

The Wug Test Revisited: A Cross-Linguistic Protocol for Comparing Morphological Productivity in Humans and Large Language Models

Zhaleh Gholami ¹ , Hamed Zarabi ^{2*} 

¹ Ph.D. Candidate of Linguistics, Islamic Azad University of Isfahan (Khorasgan) Branch, Isfahan, Iran

² Ph.D. Candidate of TEFL, Hakim Sabzevari University, Sabzevar, Iran



[10.22080/iselt.2026.32544.1218](https://doi.org/10.22080/iselt.2026.32544.1218)

Received

July 27, 2026

Accepted

September 20, 2026

Available online

September 25, 2026

Keywords:

Large Language Models, Morphological Productivity, Persian Morphology, Tokenization, Wug Test.

Abstract

The Wug Test is often treated as a compact demonstration that speakers can extend morphological patterns beyond memorized words. Its apparent simplicity, however, conceals a difficult inferential problem when the same task is administered to large language models (LLMs): a formally correct nonce-word response may arise from productive abstraction, lexical analogy, subword completion, prompt-local imitation, or contamination from training data. This article develops a falsifiable comparative protocol for distinguishing these mechanisms across adult native speakers, children, second language learners, and contemporary LLMs. English is paired with Persian, whose productive and restricted plural systems, derivational resources, script, and uneven representation in model training create a stringent cross-linguistic test. The design orthogonally manipulates phonotactic probability, lexical-neighborhood density, morphological family support, tokenizer alignment, script, prompt regime, and inflectional or derivational complexity. Responses are evaluated not only for target accuracy, but also for error topology, calibration, prompt stability, and transfer to held-out morphological families. Hierarchical models estimate group-specific sensitivities while contamination audits and repeated model sampling constrain claims about novelty. The central argument is methodological: human-like aggregate accuracy is neither necessary nor sufficient evidence of human-like productivity. A model warrants that attribution only when it reproduces the structured generalizations, dissociations, and error distributions that make nonce-word evidence theoretically informative in human speakers. If implemented, the protocol would provide a reproducible standard for evaluating morphological claims about LLMs and for designing multilingual benchmarks that distinguish robust generalization from prompt- or tokenizer-dependent success.

* **Corresponding Author:** Hamed Zarabi, Ph.D. Candidate of TEFL, Hakim Sabzevari University, Sabzevar, Iran, Email: hamed.zarabi@hsu.ac.ir



Copyright © 2027 The Author(s). Published by the Faculty of Persian Literature and Foreign Languages at the University of Mazandaran. This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/) license.

1. INTRODUCTION

Few experimental objects in linguistics have acquired the iconic compactness of the wug. A child sees an unfamiliar creature, hears that it is “a wug,” and is then asked to name two of them. The elicited plural is conventionally read as evidence that the child has done something more consequential than repeat a stored expression: the child has extended a pattern to an item for which no inflected token has previously been supplied. Berko’s formulation was deliberately methodological rather than theatrical. Her premise was that “to test for knowledge of morphological rules, we use nonsense materials” (Berko, 1958, p. 150), and the experiment’s force lay in the disciplined removal of one evidential shortcut—direct memorization of the target word-form pair. The resulting paradigm became a durable point of contact among acquisition, psycholinguistics, morphology, phonology, and computational modeling because it compressed a large theoretical question into an observable decision: what counts as evidence that a linguistic system can make a licensed form that it has not merely retrieved? Yet the wug test has never been a transparent window onto a single faculty. Human responses reflect phonotactic probability, lexical neighborhoods, type and token frequency, morphophonological opacity, semantic interpretation, literacy, task demands, and developmental history (Albright & Hayes, 2003; Alegre & Gordon, 1999; Bybee & Moder, 1983; Hay & Baayen, 2005; Keuleers & Brysbaert, 2010; Marchman & Bates, 1994; Marcus et al., 1992; Prasada & Pinker, 1993). The test is, therefore, valuable not because every correct answer is a direct signature of a symbolic rule, but because carefully structured contrasts can reveal what information a speaker’s generalization tracks. In Berko’s own summary, “the picture that emerged was one of consistency, regularity, and simplicity” (1958, p. 177); subsequent research has shown that this consistency is frequently gradient, lexically conditioned, and distributed across representational levels rather than reducible to a single all-or-none operation (Baayen, 1992, 1993; Blevins, 2006, 2016; Booij, 2010; Bybee, 1995; Ernestus & Baayen, 2003).

The arrival of generative large language models (LLMs) appears to restore the original experimental drama in a technologically intensified form. A model is shown a nonce stem, given a grammatical context or a set of examples, and asked to produce an inflected or derived word. Correct outputs are then described as evidence that the model has “learned morphology,” “acquired rules,” or attained a human-like capacity for linguistic productivity. This inference is tempting because the output resembles the human response and because LLMs can generate strings that are absent from a particular prompt. It is nevertheless underdetermined. A model’s answer can be licensed by at least five distinguishable mechanisms: a productive representation of a morphological relation, analogy to one or more lexical neighbors, recombination of memorized subword units, copying of a prompt-local transformation, or retrieval of a form, stimulus, or close analogue encountered during pretraining. These mechanisms are not mutually exclusive, and their outputs can converge on the same orthographic string. Recent studies have supplied important evidence that the convergence is unstable. Weissweiler et al. (2023) found substantial cross-linguistic and prompt-conditioned variation in ChatGPT’s wug performance, including real-word substitutions and marked weakness in Tamil, Anh et al. (2024) showed that performance across six languages decreased with morphological information complexity, Ismayilzada et al. (2025) separated morphological productivity from systematicity in Turkish and Finnish and found large gaps between models and human annotators, and Lerner and Yvon (2025) demonstrated that Byte Pair Encoding (BPE) segmentation can make prefixation substantially more difficult than suffixation. These results are not evidence that LLMs simply lack morphology. They show that the construct being measured is entangled with tokenization, prompt design, language coverage, and decoding. Claims of equivalence are, therefore, premature when they rest on answer accuracy

alone- a conclusion consistent with Weissweiler et al.'s characterization of broad human-likeness claims as "premature and misleading" (2023, p. 6508).

The inferential asymmetry between human and model wug testing is deeper than the familiar distinction between biological and artificial learners. Human participants arrive with opaque, but developmentally constrained histories of exposure, interaction, embodiment, literacy, and correction. Researchers cannot enumerate every word a speaker has heard, but they can make principled assumptions about communicative experience, age of acquisition, and the absence of direct exposure to newly coined laboratory items. LLMs arrive with training corpora that may contain test templates, nonce strings, online descriptions of the Wug Test, morphological tables, dictionaries, and near-duplicate examples; their tokenizers may also decompose an apparently novel word into highly familiar pieces. The same nonce item can, thus, be novel as a whole string, non-novel as a token sequence, and contaminated as an experimental schema. Furthermore, an adult or child normally contributes one response in a stable interaction, whereas an LLM defines a conditional distribution from which many outputs can be sampled under adjustable temperature, system instructions, and context windows. Treating one greedy output as the computational counterpart of one human answer collapses a stochastic system into an arbitrary point estimate. Conversely, averaging hundreds of model generations and comparing that distribution with one response per person gives the model an evidential advantage. A valid comparison must, therefore, match units of observation, distinguish system-level from sample-level variability, and preserve the difference between uncertainty within a model and heterogeneity across people (Barr et al., 2013; Brysbaert & Stevens, 2018; Jaeger, 2008; Westfall et al., 2014).

A further difficulty concerns what morphology is taken to be. Hockett defined a model of grammatical description as "a frame of reference within which an analyst approaches the grammatical phase of a language" (1954, p. 210). The apparently neutral act of scoring a nonce response presupposes such a frame. An item-and-process analysis may treat the response as the application of an operation to a base, a word-and-paradigm account may treat it as the realization of a cell in a structured system, a constructional account may treat it as the extension of a form-meaning schema, and exemplar or discriminative approaches may derive it from similarity-weighted mappings over lexical experience (Aranovich, 2026; Baayen & Lieber, 1991; Blevins, 2006; Booij, 2010; Bybee, 1985; Clahsen, 1999; Hockett, 1954; Pinker & Prince, 1988; Yang, 2016). LLM evaluation often avoids this theoretical choice by using exact match, but exact match does not eliminate a theory- it silently equates morphological competence with production of a reference string. This is particularly problematic where multiple outputs are licensed, where register and animacy condition competing plural markers, or where a derivation is structurally possible, but pragmatically anomalous. A human speaker who rejects a nonce derivative may display lexical or semantic sensitivity rather than deficient productivity; a model that produces every formally concatenable derivative may display morphotactic permissiveness rather than linguistic creativity. The proposed comparison, therefore, treats productivity as a profile of constrained extension, not a count of correct suffixes.

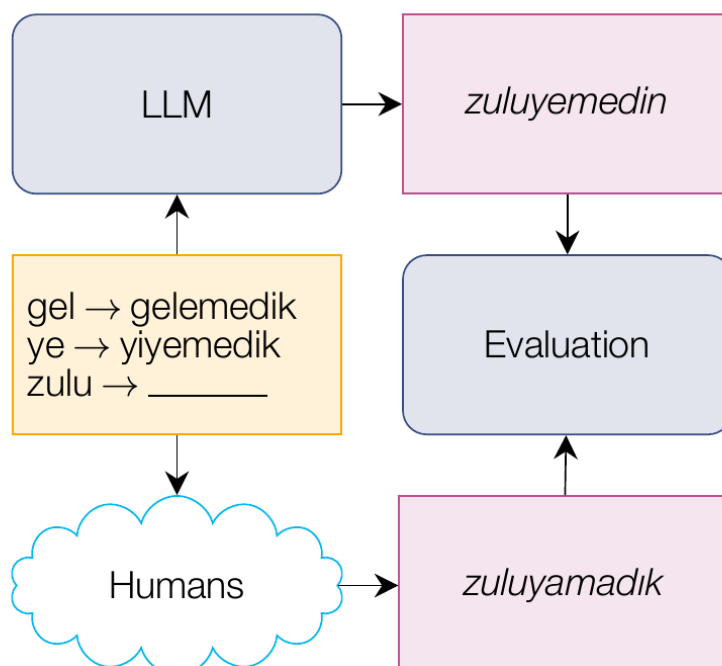


Figure 1. Experimental paradigm for multilingual LLM-human nonce inflection. Reproduced from Weissweiler et al. (2023, Figure 1), CC BY 4.0.

The present article advances three linked claims. First, a wug response is theoretically informative only relative to a contrast set that separates productive extension from lexical, distributional, and tokenization-based alternatives. Second, human–LLM comparison requires explanatory parity: the same nuisance variables, response options, and uncertainty measures must be modeled for both populations, while variables unique to one system, such as tokenizer boundaries or prompt sensitivity, must be tested rather than ignored. Third, a genuinely cross-linguistic design must include a language that disrupts assumptions built into English-centered evaluation. Persian is especially revealing. Its productive colloquial/general plural marker *-hâ* competes with the more lexically and semantically restricted *-ân*, plural marking interacts with definiteness and noun-phrase structure, comparative and superlative formations use transparent suffixes, and the Perso-Arabic script omits some phonological information while inducing tokenization patterns very different from Latin-script English (Ghameshi, 1997, 2003; Lazard, 1992; Mahootian, 1997; Sagot & Walther, 2010; Samvelian, 2012). Persian also permits a script manipulation unavailable in English: identical lexical and morphological materials can be presented in native script and controlled transliteration, enabling direct separation of linguistic generalization from script-specific subword access. The recent development of a Persian Wug Test for children confirms the feasibility and local relevance of nonce morphology in this language, although that instrument and its psychometric claims require independent replication (Mehranirad et al., 2025).

This study is formulated as a complete, preregisterable comparative protocol rather than as an article containing invented participant or model results. It asks five questions.

RQ1: Do adult native speakers, children, second-language learners, and LLMs converge on the same productive exponents when lexical-neighborhood support and phonotactic probability are controlled?

RQ2: Are group differences reducible to aggregate accuracy, or do they persist in error distributions, confidence, item sensitivity, and stability across elicitation contexts?

RQ3: How strongly do tokenizer alignment, script, and prompt regime predict LLM responses relative to the linguistic predictors that organize human responses?

RQ4: Can LLMs generalize across held-out morphological families rather than reproduce transformations demonstrated in context?

RQ5: What pattern of convergent evidence would justify the restricted claim that a model exhibits human-like morphological productivity?

The protocol's answer to the last question is intentionally demanding: similar mean accuracy is insufficient. Human-likeness requires matched direction and approximate magnitude of sensitivity to linguistic variables, comparable treatment of licensed competitors, a partially aligned error topology, robustness to prompt paraphrase, and transfer that survives removal of close lexical and subword analogues.

The implications extend beyond a single morphological benchmark. For theories of productivity, the protocol specifies the behavioral evidence that can and cannot support claims about abstraction, for multilingual language-model evaluation, it identifies the controls needed before cross-language score differences are interpreted as differences in competence, and for language research, assessment, and pedagogy, it cautions against treating fluent nonce-word completion as stable morphological knowledge without evidence of transfer, restraint, and robustness. The resulting framework is intended to make positive, negative, and dissociative findings equally interpretable rather than presuming either human equivalence or model failure in advance.

2. LITERATURE REVIEW

From nonce-word completion to a theory of productivity

Morphological productivity is not the unrestricted possibility of attaching an exponent to any base. It is the graded availability of a morphological pattern for extending the lexicon or realizing a paradigm under structural, phonological, semantic, and pragmatic constraints. Corpus measures such as type frequency, hapax-conditioned productivity, and category-conditioned potential productivity capture different aspects of this availability (Baayen, 1992, 1993; Baayen & Renouf, 1996; Bauer, 2001; Plag, 1999), while experimental measures reveal how speakers treat novel bases under controlled conditions. The two traditions should not be collapsed. A process may be productive in corpora because it supports many low-frequency formations, but still be disfavored for a phonotactically unusual nonce base; conversely, participants may extend a pattern in an explicit metalinguistic task that is rare in spontaneous language. The classical English past-tense debate made this distinction unavoidable. Dual-route accounts treated regular past tense as rule-governed and irregulars as lexically represented, whereas connectionist and analogical accounts showed how distributional structure could generate graded generalization without a categorical symbolic default (Albright & Hayes, 2003; McClelland & Patterson, 2002; Pinker & Prince, 1988; Prasada & Pinker, 1993; Rumelhart & McClelland, 1986; Ullman, 1999). The empirical legacy is not a simple victory for either “rules” or “analogy.” Speakers display default-like behavior, frequency effects, neighborhood effects, and morphophonological sensitivity simultaneously. A contemporary wug design must, therefore, manipulate the information that supports each account rather than use a single novel word as a philosophical referendum.

The evidential unit should be the response profile over systematically related items. Suppose a participant pluralizes an English nonce noun ending in /s/ with /ɪz/. The response is compatible with the phonological allomorphy, analogy to buses and glasses, or a memorized orthographic heuristic. To discriminate among these accounts, the item set must vary the final

segment, spelling, neighborhood density, and morphological family support while preserving the intended feature contrast. Analogous logic applies to derivation. A nonce base may accept -ness because the suffix is broadly productive, because the base resembles a dense family of adjectives in -y, or because the prompt has just demonstrated the same transformation. Phonotactic well-formedness is itself learned and gradient (Daland et al., 2011; Hayes & Wilson, 2008), and morphological alternations can remain grammatically active even when their lexical distributions are sparse (Gouskova & Becker, 2013; Hayes & Londe, 2006). Accordingly, the proposed stimulus space crosses phonotactic probability with lexical-neighborhood density and morphological support. High-probability/low-neighborhood items test extension beyond close analogues; low-probability/high-neighborhood items test whether lexical similarity overrides phonological markedness; and low-low items provide the strongest, though not pure, challenge to generalization. No single cell is privileged as the “true” wug condition. The diagnostic information lies in the slopes and interactions.

Developmental and learner populations are essential because adult native speakers are not a neutral human baseline. Children’s overregularizations and age-related changes reveal how productivity emerges from unequal exposure and developing phonological representations (Kuczaj, 1977; Marchman & Bates, 1994; Marcus et al., 1992). Adults contribute entrenched lexical knowledge, literacy, metalinguistic strategies, and sensitivity to register. Second language (L2) learners introduce a different decomposition of experience: target-language type frequency interacts with proficiency, instruction, and transfer from the first language. A comparison restricted to highly educated adult English speakers and English-dominant LLMs would, therefore, confound human morphology with one mature literate population. The present design uses three human groups per language—native-speaking adults, native-speaking children, and adult L2 learners—and treats group as an explanatory factor rather than a ranking. Children are not expected to be miniature adults, and L2 learners are not expected to approximate native speakers uniformly. The question is whether each population’s response structure can be described by the same linguistic predictors and whether LLM behavior occupies any coherent location in that multidimensional space.

What current LLM Wug Tests establish—and what they leave open

The emerging literature has decisively improved on anecdotal prompting by constructing multilingual datasets, recording prompt conditions, and comparing model outputs with human judgments. This work builds on a mature computational-morphology tradition in reinflection, character-level transduction, data augmentation, cognitively plausible minimal generalization, and copying-bias control (Cotterell et al., 2016; Kann et al., 2017; Kann & Schütze, 2016; Kirov & Cotterell, 2018; Liu & Hulden, 2022; Makarov & Clematide, 2018; Silfverberg et al., 2017; Vylomova et al., 2020; Wilson & Li, 2021; Wu et al., 2021). Weissweiler et al. (2023) tested reinflection and nonce inflection in English, German, Tamil, and Turkish, showing that prompt length, number of demonstrations, and language materially affected performance. Their experimental paradigm, reproduced in Figure 1, makes a crucial advance: model responses are not treated as self-validating, but are compared with human responses or ratings; yet the same figure also exposes a residual asymmetry. The LLM receives examples in a textual prompt, whereas human annotators may rely on a lifetime of exposure and a task explanation. The number and structure of demonstrations can induce a local mapping that is neither equivalent to child acquisition nor irrelevant to it. Figure 2 shows that results changed across zero-shot, one-shot, and few-shot settings, with the direction and size of the change varying by language. Prompting is, thus, not a superficial interface choice; it is part of the learning evidence available at test time. Any

claim about model productivity must specify whether it concerns pretrained competence, in-context adaptation, or their interaction. The methodological importance of prompt specification is also evident in human-AI comparisons outside morphological testing. [Bozorgian and Rahimi \(2025\)](#) pilot-tested alternative prompts before adopting a standardized open-ended prompt intended to sample ChatGPT-4o’s comparatively unengineered response behavior under realistic conditions. This procedure underscores why prompt wording, demonstrations, and formatting must be treated as documented components of the elicitation design.

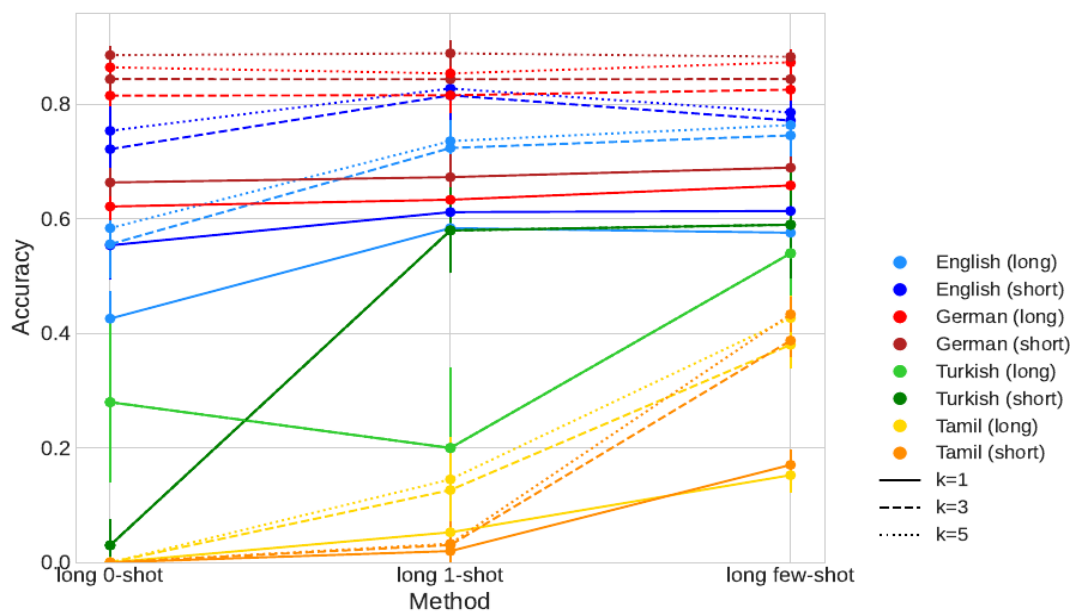


Figure 2. Accuracy across prompt scenarios, formats, languages, and values of k. Reproduced from [Weissweiler et al. \(2023, Figure 2\)](#), CC BY 4.0.

[Anh et al.’s \(2024\)](#) multilingual Wug Test extends this logic by translating and validating stimuli in Vietnamese, French, Spanish, Romanian, Portuguese, and German, collecting human responses, and asking speakers to rate model answers. Their workflow, shown in [Figure 3](#), is exemplary in treating translation as an expert-mediated experimental operation rather than assuming that English prompts are cross-linguistically neutral. Their analyses further indicate that morphological information complexity predicts performance more strongly than raw training-data quantity in some conditions. [Figure 4](#) summarizes the reported direction of these relationships: higher I-complexity was associated with reduced binary accuracy, whereas E-complexity and training data showed weaker patterns. This finding is theoretically important because it suggests that “low-resource” and “morphologically rich” are not interchangeable explanations. Nevertheless, typological aggregate measures cannot identify the mechanism behind a response to a particular nonce form. A language can be globally complex while a tested construction is transparent, or globally less complex while a specific allomorphy is lexically restricted. The proposed design, therefore, moves from language-level complexity scores to item-level contrasts whose morphological, phonological, and tokenization properties can be estimated directly.

For each of six languages, Vietnamese, French, Spanish, Romanian, Portuguese, German:

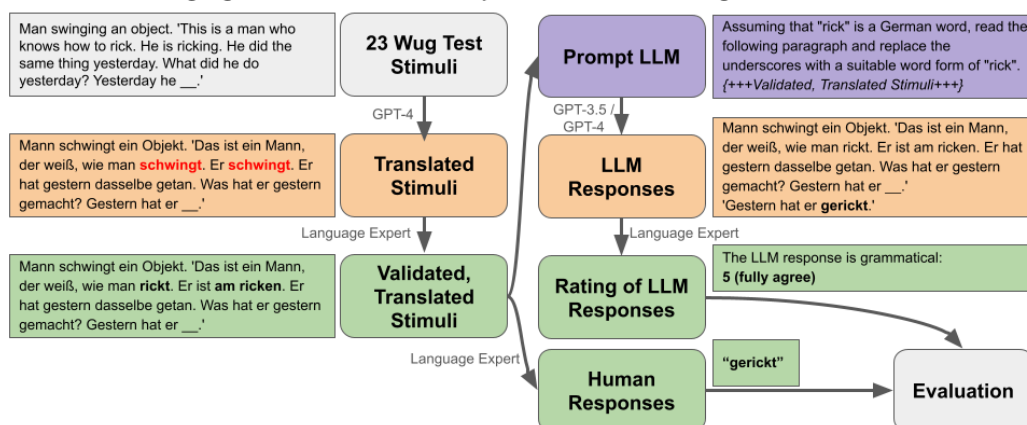


Figure 3. Multilingual translation, validation, prompting, human-response, and evaluation workflow. Reproduced from Anh et al. (2024, Figure 1), CC BY 4.0.

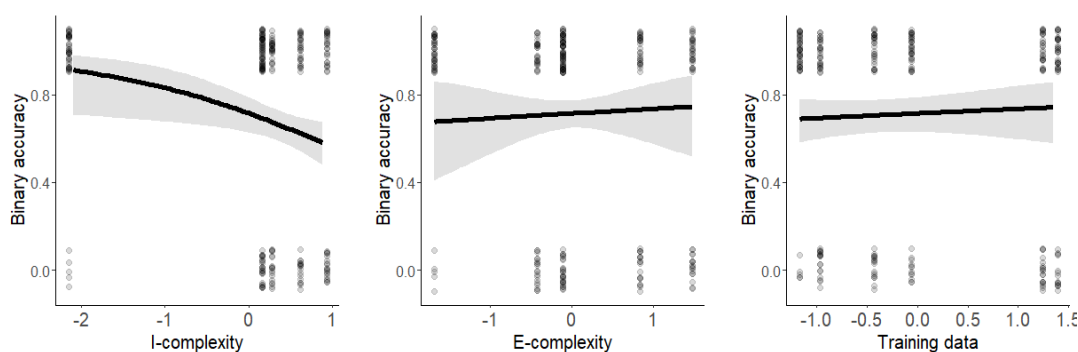
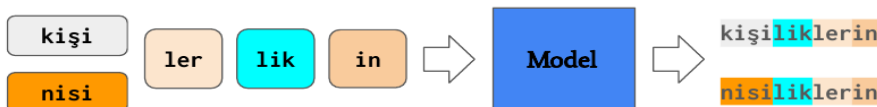


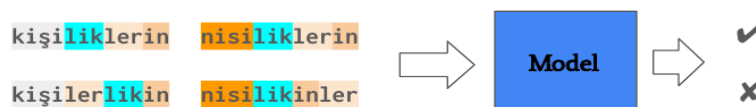
Figure 4. Relationships between binary accuracy and I-complexity, E-complexity, and training-data percentage. Adapted from Anh et al. (2024, Figure 2), CC BY 4.0.

Ismayilzada et al. (2025) introduced an especially useful distinction between productivity and systematicity. Their productivity task requires generation of an inflected form for an out-of-distribution base, while their systematicity task evaluates whether models apply morphological combinations coherently across related inputs (Figure 5). This distinction addresses a central weakness in single-item wug testing: a model can succeed on one form, yet fail to preserve the same analysis across a paradigm. Their Turkish and Finnish results (Figure 6) show that human performance remained substantially stronger, particularly as morpheme sequences became more complex. The implication is not that concatenative morphology is inaccessible to LLMs; rather, local success does not guarantee paradigm-level compositionality. The present protocol adopts this distinction, but adds a third dimension, contrastive restraint: whether a system avoids applying a formally available pattern where lexical, semantic, or morphosyntactic restrictions make it inappropriate. Productivity without systematicity is brittle; systematicity without restraint is overgeneration. Human-like morphology requires all three.

Morphological Productivity



Morphological Systematicity



□ ID word □ OOD word □ Bound Morphemes

nisi = kişi = person
nisiliklerin = of the personalities

Figure 5. Productivity and systematicity tasks illustrated with Turkish morphemes. Reproduced from Ismayilzada et al. (2025, Figure 1), CC BY 4.0.

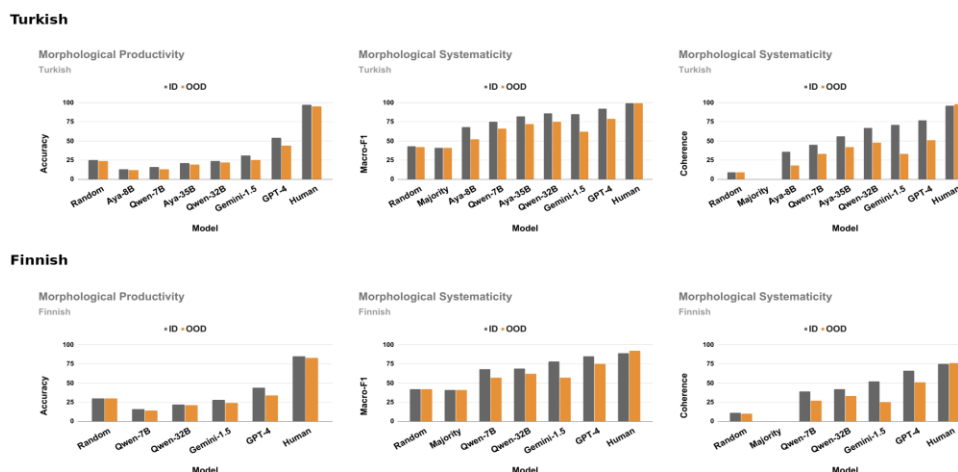


Figure 6. Morphological productivity and systematicity results for Turkish and Finnish. Adapted from Ismayilzada et al. (2025, Figures 2-3), CC BY 4.0.

Two recent studies sharpen this methodological problem. Pantelidou et al. (2026) compared six models across Catalan, English, Greek, and Spanish and found that community size and the associated availability of linguistic resources predicted Wug Test accuracy more strongly than aggregate grammatical complexity. Their result reinforces the need to separate construction-level linguistic difficulty from exposure. Xiong and Wu (2025) further showed, in morphological-learning tasks, that LLMs did not reproduce human benefits from interleaved and spaced practice uniformly, indicating that superficially similar learning gains can be organized by different sensitivities to input scheduling. Together, these findings favor item-level controls and repeated elicitation conditions over language-level accuracy rankings.

A separate line of work demonstrates why tokenization must be treated as a causal variable rather than a post hoc explanation. Standard subword tokenizers optimize compression or likelihood, not morphological transparency. A nonce word may be segmented at a morpheme boundary, fragmented within a morpheme, or represented partly as bytes; different scripts and model vocabularies produce different segmentations of the same linguistic structure. [Bostrom and Durrett \(2020\)](#) showed that common BPE tokenization can be suboptimal for pretrained language models, [Rust et al. \(2021\)](#) documented cross-lingual disparities in tokenizer quality, and [Truong et al. \(2024\)](#) found extensive under- and over-segmentation of English affixal negation across major models ([Figure 7](#)). [Lerner and Yvon \(2025\)](#) then showed a striking prefix–suffix asymmetry: under BPE conventions, word-initial marking can impede prefix generation even when suffixation succeeds, while morphologically informed segmentation largely removes the gap ([Figure 8](#)). These results make a strong methodological prediction for Persian. If native-script and transliterated items differ sharply for LLMs after linguistic form is held constant, but not for literate bilingual humans to the same degree, the difference is evidence of representational access, not of Persian morphology in the abstract.

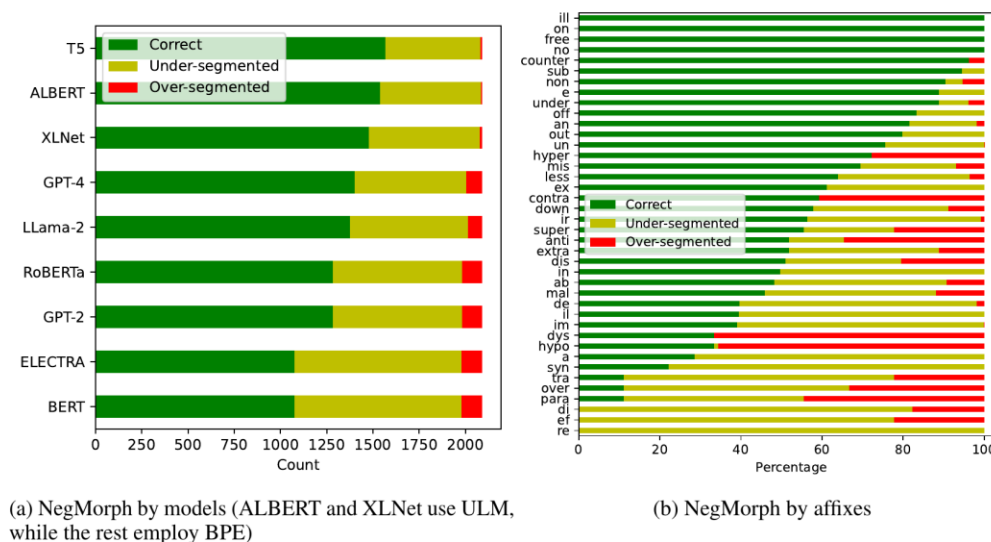


Figure 7. Correct, under-segmented, and over-segmented analyses of affixal negation across models and affixes. Adapted from [Truong et al. \(2024, Figure 2\)](#), CC BY 4.0.

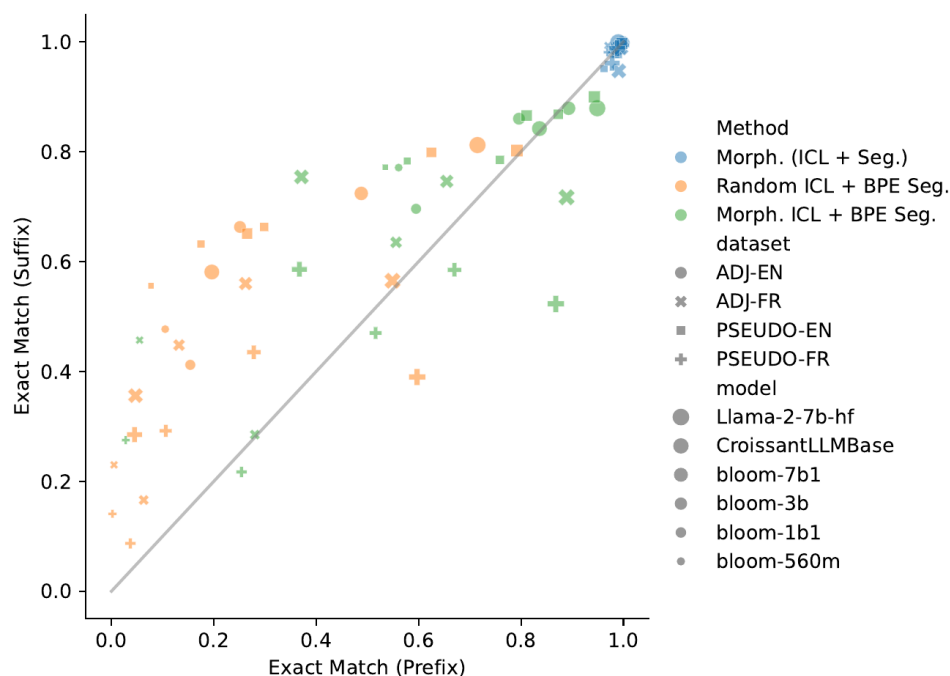


Figure 8. Exact-match performance for prefixation and suffixation under morphological and BPE segmentation. Adapted from Lerner and Yvon (2025, Figure 2), CC BY 4.0.

Lexical creativity studies further complicate the interpretation of novel output. McCoy et al. (2023) show that language models can generate novel n-grams, syntactic structures, and morphological combinations rather than merely copying complete passages, but they also distinguish novelty from the originality of the underlying pattern. Ciaccio et al. (2025) similarly evaluate lexical proficiency through reverse dictionary, definition modeling, and contextual use of lexicalized words, neologisms, and nonce forms (Figure 9). The broader lesson is that generation beyond a training string is a necessary, but weak, condition for productivity. A system can generate an unattested word by recombining familiar tokens in a way that is semantically incoherent; it can also reproduce a familiar morphological relation in a genuinely new lexical context. The protocol, therefore, treats string novelty, morphological licensing, semantic fit, and distributional distance as separate annotations. Contamination audits search exact stimuli and templates where possible, but they do not claim to prove absence from inaccessible corpora. Instead, stimulus families are generated after model training cut-offs when known, include cryptographic preregistration timestamps, and are evaluated at multiple distance levels from attested lexemes. The strongest evidence comes from patterns that persist across many newly generated families, not from a claim that any single nonce string is metaphysically unseen.

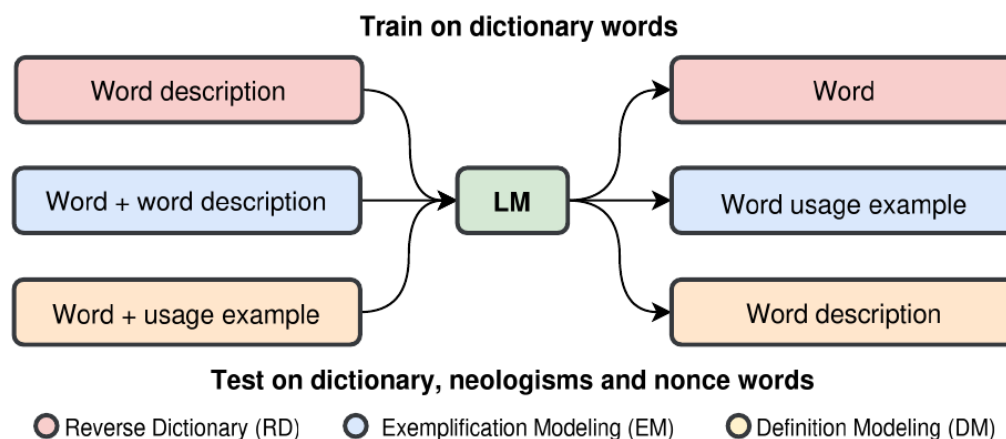


Figure 9. Framework for evaluating lexical proficiency through reverse dictionary, exemplification, and definition modeling. Reproduced from Ciaccio et al. (2025, Figure 1), CC BY 4.0.

Taken together, previous research has established three points: LLMs can extend some inflectional and derivational patterns to nonce forms; their success varies across languages, prompt regimes, and morphological structures; and tokenization and exposure can reorganize apparently linguistic performance. The remaining limitation is methodological. No existing comparison simultaneously controls construction-specific morphology, phonotactic probability, lexical neighborhoods, morphological-family support, tokenization, script, prompt information, exposure risk, and matched human response distributions. Aggregate exact-match results, therefore, cannot distinguish productive abstraction from local analogy, in-context copying, or representational convenience. The present protocol addresses this limitation by crossing these factors within English and Persian, separating production, systematicity, and restraint, auditing contamination, and requiring multivariate equivalence across response probability, predictor sensitivities, error topology, and robustness.

Why English and Persian form a diagnostic pair

English is indispensable because the original Wug Test and much subsequent productivity research provide well-characterized contrasts in plural allomorphy, past tense, third-person singular, possessive marking, comparative formation, and derivation. It is also the language in which contemporary LLMs generally have the greatest training exposure, making high performance plausible, but contamination and lexical support difficult to exclude. Modern Persian contributes a different diagnostic geometry. It is predominantly suffixing in the selected nominal and adjectival domains, but broadly productive patterns coexist with restricted, lexically stratified, and register-sensitive alternatives (Lazard, 1992; Mahootian, 1997). The comparison is, therefore, not English-as-simple versus Persian-as-complex; it contrasts phonologically conditioned allomorphy with distributionally and semantically conditioned competition.

In the nominal system, general plural *-hâ* is available across a broad range of human and nonhuman nouns, whereas *-ân* is a restricted and lexically conditioned competitor with stronger human/animate and formal or literary associations; plural choice also interacts with the syntax and interpretation of nominal expressions (Ghomeshi, 2003). In the adjectival system, *-tar* and *-tarin* realize comparative and superlative degree only with semantically gradable bases, permitting transparent concatenation to be separated from semantic restraint. In derivation, *-gar* primarily

forms occupation or agent nouns from compatible nominal or activity-denoting bases, whereas *-ande* attaches to a verbal present stem to form agentive or participial nouns and adjectives; neither suffix licenses unrestricted attachment (Lazard, 1992; Mahootian, 1997). Persian complex predicates and light-verb constructions are excluded from the primary task because their lexical and event-structural dependencies would confound a first comparison of word-level productivity (Karimi-Doostan, 2005). *Ezafe* is recorded as a contextual factor, but is not treated as an overt production target because it is often absent from ordinary Persian orthography despite its syntactic and phonological role (Ghameshi, 1997; Samvelian, 2012).

The script manipulation is central. Each Persian nonce item is constructed in phonemic transliteration, validated by native speakers, and then presented in two matched orthographic versions: standard Persian script and a preregistered Latin transliteration that preserves segmental distinctions relevant to the target morphology. Human Persian readers complete both versions in counterbalanced blocks; children complete native-script items only unless their literacy profile supports the transliteration condition. LLMs receive both forms under matched prompts. For each model, the study records token count, token identities when available, morpheme-boundary alignment, within-morpheme fragmentation, and whether suffixes are represented consistently across stems. This yields a direct test of whether a model’s productivity is invariant under representational recoding. Human responses may also show script effects because orthography shapes lexical access, but the prediction is not that humans are script-independent. It is that human variation will be better explained by literacy and phonological well-formedness, whereas model variation will show an additional, large contribution from tokenizer alignment and vocabulary coverage (Kudo, 2018; Kudo & Richardson, 2018; Lerner & Yvon, 2025; Rust et al., 2021; Sennrich et al., 2016; Truong et al., 2024).

Table 1 specifies the construction-level contrasts used in the Persian arm and distinguishes a target from a licensed or illicit competitor before scoring decisions are made.

Table 1. Persian morphological domains and diagnostic contrasts

Morphologic al domain	Target construction	Competitor construction	Linguistic restriction	Reason for inclusion
General nominal plural	Noun + <i>-hâ</i>	Noun + <i>-ân</i>	<i>-hâ</i> is broadly productive; <i>-ân</i> is lexically restricted and favors human/animate and formal or literary contexts.	Tests extension of a productive default and avoidance of a restricted competitor.
Restricted nominal plural	Prevalidated noun + <i>-ân</i>	Noun + <i>-hâ</i>	<i>-hâ</i> may remain grammatical but can mismatch the intended lexical or register preference.	Tests graded competition and treatment of a licensed alternative rather than binary accuracy.
Comparative degree	Gradable adjective + <i>-tar</i>	Bare adjective or analytic <i>bištar</i> + adjective	The base must support a scalar interpretation; some lexical classes prefer an analytic strategy.	Tests transparent suffixation under semantic selection.
Superlative degree	Gradable adjective + <i>-tarin</i>	Comparative/peri phrastic superlative	Gradability and attributive or predicative syntax constrain the realized form.	Tests comparative- superlative paradigm coherence.
Nominal-base agentive	Compatible base + <i>-gar</i>	<i>-či</i> or <i>-kâr</i> where independently licensed	Base semantics, register, and lexicalization restrict attachment; <i>-gar</i> is not licensed on every noun.	Tests derivational productivity together with restraint.
Deverbal agentive or participle	Present stem + <i>-ande</i>	Nominal-base <i>-gar</i> or bare stem	Requires a licensed present stem and a compatible agent or process interpretation; spelling adjustments are preregistered.	Tests stem selection, category sensitivity, and morphographic consistency.

3. METHODS

Design and populations

The design is a crossed, multilingual nonce-word production experiment with parallel human and model arms. The human arm targets, in each language, 120 adult native speakers, 80 native-speaking children distributed across two preregistered age bands, and 120 adult L2 learners divided by independently measured proficiency. These numbers are planning targets rather than claims about completed recruitment. Final sample sizes will be determined by simulation-based power analysis for the smallest theoretically consequential interaction— group by morphological support— using conservative variance estimates from a pilot that is excluded from confirmatory analysis (Brysbaert & Stevens, 2018; Lakens, 2022; Westfall et al., 2014). The hypotheses, exclusions, and analysis code are preregistered to constrain undisclosed analytic flexibility (Munafò et al., 2017; Nosek et al., 2018; Simmons et al., 2011). Adults provide informed consent; children provide assent, and a parent or guardian provides consent. Tasks are designed to avoid diagnostic labeling, and no individual child score is returned as a clinical judgment. The model

arm includes at least six model families representing proprietary and openly documented systems, with exact model identifiers, access dates, context-window settings, tokenizer versions, and generation parameters archived. Model inclusion is frozen at preregistration; later systems may be reported as a labeled exploratory extension rather than silently substituted. Because a model endpoint can change without a public version change, a reproducibility bundle stores prompts, request metadata, raw responses, and timestamps.

Morphological domains and stimulus construction

The English task includes noun plural, possessive, verbal past tense, third-person singular, comparative/superlative, agentive *-er*, and nominalizing *-ness*. As specified in [Table 1](#), the Persian task includes general and restricted plural contexts for the *-hâ/-ân* contrast, comparative *-tar*, superlative *-tarin*, nominal-base agentive *-gar*, and present-stem agentive/participial *-ande*. The two derivational constructions are retained only when an independent native-speaker norming sample confirms both the intended base category and an interpretable target meaning for attested controls and the planned retention threshold is 80% agreement. Items for which a listed competitor is also licensed are not forced into binary exact-match scoring: target preference, competitor licensing, register, and semantic fit are coded separately. Each domain contains attested controls, phonotactically legal nonce bases, and adversarial nonce bases. Nonce bases are generated with language-specific phonotactic models and manually screened to exclude offensive strings, obvious proper names, high-frequency abbreviations, and close homophones. Wuggy-style generation and maximum-entropy phonotactic estimates supply candidates, while a preregistered optimization balances length, syllable structure, phonotactic probability, orthographic neighborhood, semantic cueing, and tokenizer properties ([Hayes & Wilson, 2008](#); [Keuleers & Brysbaert, 2010](#)).

[Table 2](#) summarizes the factorial logic. Within each morphological domain, items cross high versus low phonotactic probability, dense versus sparse lexical neighborhoods, high versus low morphological-family support, and aligned versus misaligned tokenization for each model class. Perfect orthogonality across model tokenizers is impossible; tokenizer alignment is, therefore, a measured, model-specific predictor rather than a universal item label. A subset of stems is constructed so that the full nonce form is a single token in one model and multiple tokens in another, while another subset matches token count, but differs in whether the morpheme boundary is respected. For Persian, script is crossed within item. For English, an orthographic manipulation contrasts phonologically matched spellings where this can be done without introducing implausible grapheme sequences. Critical items are organized into families so that one stem appears in only one target construction for a participant or model session, preventing direct within-session recycling. Parallel lists rotate stems across constructions between participants.

Table 2. Core manipulations and their diagnostic interpretation

Manipulation	Human diagnostic	LLM-specific diagnostic	Principal rival explanation constrained
Phonotactic probability	Sensitivity to language-specific well-formedness	Sensitivity independent of token frequency	Unstructured suffix copying
Lexical-neighborhood density	Analogy and lexical entrenchment	Retrieval of close subword/word patterns	Pure categorical rule account
Morphological-family support	Type-based productivity	Distributional support in model representations	Single-neighbor analogy
Tokenizer boundary alignment	No direct analogue; orthographic parsing effects	Morpheme access under subword segmentation	Linguistic competence inferred from token convenience
Script (Persian)	Literacy and lexical-access effects	Vocabulary and segmentation dependence	Language-level deficit explanation
Prompt regime	Task interpretation and local priming	In-context learning and instruction sensitivity	Stable pretrained competence
Held-out family	Transfer beyond demonstrated exponents	Compositional generalization	Prompt-local transformation copying

Procedures for human participants

Human testing uses a picture-supported elicitation format for children and a matched sentence-completion format for adults. Every critical trial introduces the nonce base in an uninflected or minimally inflected context, confirms comprehension through a non-morphological selection question, and elicits the target in a new context. Audio presentation accompanies text for adults in the primary condition so that orthographic decoding does not define the task; children receive audio plus pictures, with text included only in age-appropriate literacy blocks. Responses are recorded and transcribed independently by two trained annotators. A delayed recognition check asks whether participants believe they had known the word before the experiment; affirmative or uncertain responses trigger a sensitivity analysis, but not automatic deletion. L2 learners complete standardized vocabulary and grammar measures, a language-history questionnaire, and a brief phonological discrimination task. Confidence is collected from adults on a four-point scale after production; children use a pictorial certainty scale validated in piloting. The task contains filler trials that require semantic choices rather than morphology, reducing the salience of suffix manipulation and detecting indiscriminate pattern application.

Procedures for language models

Models receive prompts that mirror the information available to human participants without pretending that the modalities are identical. Three preregistered regimes are used: zero-demonstration instruction, one structurally relevant demonstration with a different exponent, and four demonstrations balanced across target and competitor patterns. Demonstrations never use the same stem family as a critical item. Each regime has eight paraphrased prompt templates whose semantic content is held constant; template is a random factor. Models are tested at deterministic decoding where available and at a moderate sampling temperature, with 20 independent samples per item–template combination. The first orthographic word or explicitly delimited answer is scored as the primary response; explanations are stored, but excluded from scoring because chain-of-thought text is neither required nor treated as privileged evidence. When a model refuses,

comments on the word's fictional status, or supplies multiple alternatives, the response is coded rather than repaired. A separate forced-choice block presents target, licensed competitor, illicit form, and "none of these," but confirmatory claims are based on production; forced choice diagnoses recognition–production dissociations.

Figure 3 illustrates a multilingual workflow in which translated stimuli are expert validated and model responses are judged by humans. The present protocol extends that workflow by keeping validation separate from outcome adjudication. Native-speaker panels first determine whether each intended response is morphologically licensed and whether competitors differ in register, animacy, or semantic fit. A second, blinded panel later judges generated responses without knowing the producing group or model. This prevents reference answers from converting gradient linguistic judgments into mechanical exact match. For transparent inflectional domains, automatic normalization and morphological parsing precede manual adjudication; for derivation, semantic fit and lexical plausibility are rated separately. Inter-annotator agreement is reported for each coding layer, and disagreements are resolved by an adjudicator who sees the linguistic context, but not the system identity.

Contamination, lexical similarity, and tokenization audits

No audit can prove that a proprietary model has never encountered an item or experimental template. The protocol, therefore, avoids the binary labels *seen* and *unseen* unless training data are inspectable. It measures graded exposure risk. Exact nonce strings, prompt fragments, and inflected targets are searched in accessible web corpora and model-associated datasets; lexical and subword similarities are calculated against large word lists and newly generated item families are timestamped after model training cut-offs where those cut-offs are public. Items that occur online before the freeze date are replaced. Near-neighbor exposure is quantified by edit distance, phonological distance, embedding similarity, and shared tokenizer sequences. A model's output is flagged as a real-word substitution when it replaces the nonce stem with an attested neighbor, a pattern reported in prior LLM wug work (Weissweiler et al., 2023). Deduplication research shows why this caution matters: repeated data increase memorization and can inflate apparent generalization, while exact string novelty does not exclude abstraction from near duplicates (Carlini et al., 2023; Dodge et al., 2021; Kandpal et al., 2022; Lee et al., 2022; Razeghi et al., 2022). The audit variables enter statistical models rather than serving only as exclusions.

Response taxonomy and preregistered hypotheses

Every response is assigned to one primary and any applicable secondary categories: target-consistent realization, licensed competitor, overregularization, irregular or lexicalized analogue, illicit allomorph, null marking, stem distortion, real-word substitution, script or spacing error, language fallback, multiple-answer response, refusal/metalinguistic answer, and uninterpretable output. The category system is language-specific at the surface, but linked to cross-linguistic superordinate labels. For example, Persian *-ân* on a semantically incompatible nonce noun may be coded as an illicit restricted plural, whereas *-hâ* in a context favoring *-ân* may be a licensed general competitor with a register mismatch. Exact-match accuracy is reported for comparability with prior work, but the primary outcomes are target-or-licensed response probability, entropy over response categories, sensitivity to experimental predictors, and family-level coherence. Ismayilzada et al.'s (2025) distinction between productivity and systematicity motivates the coherence measure: a system receives high coherence only when its responses across a held-out family instantiate a consistent morphological analysis rather than a mixture of individually plausible forms.

Six hypotheses (H) are confirmatory. H1 (structured human generalization): all human groups will show effects of phonotactic probability and morphological-family support, with adults showing stronger lexical-neighborhood effects and children more frequent default-like overregularization. H2 (tokenization dependence): LLM performance will improve when morpheme boundaries align with tokenizer boundaries, even after token frequency and string length are controlled; this effect will be larger than any analogous orthographic segmentation effect in humans. H3 (script dissociation): Persian LLM performance will differ more strongly between native script and transliteration than adult human performance, and the difference will be mediated by tokenization measures. H4 (prompt instability): LLM response distributions will vary more across semantically equivalent prompt templates than human distributions vary across matched instruction paraphrases. H5 (systematicity gap): at matched target accuracy, LLMs will exhibit lower family-level coherence and more real-word substitutions than adult native speakers. H6 (non-equivalence of error topology): no model will qualify as human-like if its aggregate accuracy is similar but the rank order and conditional distribution of errors differ materially from all human groups. These hypotheses are not arranged so that humans must win. A model may exceed adults in exact match or systematicity; such a result would be theoretically important, but would not by itself identify the mechanism.

Statistical analysis and criteria for human-like productivity

The confirmatory analysis uses Bayesian hierarchical models with participant or model-run, item, stem family, prompt template, and morphological domain as varying factors. Binary target-or-licensed outcomes are analyzed with logistic models, response categories with hierarchical multinomial models, confidence and acceptability with ordinal models, and response entropy and coherence with distributionally appropriate generalized models. Fixed effects include group, language, morphological domain, phonotactic probability, neighborhood density, family support, script, prompt regime, token count, boundary alignment, exposure-risk measures, and preregistered interactions. Continuous predictors are standardized within language where their scales are not directly comparable. Models include the maximal justified random-slope structure, simplified only under preregistered diagnostics rather than significance-driven selection (Barr et al., 2013; Bates et al., 2015). Posterior predictive checks evaluate whether the fitted model reproduces item-level dispersion and rare error categories. Frequentist mixed-effects analyses are supplied as robustness checks, not as competing opportunities to select a preferred result.

Human-like is operationalized as a multivariate equivalence claim. A model must meet four gates: performance gate- its target-or-licensed response probability falls within a preregistered equivalence region around at least one human group in both languages and in held-out families, sensitivity gate- posterior contrasts for phonotactic probability, lexical neighborhood, and morphological support have the same direction and fall within broad, but finite equivalence bounds, error-topology gate- the model's conditional error distribution is closer to that human group than to a frequency-matched baseline and remains so after real-word substitutions are separated, robustness gate- the classification survives prompt paraphrase, script recoding, and removal of items with high exposure risk. Failure at one gate does not entail that the model has no morphological knowledge; it restricts the conclusion to the dimensions actually supported. This gate structure prevents a familiar inferential slide from *the model produced the expected suffix* to *the model learned morphology as humans do*. It also permits an informative opposite result: models may instantiate a nonhuman, but systematic form of morphological computation.

4. PROJECTED OUTCOMES AND THEORETICAL PREDICTIONS

Predicted diagnostic patterns

Several possible outcomes would discriminate among current accounts. If models and adults show similar effects of phonotactic probability and family support, low real-word substitution, and stable held-out transfer, then an explanation based solely on prompt imitation or token concatenation would be inadequate. If model accuracy is high, but collapses under transliteration or boundary-misaligned items, tokenization would be implicated as a bottleneck or scaffold. If children and models both overgeneralize the dominant pattern, but diverge in which items trigger overgeneralization, surface similarity would conceal distinct internal organizations. If L2 learners pattern with models in aggregate, yet show effects of proficiency and first language morphology that models lack, a generic *less competent than natives* ordering would be theoretically empty. The central comparisons are, therefore, relational. A model does not become human-like by landing between children and adults on a bar chart; it becomes a candidate for a human-like explanation only when the same manipulations reorganize its behavior in comparable ways.

The design also bears on the rule–analogy opposition. Strong lexical-neighborhood effects in humans and models would not eliminate productive abstraction, because analogical support can coexist with default extension and paradigm structure. Conversely, success on low-neighborhood items would not prove a context-free symbolic rule if tokenizer sequences or latent distributional neighbors remain available. The relevant question is whether a compact model of response behavior requires discrete morphological representations, graded similarity, or both. Word-and-paradigm and constructional perspectives predict that family-level coherence and form–meaning relations will matter beyond local affix selection (Blevins, 2016; Booij, 2010). Discriminative and usage-based approaches predict continuous effects of distribution and entrenchment (Baayen, 1992, 1993; Baayen & Renouf, 1996; Bybee, 1995). The proposed data cannot reveal the internal representation of a model by behavioral fiat, but they can make representational claims answerable to more than one output string.

A cross-linguistic result would also challenge the habit of treating English as the control language and morphologically richer languages as stress tests. English is unusually favorable to contamination and subword support because of its dominant representation in training data. Persian is not simply “harder”; it rearranges the sources of difficulty. Its general and restricted plural exponents create competition not reducible to phonological allomorphy, its script allows a representational intervention, and its derivational choices test semantic restraint. Turkish and Finnish results already show that concatenative transparency does not guarantee systematicity (Ismayilzada et al., 2025), while multilingual wug studies show that global training-data size does not exhaust cross-linguistic variation (Anh et al., 2024). The English–Persian comparison, thus, replaces a resource hierarchy with a mechanism contrast.

No empirical participant or model-output results are reported in this protocol manuscript; this section, therefore, defines the preregistered interpretive outcome space and theoretical predictions rather than presenting completed statistical findings.

5. DISCUSSION

Theoretical implications

Because the study is a protocol, the discussion concerns the inferences that each projected pattern would license rather than interpreting observed effects. The first theoretical implication is that the

rule-analogy opposition cannot be adjudicated by a correct nonce form alone. Evidence for productive abstraction would require stable transfer across held-out families together with graded sensitivity to phonotactics, lexical neighborhoods, and form-meaning compatibility. A model that preserves the target suffix, but changes its distribution when token boundaries or prompt wording change, would demonstrate conditional string production, not representation-invariant morphological productivity. Conversely, a model that meets the four equivalence gates would support a circumscribed behavioral claim even if its internal implementation differed from human cognition.

The framework also recasts human-likeness as an equivalence problem rather than a leaderboard. Children, adult L2 learners, and adult native speakers provide theoretically distinct response distributions, not progressively higher rungs on a single scale. Matching one group in mean accuracy while diverging in error topology or conditional sensitivities would be evidence of partial resemblance, whereas exceeding all human groups on exact match could coexist with nonhuman overgeneration or prompt dependence. This interpretation guards against both anthropomorphic inflation and human exceptionalism: the analysis can recognize systematic model behavior without treating output similarity as proof of shared learning mechanisms.

Methodological implications

The second implication concerns experimental control. Prompt regime, decoding, model version, tokenizer behavior, and exposure risk must be reported as properties of the test rather than relegated to implementation details. Parallel human and model arms do not require identical modalities, which would be impossible, but they do require explanatory parity: both arms should face matched informational demands, use commensurable observational units, and be evaluated against the same licensed response space. The protocol, therefore, treats model run, prompt template, item, and morphological family as crossed sources of variation and separates production from recognition. These provisions make null, positive, and dissociative outcomes interpretable.

The English-Persian design further shows why multilingual evaluation should be construction-specific. A global language-complexity score cannot explain whether a model mishandles Persian *-ân* because of semantic restriction, lexical stratification, sparse exposure, or token fragmentation. Crossing native script with transliteration and specifying the target-competitor relation for each domain makes these alternatives empirically separable. The same logic can be extended to other language pairs, provided that the L2 contributes a genuine diagnostic contrast rather than serving as a generic low-resource difficulty condition.

Practical implications

The practical implication is a stricter standard for claims about LLM support in language research, assessment, and pedagogy. Before a model is described as having acquired a morphological pattern, developers and educators should examine whether its behavior transfers to new lexical families, remains stable across semantically equivalent instructions, respects restrictions as well as productive defaults, and survives changes in script or token segmentation. The proposed response taxonomy and reporting bundle offer a reusable audit format for benchmark designers and language-technology evaluators. They do not authorize direct instructional or assessment use by themselves; instead, they identify the evidence needed to distinguish reliable generalization from fluent, but fragile completion.

Limitations

This protocol has several limitations. Human exposure histories remain only partially observable, Persian norming resources are thinner than the English ones, model providers may change systems or withhold details, child tasks differ in modality and duration, transliteration is not neutral, and derivational judgments may vary legitimately. In addition, the proposed sample targets and equivalence regions must be recalibrated after pilot estimates are available, and contamination audits cannot establish non-exposure for proprietary corpora. These constraints limit the strength of causal and cognitive interpretations even when the protocol is implemented. They are addressed through pretesting, counterbalancing, uncertainty modeling, versioned model records, sensitivity analyses, and explicit treatment of licensed variation, but they cannot be eliminated. The contribution of the present article is consequently methodological: it specifies a design and adjudication framework that can support stronger inferences once empirical data are collected.

Conclusion

The Wug Test endures because it asks a precise question of an unfamiliar form, not because a nonce word magically isolates a rule. Applied to LLMs, the test becomes more— not less—demanding. Novel strings can be generated through productive morphology, analogy, subword recombination, prompt adaptation, or exposure to near duplicates, and these routes often converge on the same answer. A defensible human–model comparison must, therefore, examine structured contrasts, error distributions, systematicity, restraint, prompt stability, tokenization, and cross-script transfer. Pairing English with Persian makes these distinctions empirically visible. The proposed protocol treats human-like productivity as a falsifiable multivariate claim rather than a rhetorical gloss on accuracy. Its standard is deliberately strict: comparable behavior must survive the removal of the conveniences that allow a model to look productive without generalizing in the way the experiment is intended to test.

Funding

No external funding was received for the preparation of this protocol manuscript.

Disclosure statement

The author(s) report no competing interests to declare.

Ethics statement

No human participants were recruited and no new participant data were analyzed for this protocol manuscript. The proposed adult and child studies require approval from the relevant institutional review board or research ethics committee before recruitment, together with informed consent, parental consent, and child assent as applicable.

Data and materials availability

No new data are reported. Upon implementation, the preregistration, de-identified response data where ethically permissible, analysis code, prompts, model metadata, stimulus-generation code, and adjudication manual should be deposited in a persistent repository. Proprietary model outputs should be shared subject to applicable terms and privacy constraints.

Declaration of generative AI use

Generative AI tools were used to assist with language drafting, literature organization, and document formatting. The authors verified the claims, quotations, page numbers, references, methodological decisions, and final wording and retain full responsibility for the manuscript.

References

- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2), 119–161.
[https://doi.org/10.1016/S0010-0277\(03\)00146-X](https://doi.org/10.1016/S0010-0277(03)00146-X)
- Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40(1), 41–61.
<https://doi.org/10.1006/jmla.1998.2607>
- Anh, D., Raviv, L., & Galke, L. (2024). Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a Wug Test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 177–188). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2024.cmcl-1.15>
- Aranovich, R. (2026). Charles Hockett’s “Two Models of Grammatical Description”: A neo-Bloomfieldian searching for alternatives. *WORD*, 72(1), 5–17.
<https://doi.org/10.1080/00437956.2026.2621441>
- Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Kluwer.
- Baayen, R. H. (1993). On frequency, transparency and productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1992* (pp. 181–208). Kluwer.
- Baayen, R. H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, 29(5), 801–843. <https://doi.org/10.1515/ling.1991.29.5.801>
- Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, 72(1), 69–96. <https://doi.org/10.2307/416794>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
<https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bauer, L. (2001). *Morphological productivity*. Cambridge University Press.
- Berko, J. (1958). The child’s learning of English morphology. *WORD*, 14(2–3), 150–177.
<https://doi.org/10.1080/00437956.1958.11659661>
- Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42(3), 531–573.
<https://doi.org/10.1017/S0022226706004191>
- Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press.
- Booij, G. (2010). *Construction morphology*. Oxford University Press.
- Bostrom, K., & Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4617–4624). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.findings-emnlp.414>

- Bozorgian, H., & Rahimi, H. (2025). Peer e-feedback and ChatGPT-4o in EFL writing: A cognitive-interpersonal comparison based on EFL students. *Journal of Language and Education*, 11(4), 51–65. <https://doi.org/10.17323/jle.2025.27195>
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), Article 9. <https://doi.org/10.5334/joc.10>
- Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins.
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5), 425–455. <https://doi.org/10.1080/01690969508407111>
- Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, 59(2), 251–270. <https://doi.org/10.2307/413574>
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying memorization across neural language models. In *International Conference on Learning Representations*.
- Ciaccio, C., Miaschi, A., & Dell’Orletta, F. (2025). Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1267–1286). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.64>
- Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22(6), 991–1013. <https://doi.org/10.1017/S0140525X99002228>
- Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The SIGMORPHON 2016 shared task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 10–22). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-2002>
- Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, 28(2), 197–234. <https://doi.org/10.1017/S0952675711000145>
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1286–1305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.98>
- Ernestus, M., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79(1), 5–38. <https://doi.org/10.1353/lan.2003.0076>
- Ghomeshi, J. (1997). Non-projecting nouns and the Ezafe construction in Persian. *Natural Language & Linguistic Theory*, 15(4), 729–788. <https://doi.org/10.1023/A:1005886709040>
- Ghomeshi, J. (2003). Plural marking, indefiniteness, and the noun phrase. *Studia Linguistica*, 57(2), 47–74. <https://doi.org/10.1111/1467-9582.00099>
- Gouskova, M., & Becker, M. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory*, 31(3), 735–765. <https://doi.org/10.1007/s11049-013-9197-5>
- Hay, J., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, 9(7), 342–348. <https://doi.org/10.1016/j.tics.2005.04.002>

- Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23(1), 59–104. <https://doi.org/10.1017/S0952675706000765>
- Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3), 379–440. <https://doi.org/10.1162/ling.2008.39.3.379>
- Hockett, C. F. (1954). Two models of grammatical description. *WORD*, 10(2–3), 210–234. <https://doi.org/10.1080/00437956.1954.11659524>
- Ismayilzada, M., Circi, D., Sälevä, J., Sirin, H., Köksal, A., Dhingra, B., Bosselut, A., Ataman, D., & van der Plas, L. (2025). Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 1, Long Papers* (pp. 1270–1305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.59>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 10697–10707). PMLR.
- Kann, K., & Schütze, H. (2016). MED: The LMU system for the SIGMORPHON 2016 shared task on morphological inflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 62–70). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-2010>
- Kann, K., Cotterell, R., & Schütze, H. (2017). Neural multi-source morphological inflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 1, pp. 514–524). Association for Computational Linguistics. <https://doi.org/10.18653/v1/E17-1049>
- Karimi-Doostan, G. (2005). Light verbs and structural case. *Lingua*, 115(12), 1737–1756. <https://doi.org/10.1016/j.lingua.2004.08.002>
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633. <https://doi.org/10.3758/BRM.42.3.627>
- Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6, 651–665. https://doi.org/10.1162/tacl_a_00247
- Kuczaj, S. A., II. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 589–600. [https://doi.org/10.1016/S0022-5371\(77\)80021-2](https://doi.org/10.1016/S0022-5371(77)80021-2)
- Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 66–75). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1007>
- Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-2012>
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), Article 33267. <https://doi.org/10.1525/collabra.33267>
- Lazard, G. (1992). *A grammar of contemporary Persian*. Mazda Publishers.

- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 8424–8445). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.577>
- Lerner, P., & Yvon, F. (2025). Unlike “likely,” “unlikely” is unlikely: BPE-based segmentation hurts morphological derivations in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5181–5190). Association for Computational Linguistics.
- Liu, L., & Hulden, M. (2022). Can a transformer pass the Wug Test? Tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 739–749). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-short.82>
- Mahootian, S. (1997). *Persian*. Routledge.
- Makarov, P., & Clematide, S. (2018). UZH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task* (pp. 69–75). Association for Computational Linguistics. <https://doi.org/10.18653/v1/K18-3008>
- Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21(2), 339–366. <https://doi.org/10.1017/S0305000900009302>
- Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4), i–178. <https://doi.org/10.2307/1166115>
- McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11), 465–472. [https://doi.org/10.1016/S1364-6613\(02\)01993-9](https://doi.org/10.1016/S1364-6613(02)01993-9)
- McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11, 652–670. https://doi.org/10.1162/tacl_a_00567
- Mehranirad, M., Arabpour, M., & Emami Neyshaburi, E. (2025). Developing a Persian version of the Wug Test to assess children’s morphological knowledge. *Journal of Child Language Acquisition and Development*, 1198–1214. <https://doi.org/10.5281/zenodo.17008773>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, Article 0021. <https://doi.org/10.1038/s41562-016-0021>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Pantelidou, N., Leivada, E., Montero, R., & Morosi, P. (2026). Community size rather than grammatical complexity better predicts large language model accuracy in a novel Wug Test. *PLOS ONE*, 21(3), e0343164. <https://doi.org/10.1371/journal.pone.0343164>
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1–2), 73–193. [https://doi.org/10.1016/0010-0277\(88\)90032-7](https://doi.org/10.1016/0010-0277(88)90032-7)

- Plag, I. (1999). *Morphological productivity: Structural constraints in English derivation*. Mouton de Gruyter.
- Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8(1), 1–56.
<https://doi.org/10.1080/01690969308406948>
- Razeghi, Y., Logan, R. L., IV, Gardner, M., & Singh, S. (2022). Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 840–854). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.59>
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). MIT Press.
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 3118–3135). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.243>
- Sagot, B., & Walther, G. (2010). A morphological lexicon for the Persian language. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 3707–3714). European Language Resources Association.
- Samvelian, P. (2012). *The syntax of the Persian noun phrase*. John Benjamins.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1715–1725). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P16-1162>
- Silfverberg, M., Wiemerslage, A., Liu, L., & Mao, L. J. (2017). Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection* (pp. 90–99). Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-2012>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Truong, T. H., Otmakhova, Y., Verspoor, K., Cohn, T., & Baldwin, T. (2024). Revisiting subword tokenization: A case study on affixal negation in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 5082–5095). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.284>
- Ullman, M. T. (1999). Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighborhood effects. *Language and Cognitive Processes*, 14(1), 47–67.
<https://doi.org/10.1080/016909699386383>
- Vylomova, E., White, J., Salesky, E., Mielke, S. J., Wu, S., Ponti, E. M., Maudslay, R. H., Zmigrod, R., Valvoda, J., Toldova, S., Tyers, F., Klyachko, E., Yegorov, I., Krizhanovsky, N., Czarnowska, P., Nikkarinen, I., Krizhanovsky, A., Pimentel, T., Hennigen, L. T., ... Hulden, M. (2020). SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 1–39).

- Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.sigmorphon-1.1>
- Weissweiler, L., Hofmann, V., Kantharuban, A., Cai, A., Dutt, R., Hengle, A., Kabra, A., Kulkarni, A., Vijayakumar, A., Yu, H., Schütze, H., Oflazer, K., & Mortensen, D. R. (2023). Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6508–6524). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.401>
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020–2045. <https://doi.org/10.1037/xge0000014>
- Wilson, C., & Li, J. S. Y. (2021). Were we there already? Applying minimal generalization to the SIGMORPHON-UniMorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 283–291). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.sigmorphon-1.31>
- Wu, S., Cotterell, R., & Hulden, M. (2021). Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1901–1907). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.163>
- Xiong, Y., & Wu, S. (2025). Do large language models learn like humans? Interleaved and spaced practice in morphological learning. *Acta Psychologica*, 260, 105518. <https://doi.org/10.1016/j.actpsy.2025.105518>
- Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.