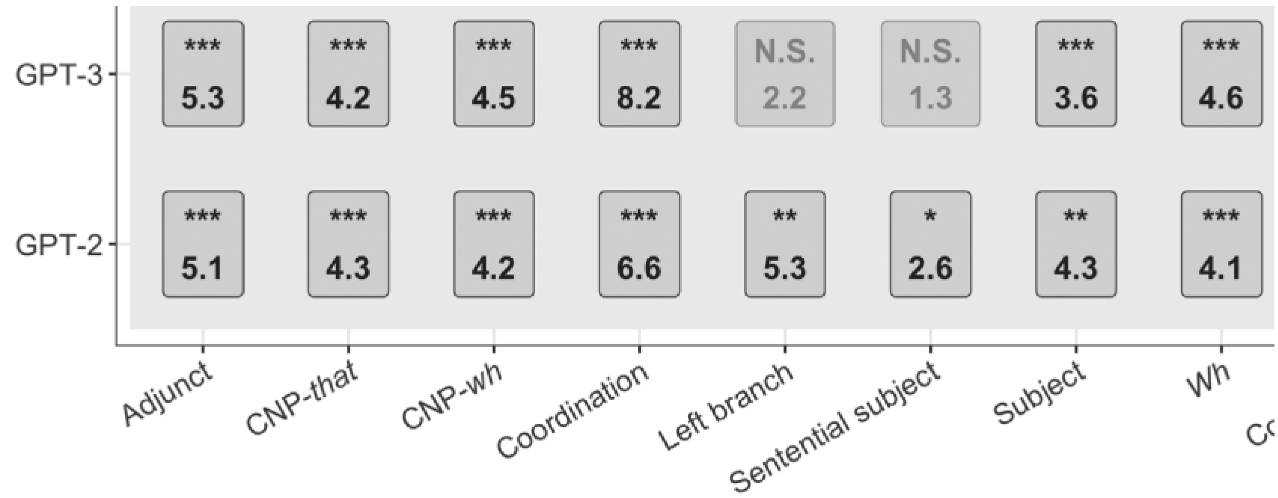


Figure 4: Summary of wh-effects across island test sets for GPT-2 and GPT-3

Category	Sentence
C-command	A lot of patients who can sell some couch didn't investigate **themselves/*itself**.
Principle A Case 1	The teenagers explain that **they/*themselves** aren't breaking all glasses.
Principle A Case 2	Eric imagines himself **taking/*took** every rug.
Domain 1	Carla had explained that Samuel has discussed **her/*herself**.
Domain 2	Donald can imagine those college campuses are boring **themselves/*himself**.
Domain 3	Steven explains Kayla won't hurt herself v Kayla explains Steven won't hurt herself.
Reconstruction	It's himself that **this cashier attacked/*attacked this cashier**.

clauses which embed an infinitival VP. We can broadly identify two types of constructions that lack an overt subject.

Raising constructions have a predicate with a syntactic argument that is naturally the semantic argument of its embedded predicate. In a sentence like *He seems to scare them.*, *seem* does not "select" the subject. Assuming a syntactic theory with movement, the argument *he* moves from the embedded clause to its subject position. Meanwhile, in **control** constructions, the matrix verb "controls" the arguments in the subordinate clause, e.g., in the sentence *John promises to help us*, the subject *John* is the controller of the arguments in *help* clause. We refer to *promise* as a control verb, which semantically selects its arguments.[13]

Note that these dependencies are not standardly represented in *basic* Universal Dependencies. **Enhanced Universal Dependencies Graphs** represents control and raising constructions via an **addi-**

[13] The syntax of control varies across formalisms. Generative linguists posit a null element PRO to formally accomodate control constructions in X-bar theoretic analyses

tional dependency (i.e. an additional `nsubj`) between a controlled verb and its controller or between an embedded verb and its raised subject.

BLiMP's datasets contain three types of raising and control constructions:

- ***tough*-movement predicates:** These are predicates involving verbs like *tough/difficult/easy* that allow the subject of the matrix clause to appear semantically as the object of the embedded clause. An example of this contrast in the BLiMP dataset is: *Julia wasn't fun to talk to.* v **Julia wasn't unlikely to talk to*

- **Existential *there*:** *there* is used to indicate that something exists, or to assert its non-existence. *William has declared there to be no guests getting fired.* v **William has obliged there to be no guests getting fired.*

- **Expletive *it*:** Dummy *it* is introduced in cases of raising (and extraposition). *Carla could declare it to be not so important that these doctors observe Rhonda.* v **Carla could convince it to be not so important that these doctors observe Rhonda.*

The syntax of control and raising extends beyond these simple cases of subject raising and subject control. English can have **object raising** and **object control**.[14] These cases are not explicitly handled in BLiMP.

Ellipsis

BLiMP does not offer full coverage of ellipsis, since it only considers sentences of equal length. The ellipsis paradigms cover special cases of NP ellipsis (or more, precisely, in X-bar terms **N-bar Ellipsis**) that meet this practical constraint:
Brad passed one big museum and Eva passed several. v ** Brad passed one museum and Eva passed several big.*

It is worth mentioning that English has several forms of **predicate/VP ellipsis (VPE)**:

- Auxiliary VPE: Susan has read War and Peace, but Maria hasn't.

- Modifier VPE: Susan can speak French, and Maria can too.

- Pseudogapping: Susan doesn't eat pasta, but she does pizza.

- Antecedent Contained Deletion: Susan has read every book Maria has.

Typologically, we can note that many Romance and Germanic languages lack Auxiliary VPE, although they do have Auxiliary VPE,

[14] Raising to Object verbs are also known as Exceptional Case Marking (ECM) Verbs. These are infinitives that have embedded accusative subjects, e.g., *Rosie believed him to be innocent.*

and pseudogapping is also more marginal here. Syntacticians typically attribute these differences to the nature of the English auxiliary system.

Syntax-Semantic Interface

BLiMP contains three "interface" phenomena:

- ARGUMENT STRUCTURE: the ability of different verbs to appear with different types of arguments. BLIMP's Argument Structure consists of verbs that appear with a direct object, participate in a **causative alternation** (*the boy broke the window* v *the window broke*), or take an inanimate argument.
- NPI LICENSING: restrictions on the distribution of **negative polarity items** like *any* and *ever*. limited to, e.g., the scope of negation and *only*.
- QUANTIFIERS: restrictions on the distribution of quantifiers. We cover two such restrictions: superlative quantifiers (e.g., *at least*) cannot embed under negation, and definite quantifiers and determiners cannot be subjects in existential-*there* constructions.

BLiMP Supplement

BLiMP Supplement was unofficially released for the BabyLM Shared Task and additionally contains minimal pairs datasets for **subject-auxiliary inversion** (e.g., *Is the novel he is putting away from the library?* v **Is the novel he putting away is from the library?*), and **hypernyms** (e.g., *If she has a dog, it must be the case that she has a mammal.* v **If she has a dog, it must be the case that she has a chihuahua.* Additionally, it contains **discourse phenomena** like **turn taking**:
"David: Should you quit? Sarah: No, I shouldn't."
*? *"David: Should she quit? Sarah: No, I shouldn't."*

Additionally, it contains datasets about **question-answering congruence**. This is the only minimal pairs dataset that is split into **difficulty levels** (EASY/HARD). An inimate v animate contrast is meant to be *easy*, while an animate vs. inanimate is meant to be *tricky*.
Easy: *"What did you get? I got a chair."* v *? *"What did you get? I got a teacher."*
Tricky: *"Who cleaned? David cleaned."* v *? *"Who cleaned? The patio cleaned."*

Syntactic Typology

Morpho-Syntactic Typology for NLP

Morpho-syntactic typology investigates how languages differ in the grammatical systems that determine how words combine into larger structures and how grammatical information is encoded. These differences are systematic, and they shape how prediction, generalisation, and compositional reasoning operate in human language. Morpho-syntactic typology provides both a catalogue of grammatical design possibilities and a theory of the cognitive and communicative pressures that shape them. This typological variation defines a challenge space for language technology beyond English: word order, agreement, morphology, alignment systems, and argument structure all influence what information must be represented, where, and how robustly.

- A core dimension of typological variation concerns *word order*. Greenberg's universals show that languages cluster around a small set of ordered patterns, and disharmonic word orders are rare and harder for both humans and models to acquire. Cognitive and communicative pressures have been proposed to explain these universals, and multilingual LLM performance often reflects similar pressures.

- Word order interacts with *alignment* ((how different languages define what counts as a "subject")—whether a language treats the subject of transitives and intransitives alike (nominative–accusative) or aligns subjects of intransitives with objects (ergative–absolutive). The acquisition and processing of ergative systems and indirect object marking (e.g., datives) reveal how argument structure is encoded crosslinguistically, and provide targeted tests for LM syntactic generalisation.

- *Relative clause* and *subordination* structures highlight how structural complexity varies crosslinguistically. Head-internal relative clauses can function as developmental precursors to externally headed relatives, and the Noun Phrase Accessibility Hierarchy predicts crosslinguistic difficulty patterns for both humans and models—e.g., subject relatives are generally easier than object relatives. These structures offer controlled diagnostics for long-distance dependencies inside LLMs.

- Variation in *argument realisation* describes how grammars distributes information between morphology and syntax. Some languages permit radical argument drop or pronominal arguments (e.g., polysynthetic systems), whereas others use rich case and agreement morphology or rely on strict word order.

- Variation in tense–aspect–mood (TAM) marking, and in whether morphology is free or bound, poses challenges for sequence modelling and generalisation. Similarly, isolating languages with little inflection test whether models can infer structure without overt morphological cues.
- Nominal systems add further complexity: *count vs. mass distinctions*, classifier systems, number marking, definiteness, and grammatical gender all shape reference and quantification. Typologically distant nominal systems are known to cause strong L2 transfer effects in humans, and provide a natural way to study linguistic distance measures and cross-lingual interference in LLMs.
- Interrogatives, negation, and discourse-sensitive categories contribute additional axes of variation, particularly relevant for instruction-tuned systems. More broadly, multilingual language models must contend with *crosslinguistic influence*: structural transfer, positive or negative, can reveal whether models rely on universal or language-specific features.

A Case Study: Deep Subjecthood in mBERT (Papadimitriou et al (2021)

Papadimitriou et al. (2021) investigate whether multilingual BERT (mBERT) encodes **high-level grammatical features** that cannot be inferred from individual sentences alone. Their focus is **morphosyntactic alignment**—a typological property that determines how languages identify the grammatical "subject." Some languages group intransitive subjects (S) with transitive subjects (A), forming a **nominative–accusative** system (e.g., English), while others group S with objects (O), forming an **ergative–absolutive** system (e.g., Basque).

An **ergative language** is characterized by a mapping of grammatical roles $\{\text{AGENT}_{\text{Transitive}}, \text{SUBJECT}_{\text{Intransitive}}, \text{PATIENT}_{\text{Transitive}}\}$ such that the **ergative transitive subject** occupies a privileged position: it patterns distinctly from intransitive subjects, reflecting the language's ergative-absolutive alignment.

The following examples from Basque demonstrate an ergative–absolutive case marking system:

Sentence	**Martin etorri da.**		**Martinek Diego ikusi du.**		
Word	Martin-Ø	etorri da	Martin-ek	Diego-Ø	ikusi du
Gloss	Martin-ABS	has arrived	Martin-ERG	Diego-ABS	has seen
Function	S	VERB$_{\text{intrans}}$	A	O	VERB$_{\text{trans}}$
Translation	"Martin has arrived."		"Martin has seen Diego."		

In these examples, Basque demonstrates an ergative–absolutive alignment. In the intransitive sentence *Martin etorri da* "Martin has arrived," the subject *Martin* takes the absolutive case, marked here with a zero morpheme (−) because proper nouns are unmarked for absolutive. In the transitive sentence *Martinek Diego ikusi du* "Martin has seen Diego," the transitive subject *Martin* is marked with the ergative suffix `-ek`, while the object *Diego* remains in the absolutive. Basque nouns generally require a determiner: the singular absolutive determiner is `-a` and the plural is `-ak`. When the ergative case is applied together with the determiner, suffixes combine according to phonological rules, producing forms like `gizon-a` (man-the.sing.abs), `gizon-ak` (man-the.pl.abs), `gizon-ek` (man-the.pl.erg), and `gizon-ak` (man-the.sing.erg). Thus, in these sentences, the ergative case signals the agent of a transitive verb, whereas the absolutive marks both intransitive subjects and transitive objects.

Because this property emerges across sentences rather than from any single utterance, ergativity offers a strong test of whether neural language models internalise abstract grammatical structure.

mBERT embeddings **implicitly encode high-order morphosyntactic alignment**, including nominative–accusative versus ergative–absolutive distinctions, in a **distributed, emergent manner** that is not locally identifiable in individual sentences. mBERT represents subjecthood and objecthood robustly and probabilistically. Representations are general enough such that mBERT can transfer across languages, but also language-specific enough that mBERT learns language-specific abstract grammatical features. [15]

mBERT represents **continuous, probabilistic subjecthood** in a cross-linguistically aligned embedding space, demonstrating that typologically meaningful grammatical patterns emerge **from distributed representations rather than local token-level cues**.

Method: Train classifiers to predict grammatical subjecthood from mBERT contextual embeddings and examining their behavior both within and across languages. For each language and for each mBERT layer l, the authors train a two-layer perceptron (hidden size 64) on balanced embeddings of 1,012 transitive subjects (A) and 1,012 transitive objects (O) extracted from Universal Dependencies (UD) treebanks for 24 languages. Once trained, classifiers are tested on A, O, and intransitive subjects (S), allowing us to probe how subjecthood is represented in distributed embedding space and how language-specific morphosyntactic alignment patterns are encoded.

This framework is used to conducted systematic linguistic experiments with mBERT:

- **Does mBERT encodes nominative-accusative or ergative-absolutive alignment?** Train classifiers on a single language to distinguish

[15] Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.215. URL `https://aclanthology.org/2021.eacl-main.215/`

transitive subjects (A) from objects (O) so the classifier can capture whatever features mBERT uses to encode subjecthood in that language. Test on intransitive subject (S) nouns, which are not in the training data. Classifier labels reveal underlying alignment patterns: S nouns in nominative languages are mostly classified as A, while in ergative and split-ergative languages (e.g., Basque, Hindi, Urdu), the probability of S being classified as O is higher, reflecting the emergent encoding of alignment patterns in mBERT. Classifiers achieve high accuracy (90%+) in layers 7–10, demonstrating that contextual embeddings reliably encode subjecthood.

- **Do these representations generalise across languages?** Authors train a classifier on one language and zero-shot evaluate on another. Transfer accuracy averages 82.6% across language pairs, with some pairs exceeding 90%. Notably, classifiers trained on ergative languages assign higher probabilities of being O to S nouns in other languages, indicating that alignment properties are encoded in a language-specific yet interlingually generalizable way. This suggests that mBERT's multilingual embedding space organizes syntactic and semantic relations in parallel, transferable representations.

- **How do semantic and syntactic features influence subjecthood representations?** Authors focus on passives, animacy and case. Classifiers are ambivalent about passive subjects, reflecting that subjecthood in mBERT space is sensitive to semantic agency. Animacy is a strong predictor: animate nouns are more likely to be classified as A, and classifier probabilities vary with agentive case markings, even when controlling for syntactic role. These results demonstrate that subjecthood is encoded as a continuous, probabilistic property rather than a discrete syntactic label.

Case Study: Probing for Dependencies in mBERT

To investigate whether Multilingual BERT (mBERT) encodes universal syntactic structure, Chi et al (2020) employ the *structural probing* framework of Hewitt and Manning (2019), extended to the multilingual setting.[16] The central idea is to learn a linear transformation of the model's contextual word representations such that the squared Euclidean distance between transformed vectors approximates the syntactic distance between words in a dependency parse tree.

Hewitt and Manning (2019) introduced the *structural probe* to investigate syntactic information encoded in contextual word representations. Each dependency tree T is represented as a distance metric, where the distance $d_T(w_i, w_j)$ between words w_i and w_j is the num-

[16] Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.493. URL `https://aclanthology.org/2020.acl-main.493/`

ber of edges in the path connecting them. The structural probe seeks a linear transformation of the representation space such that squared Euclidean distances approximate tree distances across all sentences.

Formally, let $h_{1:n}$ denote the sequence of contextual representations for a sentence of length n, and let $B \in \mathbb{R}^{k \times m}$ define a k-dimensional syntactic subspace of the m-dimensional representation space. The probe defines a squared distance

$$d_B(h_i, h_j) = \|Bh_i - Bh_j\|_2^2,$$

and optimizes B to minimize the discrepancy with tree distances across the training set $\mathcal{S}$:

$$\arg\min_B \sum_{s \in \mathcal{S}} \sum_{i,j} \left| d_T(w_i, w_j) - d_B(h_i, h_j) \right|.$$

Departing from prior work, the probe-transformed vectors Bh_i themselves are analyzed, rather than only their pairwise distances. Each row of B forms a basis of the syntactic subspace; a vector Bh corresponds to a point in this subspace, with each dimension reflecting the projection of h onto a basis vector.

Experimental Settings. Authors evaluated the 110M-parameter BERT-Base, Multilingual Cased (mBERT) model on eleven languages with large Universal Dependencies (UD v2) treebanks: Arabic, Chinese, Czech, English, Farsi, Finnish, French, German, Indonesian, Latvian, and Spanish.

Two baselines were used:

- **MBERTRAND:** mBERT with randomly reinitialized attention layers, leaving subword embeddings and positional encodings unchanged.
- **LINEAR:** All sentences assigned a left-to-right chain dependency structure.

The probe's effectiveness was measured with two metrics: (1) Spearman correlation between predicted and true word pair distances ($D_{\text{Spr.}}$), and (2) undirected unlabeled attachment score (UUAS), computed by constructing a minimum spanning tree from predicted distances and comparing to the gold tree. Evaluation was restricted to sentences of length 5–50, averaging first by sentence length and then across lengths. Probes of varying maximum ranks were trained on embeddings from all 12 mBERT layers to identify the layers and subspace dimensionalities that most effectively encode syntax.

Probe Rank and Layer Selection. The authors found that syntax is most accurately recovered from the middle layers of mBERT, particularly layers 7 and 8. Increasing the probe's maximum rank beyond approximately 64–128 dimensions yielded negligible improvements, indicating that syntactic information is concentrated in a relatively low-dimensional subspace of the 768-dimensional representation space. This observation underscores the efficiency of the structural probe in isolating syntactic structure without requiring full-dimensional representations.

Cross-Lingual and Joint Subspace Probing. To assess whether mBERT's syntactic subspaces are shared across languages, the authors conducted several cross-lingual probe evaluations:

- **SINGLETRAN:** Probes were trained on a single source language chosen post hoc to maximize transfer to a target evaluation language. This tests whether syntax encoded in one language's subspace is predictive for another.
- **HOLDOUT:** Probes were trained on all languages except the target language, then evaluated on the held-out language. This measures whether a subspace learned from multiple languages generalizes to an unseen language.
- **ALLLANGS:** Probes were trained on the concatenation of data from all languages, including the evaluation language. This evaluates a joint cross-lingual syntactic subspace.

The experiments demonstrate that mBERT encodes syntactic information in subspaces that are highly transferable across languages. Here, *UUAS* (undirected unlabeled attachment score) measures the percentage of syntactic head-dependent relations correctly predicted, ignoring direction and labels, while $D_{\text{Spr.}}$ denotes the Spearman correlation between predicted and true pairwise word distances.

When evaluating cross-lingual transfer:

- **SINGLETRAN:** Probes trained on the single best source language achieved, on average, an improvement of 14 UUAS points and 0.128 $D_{\text{Spr.}}$ over the stronger of the two baselines.
- **HOLDOUT:** Probes trained on all languages except the evaluation language achieved 16 UUAS points and 0.137 $D_{\text{Spr.}}$ improvement, demonstrating generalization to unseen languages.
- **ALLLANGS:** Probes trained on the concatenation of all languages, including the evaluation language, achieved 19 UUAS points and 0.156 $D_{\text{Spr.}}$ improvement, indicating that a joint cross-lingual subspace captures shared syntactic structure.

Across most languages, the cross-lingual subspaces accounted for 62–88% of the total possible UUAS improvement over baselines, highlighting that **a shared syntactic subspace encodes substantial syntactic information even across typologically diverse languages**.

xBLiMPs: Evaluating Syntax beyond English

Minimal Pairs datasets have been introduced beyond English:

1. **CLAMS (French and German):** The Cross-Lingual Syntactic Evaluation of Word Prediction Models (CLAMS) [17] generates minimal pair datasets which we use for French and German using Attribute-Varying Grammars. The dataset assesses grammaticality in Simple Agreement, VP coordination, and across "interveners" in S-V agreement (subject/object relative clause or across a Prepositional Phrase).
2. **JBLIMP (Japanese):** JBLIMP [18] is a minimal pairs dataset for targeted syntactic evaluation of Japanese. It consists of 331 minimal pairs of syntactic acceptability judgements curated from Japanese syntax articles in the *Journal of East Asian Linguistics*. The JBLiMP Minimal Pair dataset can be found here: `https://github.com/osekilab/JBLiMP/tree/main`

   As noted by Basar et al (2025), benchmarks using a similar template-based approach as BLiMP include CLiMP (Chinese, Xiang et al., 2021), ZhoBLiMP (Chinese, Liu et al., 2024), BLiMPNL (Dutch, Suijkerbuijk et al., 2025), and for Basque/Swahili/Hindi by Kryvosheieva and Levy (2025).[19]

   Another approach is based on modifying Universal Dependency trees, which has been used by SLING (Chinese, Song et al., 2022), RuBLiMP (Russian, Taktasheva et al., 2024), and MultiBLiMP (Jumelet et al., 2025), a multilingual benchmark covering 101 languages.

   Other approaches include the extraction of minimal pairs from linguistics journals, employed by JBLiMP (Japanese, Someya and Oseki, 2023), manual creation of pairs, as done for Icelandic by Ármannsson et al. (2025), and the usage of LLMs for generating pairs, as done for Tamil and Indonesian by Leong et al. (2023)
3. **SLING (Chinese): SLING** [20] is a 38K minimal sentence pair dataset derived by applying syntactic and lexical transformations to Chinese Treebank 9.0, aiming to improve on the limitations of an earlier dataset called CLiMP [21], which had a lack of diversity in the vocabulary to generate minimal pair templates. The SLING Dataset can be found here: `https://huggingface.co/datasets/suchirsalhan/SLING`

[17] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.490. URL `https://aclanthology.org/2020.acl-main.490`

[18] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-eacl.117. URL `https://aclanthology.org/2023.findings-eacl.117`

[19] Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2506.13487*, 2025

[20] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.305. URL `https://aclanthology.org/2022.emnlp-main.305`

[21] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter*

Due to the small size of the JBLIMP minimal pairs dataset, Someya and Oseki [2023]'s recommend to compute accuracy using a SLOR score to mitigate the confounding effects of lexical frequencies and sentence lengths, which is defined as follows:

$$SLOR(X) = \frac{\log p_m(X) - \log p_u(X)}{|X|}$$

where $p_m(X)$ is the probability of a sentence for a Language Model and is the unigram probability of the sentence, estimated for each subword in the training corpus. Accuracy calculations for other languages follows dataset guidance to use unnormalised log-probabilities.

BLiMP, CLiMP, SLING and JBLiMP all use a forced-choice paradigm to validate their minimal pairs with human native speakers. All papers explore the effect of training data size – CLiMP and JBLiMP found no influence of dataset size. while SLING found that smaller models may have performed better for some. The performance gap between the LMs and the native speakers is large on these cross-lingual minimal pairs datasets (and larger than it was for English). Also, models perform better at local dependencies compared to longer-distance dependencies.

Chinese Syntax: SLING highlights a few important properties of Mandarin syntax. Chinese has a rich system of **classifiers**, so there is an additional syntactic task of **classifier-noun agreement** when a noun is modified by a numeral or demonstrative. **Chinese Definiteness Effect** is a restriction of the distribution of *zhe* (this)/*na* (that) and the quantifier *mei* (every), which may not occur in the post-verbal position of an existential *you* (there is) sentence. Chinese has perfective aspect markers *le* and *guo*. SLING contains minimal pairs that contrast these markers with the tense and the progressive marker *zai*.

Japanese Syntax: Japanese has analytic morphology, so JBLIMP generalises BLiMP's irregular forms dataset to incorporate minimal pairs on morphology in general. Japanese doesn't have explicit determiner-noun agreements, so JBLiMP drops BLiMP's determiner-noun agreement category for a more general Nominal Structure dataset.

BLiMP-NL is a carefully designed new minimal pairs dataset for Dutch, which requires the critical region must be the same for the sentences of the minimal pair, unlike BLiMP to facilitate easier evaluation of langauge models and human evaluation. They source their sentences from Dutch Syntax handbooks.[22]

RuBLiMP [23] covers subject-predicate agreement, agreement (NP agreement, floating quantifier agreement, anaphor agreement), Government (checks correct assignment of cases governed by verbs,

[22] `https://osf.io/preprints/psyarxiv/mhjbx`

[23] Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. RuBLiMP: Russian benchmark of linguistic minimal pairs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.emnlp-main.522`

prepositions, or nominalizations in Russian Case System), Negative Concord and Negative particle movement, and Reflexives (minimal pairs to assess correct use of Russian reflexive pronouns in constructions with external possessors)

TurBLiMP: This is particularly interesting because Turkish has **non-configurational word order**.[24] All 6 permutations of Subject (S), Object (O) and Verb (V) are possible. This flexibility comes from case-marking which distinguishes the subject and object. As noted by Basar et al (2025), this means we can test LMs for their robustness to different positional patterns or grammatical hierarchies, in a way that is not possible with English and other fixed-order languages that dominate the training material of current LLMs. This is something that has already been studied independently in other work.[25]

Monolingual and Multilingual Evaluation

Across 55 test tasks, there are consistent performance differences between monolingual and bilingual models on 16 tasks. Despite their smaller sizes, monolingual models perform better on 12 and worse only on 4 tasks (Zhou & Matusevych, 2025). This result suggests that bilingual Chinese + English bilingual models may suffer from negative crosslingual transfer. → Implications for Multilingual Pretraining and Cross-Lingual Transfer

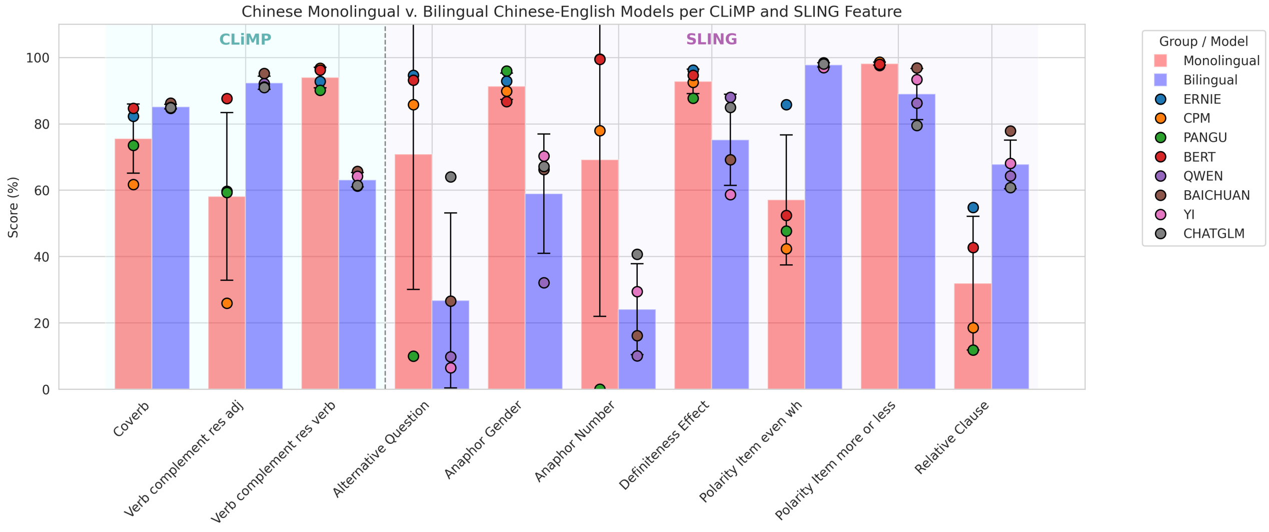


Figure 5: *Figure from Zhou & Matusevych, 2025*, `https://aclanthology.org/2025.gem-1.58/`

[24] Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2506.13487*, 2025

[25] See **Mission: Impossible Language Models** (Kallini et al 2024) for an example of one approach, https://aclanthology.org/2024.acl-long.787.pdf

Interpretability Techniques

Individual scores across datasets are not enough to understand the syntactic capabilities of a language model. The Minimal Pairs Paradigm compares the probability of two stand-alone text sequences without any explicit linguistic context. But, this is not necessarily a **naturalistic/realistic** approach, as local contextual information and discourse context can potentially influence grammaticality judgements.[26] Language Models exhibit changes in model performance that are not explainable by acceptability-preserving syntactic perturbations: LLMs have been found to have quantitative, graded effects of structural priming on string probabilities, subject to the length of the context, arising as a consequence of the in-context learning capabilities of Transformer architectures.[27] This has lead to interest in **interpretability techniques** for evaluating the syntactic capabilities of LMs.

There are two main families of Interpretability methods that extend traditional Targeted Syntactic Evaluation (TSE). **Circuit Localization:** methods to locate the most important subsets of a model for performing a given task. **Causal Variable Localization:** Featurising hidden vectors and selecting features that map. We review both, before surveying their application to **linguistic interpretability**.

A circuit C is a set of nodes and edges in the computation graph of a neural network N. Nodes typically correspond to submodules or attention heads (e.g., layer 5 MLP or attention head 10 at layer 12). Edges represent information flow between nodes. Metrics for evaluating circuits are not standardized. Two complementary goals exist:

- Measure whether C contributes positively to model performance.
- Measure whether C has any measurable effect, positive or negative, on performance.

The Mechanistic Interpretability Benchmark formalize these as two metrics. Circuit Performance Ratio (CPR) prioritizes components with a positive effect on model performance. Higher is better.

Circuit-Model Distance (CMD): CMD measures the deviation of the circuit's behavior from the full model, including negative effects. Lower is better (0 is optimal).

Given a circuit C and network N, define faithfulness:

$$f(C, N; m) = \frac{m(C) - m(\emptyset)}{m(N) - m(\emptyset)}$$

where m is a logit difference metric for counterfactual evaluation and $\emptyset$ is the empty circuit.

[26] *Syntactic Priming* is widely studied in psycholinguistics, studying the effect of linguistic contexts with only one or a small number of context sentences

[27] `https://aclanthology.org/2023.acl-long.333.pdf`

Computing CPR and CMD

1. For proportions $k \in \{0.001, 0.002, 0.005, \ldots, 1\}$, discover circuits C_k with $|C_k|/|N| \leq k$.
2. Compute faithfulness $f(C_k)$.
3. Approximate integrals using the trapezoidal rule:

$$\text{CPR} \approx \int_0^1 f(C_k)dk, \quad \text{CMD} \approx \int_0^1 |1 - f(C_k)|dk$$

Causal Variable Localization evaluates methods for localizing specific concepts along causally active paths in a model. High-level causal model H represents concepts as variables. The goal is to align variables in H with low-level features Π_X in the neural network that share the same mechanistic role. Features are obtained from hidden vectors $h \in \mathbb{R}^d$ through a mapping $F : \mathbb{R}^d \to F^k$, where Π denotes selected feature dimensions. Alignments may vary dynamically across tokens and layers.

Given base and counterfactual inputs (b, c), define:

$$\text{Faith}(X, \Pi_X, H, D) = \sum_{(b,c) \in D} \mathbf{1}[H_{X \leftarrow \text{Get}(H(c), X)}(b) = N_{\Pi_X \leftarrow \text{Get}(N(c), \Pi_X)}(b)]$$

Measures how well the neural features Π_X reproduce the effect of intervening on X in H.

For example, a Causal model for Multiple Choice Question Answering (MCQA), H_{MCQA}, has variables:

- T: text input
- X_{Order}: answer position
- O_{Answer}: answer token

SyntaxGym

SyntaxGym is a syntactic evaluation benchmark designed with more stringent evaluation criteria. For 34 different linguistic phenomena, the SyntaxGym benchmark defines test items with two to four different conditions, consisting of minimal structural variations on the same sentence which render the sentence either grammatical or ungrammatical. Model log-likelihoods are measured at a critical region within each sentence, rather than across the whole sentence, and models are expected to produce log-likelihoods that satisfy multiple inequalities across all conditions.[28]

[28] Hu et al (2020) *A Systematic Assessment of Syntactic Generalization in Neural Language Models*, https://aclanthology.org/2020.acl-main.158/. Here are the test sets: https://syntaxgym.org/

Syntactic Circuits: A Mechanistic Interpretability Approach

Unlike earlier **probing strategies**, which use a small model trained to extract linguistic information from a target model, causal interventions are the dominant methodology in current mechanistic interpretability work.[29]

Applying this to grammaticality detection, we can adopt a **causal intervention paradigm** to assess grammaticality. The core idea of an intervention is to take a base input b and a source input s and replace a given model-internal component (a "neuron"), f, with $f * (b, s)$, and assess the effect of this intervention on model output to establish causal relationships.

CausalGYM takes an input minimal pair that has an alternation that affects next-token prediction, then intervenes on the base forward pass using a pre-defined intervention function that operates on aligned representations from both inputs. Then, it is possible to determine how this intervention impacts next-token prediction probabilities. In aggregate, such interventions assess the causal role of the intervened representation on the model's behaviour.[30]

We can use **directionality** for causal effect is an intuitive test for whether they reflect features that the model uses downstream. **Distributed alignment search (DAS)** learns the **intervention direction**, potentially distributed across many neurons, that maximises the output probability of a **counterfactual label**. The counterfactual label is obtained by recasting a minimal pair, like S-V agreement, from SyntaxGYM into counterfactual pairs that elicit singular or plural verbs based on the number feature of the subject, and hold everything else (including the distractor) constant: (a) *The author near the senators → is* (b) The authors near the senators → are. One of the advantages of this paradigm is that it facilitates an analysis of **model learning dynamics** rather than analysing input/output relationships. Experiments have only been conducted for English and a limited test set, so there is scope for further empirical work in this area. Another line of research identifies **"circuits"** in LMs that handle different tasks. [31]

[29] Traditionally, if the probe could predict a target structure, then it was argued that the probe can predict a particular structure, so it is the model it's trained on has implicitly learned to encode it. However, a probe achieving high classification accuracy provides no guarantee that the model actually distinguishes those classes in downstream computations.

[30] `https://aclanthology.org/2024.acl-long.785/`

[31] `https://aclanthology.org/2024.findings-emnlp.591.pdf`

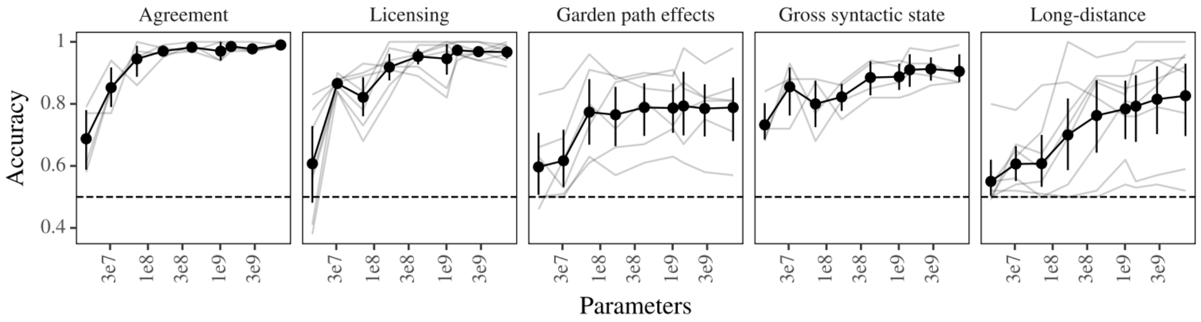


Figure 3: Accuracy of pythia-family models on the **CausalGym** tasks, grouped by type, with scale. The dashed line is random-chance accuracy (50%).

Figure 6: *Figure from Arora et al (2024)*

Syntax-Augmented Language Models

There is a more niche sub-literature focussed on **Syntactic Language Models (SLMs)**[32], which incorporate syntactic biases explicitly in pretraining.

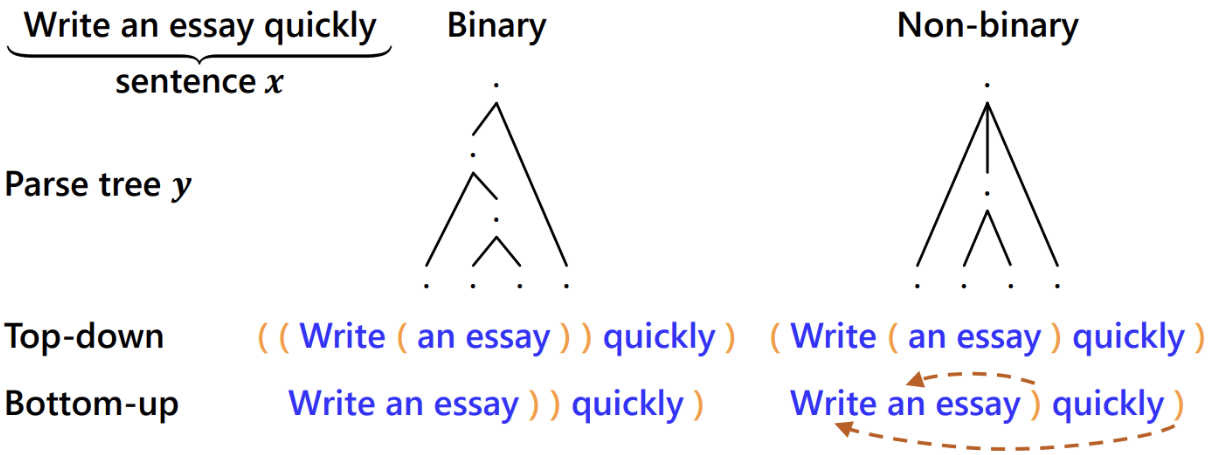

Write an essay quickly
sentence x
Binary
Non-binary
Parse tree y
Top-down
((Write (an essay)) quickly)
(Write (an essay) quickly)
Bottom-up
Write an essay)) quickly)
Write an essay) quickly)

Figure 7: *Binary and Non-Binary Parse Trees with Top-Down and Bottom-Up Linearizations. Figure from Zhao et al (2025)*

Augmenting LMs with some syntactic inductive bias has been a long-line of research in NLP: some work has tried to jointly model the distribution of sentences and their structures[33], however more recently Transformer Grammars[34] and other related works utilise a Transformer-based architecture augmented with constituency-based compositionality.

Syntactic LMs model **linearised syntactic parse trees** jointly along

[32] Not to be confused with *small language models*!

[33] See Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 199–209, San Diego, California. Association for Computational Linguistics

[34] See Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. ´ 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. Transactions of the Associa-

surface sentences. A major class of SLMs, which is referred to by Zhao et al (2025) are called compositional SLMs.[35] These are based on constituency parse trees and contain explicit composition of subconstituent representations to form constituent representations. A constituency parse can be linearised either top-down or bottom-up. Another variable is **non-binary-branching trees** – potentially **unary branches** can be eliminated. For the bottom-up linearization of the non-binary tree shown in *Figure* , arcs are used to point from ")" actions to their corresponding start positions.

Another line of work augments language models with learnable structures, such as stack-structured memory where syntax patterns are learned from data rather than being predefined.

Syntactic LMs are evaluated on Syntactic Generalization scores from SyntaxGym [36].

Beyond BLiMP-style Datasets

Neural Language Models rely heavily on the input that aligns with individual constructions and can even struggle with certain dependencies such as topicalisation. However, where they currently suffer is **generalising *beyond the input*** to learn a shared representation for a given construction. BLiMP and similar datasets cross-lingually do not necessarily characterise various aspects of human syntactic competence. If model evaluation is meant to be theory-agnostic, there is an additional criticism of whether the properties encoded in evaluation datasets are what we should be evaluating (e.g., the top-down desiderata of Construction Grammar may well differ from Generative Grammars),[37].

Additionally, minimal pairs codify an incorrect assumption that there is a strict grammaticality decision boundary. Psycholinguists, for example, have found evidence of **syntactic satiation** – comprehenders, can for example, find island-violating sentences (e.g., **What did John think a bottle of fell on the floor?*) increasingly acceptable given repeated exposure. Syntactic Acceptability is contingent on **linguistic adaptivity** and speaker-specificity; **parsability** is another important constraint. These "edge cases" emphasise the gradient nature of acceptability. Ideally, **meta-data** in test sets (e.g., including information about speaker dialects) would be a useful additional source of information, potentially facilitating more fine-grained information about morphosyntactic differences among dialects and varieties.

[35] Yida Zhao, Hao Xve, Xiang Hu, and Kewei Tu. A systematic study of compositional syntactic transformer language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7070–7083, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.350. URL https://aclanthology.org/2025.acl-long.350/

[36] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.158. URL https://aclanthology.org/2020.acl-main.158/

[37] https://aclanthology.org/2024.clasp-1.7.pdf

Log-Likelihoods and Grammaticality

Understanding the extent to which probabilistic language models (LMs) acquire grammatical knowledge is central to linguistic theory and models of language learning. While previous research has explored LM grammatical competence with varying conclusions, current methods of evaluation face conceptual challenges. Directly prompting models for grammaticality judgments conflates grammatical knowledge with meta-linguistic understanding and pragmatic competence. Conversely, evaluating string probabilities raises concerns due to the classical linguistic distinction between grammaticality and probability.

A recent paper by Hu et al (2025) develops a formal framework reconciling these tensions by proposing that the probability of a string reflects two latent factors: the probability of the intended message and whether that message is grammatically realized. Under this view, minimal pairs—grammatical and ungrammatical strings expressing (approximately) the same message—provide the most reliable diagnostic for grammatical knowledge, because they control for message probability while isolating the effect of grammatical structure.[38]

Three theoretical predictions follow: (i) log-probabilities of grammatical and ungrammatical strings within minimal pairs should be correlated due to shared message probability; (ii) differences in model log-probabilities within minimal pairs should align with human acceptability contrasts; and (iii) probability alone should fail to cleanly separate grammatical from ungrammatical sentences when broad, unpaired samples are considered, due to confounding from highly probable message content. These predictions emphasize the necessity of controlled comparisons and caution against using absolute probability thresholds to infer grammaticality.

The authors empirically validate their predictions on over 280K English and Mandarin sentence pairs across multiple benchmarks and language models, demonstrating strong support for the theoretical account. The results provide a **principled justification for minimal-pair evaluations widely used in NLP**, and offer methodological guidance for future assessments of grammatical competence in language models.

Beyond Monolingual Evaluation

There are two recent minimal pairs datasets that extend minimal pairs beyond monolingual evaluation:

- **Code-Switching (Sterner & Teufel 2025)**

[38] Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P Levy. What can string probability tell us about grammaticality? *arXiv preprint arXiv:2510.16227*, 2025

- **BLiSS 1.0 (Gao, Salhan, Caines, Buttery & Sun 2025)**

For models that aim to be cognitively plausible, we need a complementary, acquisition-focused perspective, one that inspects how grammar competence is organised and learned. This evaluation gap is important for models of Second Language Acquisition (SLA), which we refer to as L2LMs (Aoyama and Schneider, 2024). A central characteristic of the SLA process is the production of systematic errors. These deviations are not random noise, but rather structured evidence of the learneras developing internal ^ grammar, or interlanguage. This is the motivation for BLiSS 1.0 (Gao, Salhan, Caines, Buttery & Sun 2025).

The BLiSS 1.0 benchmark is a large-scale evaluation suite composed of controlled triplets designed to test a model's selective tolerance for naturalistic learner production errors. The evaluation framework for BLiSS is designed to move beyond evaluations of the formal competence of a Language Model (e.g., using broad-coverage datasets like BLiMP (Warstadt et al., 2020)) to evaluate the alignment of a language model with second language acquisition. The BLiSS 1.0 benchmark is designed to evaluate how closely a language model's outputs align with patterns observed in second language (L2) learners, particularly in terms of grammatical errors.

BLiSS thus extends evaluation beyond formal competence, providing a framework to test whether models selectively tolerate or reproduce error patterns in ways that resemble human learners. Concretely, we develop BLiSS to enable the study of:

- **Error-type sensitivity**: Whether language models recognize and react differently to common L2 errors (e.g., determiner omission, verb tense errors).
- **Position awareness:** By generating artificial errors at positions distinct from the learner's original error, we can test if language models are sensitive to the locus of grammatical deviations, not just their existence.
- **Learner-informed evaluation:** Leveraging metadata such as L1 background and proficiency level that are available in large-scale corpora allows analysis of model behavior in the context of typologically diverse learner populations.

Conclusions

Overall, there are three main takeaways from a careful analysis of BLiMP. First, it is very important to **scrutinise the datasets you are using off the shelf**. In this case, as is also the case when evaluating models on semantic evaluation datasets, a qualitative evaluation can often tell you more than reporting scores, particularly considering what the limitations of the datasets are (e.g., BINDING does not offer full coverage). Qualitative evaluation of inter-model performance and comparing models to human ceilings can tell you much more about a model than simply reporting a macro-average. Solely reporting macro-averages does not encapsulate any notion of causality in model learning and does not provide a meaningful assessment of learning trajectories in pre-training. Secondly, it is important to be able to assess what the appropriate means of **reporting scores** for a given task is – differences in accuracy calculations are contingent on model architecture and it is important to consider morphological type and frequency beyond English. Finally, like many other areas in NLP, minimal pair datasets suffer from the standard flaws of "starting from English", and subsequent non-English datasets inherit BLiMP's architecture, which may not be ideal. Even beyond English, these datasets have only been built for high-resource languages, and the extent to which the generation of minimal pairs can be feasibly conducted in low-resource regimes is currently unclear.

References

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2506.13487*, 2025.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.493. URL https://aclanthology.org/2020.acl-main.493/.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.306. URL https://aclanthology.org/2023.emnlp-main.306.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.158. URL `https://aclanthology.org/2020.acl-main.158/`.

Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P Levy. What can string probability tell us about grammaticality? *arXiv preprint arXiv:2510.16227*, 2025.

Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada, July 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-short.80. URL `https://aclanthology.org/2023.acl-short.80`.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.490. URL `https://aclanthology.org/2020.acl-main.490`.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.215. URL `https://aclanthology.org/2021.eacl-main.215/`.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.240. URL `https://aclanthology.org/2020.acl-main.240`.

Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL*

2023, pages 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-eacl.117. URL `https://aclanthology.org/2023.findings-eacl.117`.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.305. URL `https://aclanthology.org/2022.emnlp-main.305`.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. RuBLiMP: Russian benchmark of linguistic minimal pairs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.emnlp-main.522`.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.242. URL `https://aclanthology.org/2021.eacl-main.242`.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.90. URL `https://aclanthology.org/2021.acl-long.90`.

Yida Zhao, Hao Xve, Xiang Hu, and Kewei Tu. A systematic study of compositional syntactic transformer language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7070–7083, Vienna, Austria, July 2025. Association for Computational

Linguistics. ISBN 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.350. URL https://aclanthology.org/2025.acl-long.350/.