Light or Full Verb? A Minimal-Pair Dataset for Probing Phraseological Competence in Language Models

Anonymous ACL submission

Abstract

Frequent English verbs such as *have* and *make* can function either as collocates in light-verb constructions or as full lexical predicates, as in *make a decision* vs. *make a cake*. Whether language models represent this distinction remains unclear. We introduce a large-scale controlled dataset of minimally varying English sentence series in which the same context contains the same verb in light-verb and full-verb uses. Two probing experiments show that language models differentiate between these uses even in minimal contexts and exhibit separable patterns across object types. We release the dataset, generation code, and materials as a reusable resource. The framework supports extensions to broader contexts, additional verbs, and other languages.

1 Introduction

Light-verb constructions (LVCs), also known as *support-verb constructions*, have been studied for a long time across theoretical syntax, lexical semantics, phraseology, typology, and computational linguistics (Jesperen, 1942; Grimshaw and Mester, 1988; Butt, 1995, 2010; Tu and Roth, 2011; Nagy et al., 2020; Fleischhauer et al., 2025). Despite this long tradition, they remain underexploited as diagnostic material for evaluating language models. Existing computational work has largely treated LVCs as an identification or extraction problem: given a candidate verb–noun combination, the task is to determine whether it is an LVC. This has led to supervised and multilingual detection and classification experiments (Vincze et al., 2013; Chen and Palmer, 2015; Vaidya et al., 2016; Cordeiro and Candito, 2019), as well as broader overviews of LVC and verbal-MWE identification (Constant et al., 2017; Tan et al., 2021). Yet LVC identification and controlled model probing address different questions. Identification asks whether a completed expression is an LVC; probing asks whether a model distinguishes two uses of the same verb under matched syntactic and contextual conditions. This distinction matters because a model may recognize familiar LVCs through distributional association, lexical memorization, or noun-category cues, without representing the lexical-functional contrast between a support-verb use and a full lexical use. For example, *make a decision* and *make a cake* share the same transitive frame, but only the former realizes an LVC selected by the predicative nominal base *decision*.

Minimal-pair benchmarks provide a natural framework for addressing this gap: they isolate specific linguistic contrasts while holding other sentence properties constant. Targeted syntactic evaluation tests whether models assign higher probability to the expected member of minimally different sentence pairs (Marvin and Linzen, 2018; Baroni et al., 2026); BLiMP extends this approach to 67 English grammatical phenomena (Warstadt et al., 2020); and SyntaxGym provides a reproducible framework inspired by psycholinguistic experimental design (Gauthier et al., 2020). Similar resources have been developed for Chinese and Spanish (Xiang et al., 2021; Táboas García and Wanner, 2025), with recent multilingual extensions of the BLiMP paradigm (Jumelet et al., 2025). Existing minimal-pair resources, however, mostly contrast acceptable and unacceptable sentences, whereas LVC probing requires contrasting two grammatical and natural sentences that differ in the lexical-functional status of the same surface verb.

We present an extensible seed dataset for controlled probing of English LVCs, which currently covers five highly frequent light verbs. The dataset extends the contrastive logic of minimal-pair benchmarks to minimal sentence series: the same sentence context is paired with several objects, yielding LVC readings in half of the cases and full-verb readings in the other half. This design provides more directly comparable items than isolated mini-

mal pairs, improves statistical power, and reduces contextual confounds. Unlike most minimal-pair resources, which contrast acceptable and unacceptable sentences, our dataset contrasts grammatical and natural sentences that differ only in the lexical-functional status of the same surface verb. We illustrate its use through contrastive embedding analysis and surprisal-based probing in active and passive configurations, testing whether models capture both verb-to-noun and noun-to-verb expectations in LVCs.

2 Related Work

In NLP, LVCs have mainly been treated in the context of identification or classification tasks. Most often, supervised, syntax-based, lexical-knowledge-based, and multilingual classical feature-based or neural machine learning is used to distinguish LVCs from non-LVC verb–noun co-occurrences (Tan et al., 2006; Tu and Roth, 2011; Vincze et al., 2013; Chen and Palmer, 2015; Vaidya et al., 2016; Cordeiro and Candito, 2019; Nagy et al., 2020), often within a broader scope of MWE processing and verbal-MWE shared tasks (Constant et al., 2017; Savary et al., 2017; Ramisch et al., 2018). Corpus-linguistic work on collocation and collostructional analysis provides association measures for quantifying lexical co-occurrence and word–construction attraction (Stefanowitsch and Gries, 2003; Gries, 2013); we use this tradition as motivation for controlling verb–noun conventionality, while treating light-verb status as a distinct lexical-functional contrast. Methodologically, our work builds on targeted minimal-pair evaluation of language models (Marvin and Linzen, 2018; Warstadt et al., 2020; Gauthier et al., 2020; Xiang et al., 2021; Jumelet et al., 2025; Baroni et al., 2026) and on surprisal-based accounts of incremental processing (Levy, 2008; Smith and Levy, 2013; Goodkind and Bicknell, 2018). More specifically, it follows targeted test-suite approaches that evaluate models through region-level surprisal or probability contrasts (Gauthier et al., 2020; Hu et al., 2020; Wilcox et al., 2021; Beyer et al., 2021; Táboas García and Wanner, 2025). Extending this work, our benchmark uses matched grammatical sentence contrasts, rather than acceptable–unacceptable contrasts, to test whether models encode the directional expectations between predicative nouns and their light verbs.

3 Dataset Construction

We initialized the dataset with five high-frequency English verbs that commonly occur in LVCs in CollFrEn (Fisas et al., 2020): *make*, *take*, *give*, *have*, *receive*[1]; about to be included are *pay*, *hold*, *perform*, *conduct*, and *draw*.

For each verb, we retrieved a list of LVC candidates from CollFrEn. All candidates were validated according to three criteria: the noun had to be a predicative nominal base, the verb had to be semantically light in the construction, and the combination had to instantiate a canonical light-verb construction corresponding to either the Oper_1 or Oper_2 collocational relation in Explanatory Combinatorial Lexicology (Melčuk et al., 1995). Borderline cases and idiomatic expressions were excluded. From this first list, we selected 20 to 40 LVCs per verb. We then constructed a corresponding full-verb–noun set. We selected a set of nouns denoting concrete objects that could occur after each verb. The constructions in the full-verb condition had to be grammatical and natural, instantiate the same broad syntactic frame of the LVC condition, and involve a noun with which the verb expresses an independent lexical meaning. Table 1 exemplifies the resulting contrasts.

For each verb, a set of natural-language sentence contexts was generated and paired with the verb + object constructions. We produced contexts varying across several grammatical properties, including voice, tense, length, and number of specifications, allowing different degrees of stimulus variation and experimental control.[2].

The detailed specification of the resulting dataset, including its size, is provided in Appendix B.

4 Experiments

In what follows, we present the results obtained with minimal contexts, showing that even minimal contextual information is sufficient to observe discrimination effects between LVCs and fully lexical verbs. We evaluate the proposed dataset along two complementary dimensions. First, we use it for

[1] In the current version, we excluded other frequent verbs like *do* and *get* often included in LVC recognition tasks; cf., e.g., (Tu and Roth, 2011), because many of their transitive uses are generic activity predicates and do not yield clear minimal-pair contrasts between light-verb and full-verb readings.

[2] Details on the dataset creation are reported in the Appendix. Up to The dataset, along with the code and materials used for its generation are released with documentation for reuse, for submission in the uploaded zip file, upon acceptance, in a dedicated github repo

Verb	LVC use	Full-verb use
make	*make a decision*	*make a cake*
take	*take a walk*	*take a mug*
have	*have a celebration*	*have a car*
give	*give a look*	*give a coin*
receive	*receive attention*	*receive a package*

Table 1: Examples of contrastive LVC and full-verb verb–noun pairs.

Condition	Example
LVC-active	*During the afternoon, they made a decision.*
Full-active	*During the afternoon, they made a cake.*
LVC-passive	*During the afternoon, a decision was made.*
Full-passive	*During the afternoon, a cake was made.*

Table 2: Example of matched active and passive sentence-level contrasts.

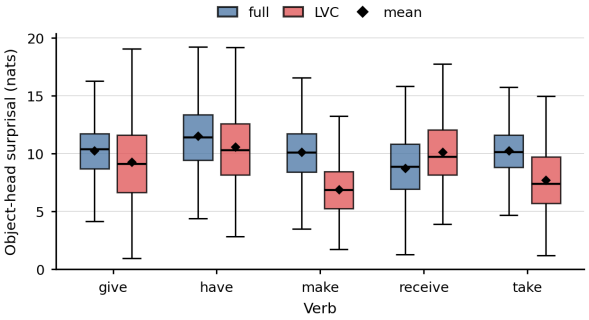

Figure 1: Final object surprisal in active sentences for LVC and full-verb controls, computed over 47,600 sentences.

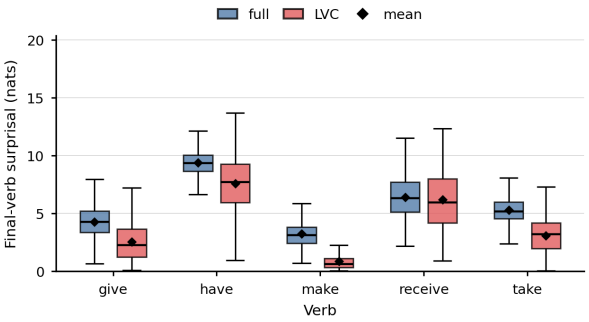

Figure 2: Final-verb surprisal in passive sentences for LVC and full-verb controls, computed over 47,600 sentences.

surprisal-based probing of language models, testing whether models assign different expectations to LVC and full-verb configurations when the surface verb and syntactic frame are controlled. Second, we conduct a contrastive embedding analysis in order to examine whether object-level contextual representations separate LVC objects from full-verb objects under matched verbal and sentential contexts. Together, these experiments are intended to assess whether the light/full-verb distinction is reflected in model expectations and model-internal representations. In the experiments reported here, we use the base `google/gemma-3-270m` checkpoint from the Gemma 3 family (Gemma Team, Google DeepMind, 2025).

4.1 Surprisal-Based Probing

For a token w_t in context $w_{<t}$, *surprisal* is defined as the negative log-probability assigned to that token by a language model:

$$S(w_t) = -\log P(w_t \mid w_{<t}).$$

We report surprisal in nits. For multi-token critical regions, we use the maximum surprisal score across the tokens.

The surprisal experiment follows the SyntaxGym evaluation format (Gauthier et al., 2020): each test item consists of matched conditions and manually defined critical regions. We compute token-level surprisals for the critical regions using autoregressive language models and compare differences across verbs and classes.

In active configurations, the critical region is the noun following the verb, as in *make a decision* versus *make a cake*. This tests whether the preceding verb and context make an LVC-compatible nominal base more or less expected than a full-verb object. In passive configurations, the critical region is the verb following the nominal base, as in *a decision was made* versus *a cake was made*. This tests whether a predicative nominal base increases the expectation for its conventional light verb. For each condition, we compare surprisal values across matched LVC and full-verb sentences.

Figure 1 shows the contrastive surprisal scores for the active-sentence setup, comparing light-verb and full-verb uses. With the exception of *receive*, the critical regions in the light-verb condition receive lower surprisal than their full-verb counterparts. This indicates that the model assigns higher contextual probability to nominal bases occurring with conventional light verbs such as *give*, *have*, *make*, and *take*. The pattern for *receive* is less clear, suggesting that its status in the tested items is more ambivalent. In the passive-sentence setup, the preceding noun provides the main cue for predicting the following verb; see Figure 2. Lower surprisal in the LVC condition therefore suggests that the model captures, at least partially, the association between predicative nominal bases and their

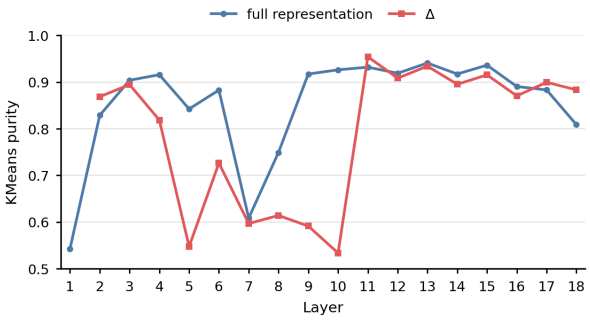

Figure 3: KMeans purity for full object-head hidden states and Δ_l, the change in the object-head representation from layer $l-1$ to layer l. Clustering is computed over 47,600 active sentences.

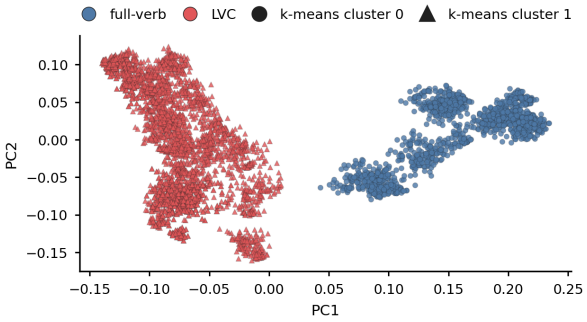

Figure 4: Example PCA projection of object-level contextual representations from layer 15 of `gemma-3-270m`, computed over 5,000 active sentences. Each point corresponds to one sentence, with the object representation obtained by averaging the embeddings of all object tokens. PC1 and PC2 explain 26.6% and 9.0% of the variance, respectively. KMeans clustering was performed on the first 50 PCA dimensions, which together retain 96.7% of the total variance. In this reduced 5,000-sentence sample, the induced clusters align with the LVC and full-verb conditions, yielding a purity of 1.00.

conventional light verbs. Variation across verbs is greater in the passive setup than in the active setup. *Make* shows the lowest surprisal values, which suggests that it is the most strongly expected support verb in the tested passive configurations.

4.2 Contextual Embedding Comparison

As a complementary analysis, we compare contextual representations at the object position in LVC and full-verb constructions. For each sentence, we extract the hidden states corresponding to the object noun phrase and average over its tokens, yielding a single object-level representation for each instance. We then ask whether these representations separate according to construction type. Since the surface verb and sentence template are controlled, such separation would indicate that the object representations contain information distinguishing LVC from full-verb contexts, although this information may also reflect correlated lexical-semantic properties of the object nouns. To assess this, we reduce the dimensionality of the object representations and apply clustering analyses. We perform this analysis both on the layer-wise object representations and on Δ_l, defined as the change in the object representation from layer $l-1$ to layer l. We then evaluate whether the induced clusters align with the LVC and full-verb conditions. This analysis tests whether the contrast observed in the surprisal experiment is also reflected in the model's contextual representation space.

Figure 3 shows that the induced clusters align strongly with the LVC and full-verb conditions. The effect is especially visible for Δ_l in later layers, suggesting that the model updates object representations differently across the two construction types. Figure 4 illustrates this pattern with a PCA projection from layer 15 of `gemma-3-270m`. These results suggest that the LVC/full-verb contrast is reflected not only in token-level surprisal, but also in the geometry of contextual object representations.

5 Discussion

Our dataset provides a controlled probing instrument for investigating differences between verbs used as LVCs, and verbs used with their full lexical meaning. The obtained surprisal scores suggest that gemma-3-270m is sensitive to LVCs. LVC-compatible nominal bases are generally less surprising after light verbs in active sentences, and predicative nominal bases make their associated support verbs more predictable in passive sentences. However, this varies across verbs, which implies the importance of the degree of conventionality of the considered LVCs.

The contextual embedding comparison experiment shows that at some point of processing the two object types are clearly separated, suggesting that the models capture at least some of their differential properties. These findings will be explored in greater depth in future analyses and with a variety of models to understand what specific lexical properties underlie this separation. Overall, the dataset is designed to support further probing experiments, including extensions to additional verbs, broader contextual settings, other languages, and alternative analytical or experimental approaches including psycholinguistic testing.

Limitations

While this database provides controlled minimal pairs, it focuses on a limited set of verbs and constructions, potentially restricting generalization across broader verb classes. Additionally, sentence contexts are semi-controlled, which may reduce ecological validity. Finally, our analyses focus on English, leaving cross-linguistic variation in light verb constructions unexplored. However, we view these choices as enabling precise comparisons, designed this dataset and methods to support future extensions across verbs, contexts, and languages.

Ethical Statement

This work introduces a dataset of constructed sentences designed for research on language processing in LLMs. All materials have been automatically generated starting from human-designed materials and do not contain nor offensive or biased content.

AI support was used for coding and writing, but not research.

References

Marco Baroni, Emily Cheng, Iria de Dios-Flores, and Francesca Franzon. 2026. Tracing the complexity profiles of different linguistic phenomena through the intrinsic dimension of llm representations. *arXiv preprint arXiv:2601.03779*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications, Stanford.

Miriam Butt. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*. Cambridge University Press, Cambridge.

Wei-Te Chen and Martha Palmer. 2015. English light verb construction identification using lexical knowledge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Carlos Ramisch van der Plas, Lonneke, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Silvio Ricardo Cordeiro and Marie Candito. 2019. Syntax-based identification of light-verb constructions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 97–104, Turku, Finland. Linköping University Electronic Press.

Beatriz Fisas, Luis Espinosa Anke, Joan Codina-Filbá, and Leo Wanner. 2020. CollFrEn: Rich bilingual English–French collocation resource. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 1–12, online. Association for Computational Linguistics.

Jens Fleischhauer, , and Anja Latrouite. 2025. *Light Verbs*. Language Sciences Press, Berlin.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Stefan Th. Gries. 2013. 50-something years of work on collocations: What is or should be next ... *International Journal of Corpus Linguistics*, 18(1):137–166.

Jane Grimshaw and Armin Mester. 1988. Light verbs and -marking. *Linguistic Inquiry*, 19(2):205–232.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.

Otto Jesperen. 1942. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. Ejnar Munksgaard, Copenhagen.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Igor A. Melčuk, André Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

István Nagy, Veronika Vincze, and Richárd Farkas. 2020. Detecting light verb constructions across languages. *Natural Language Engineering*, 26(3):319–348.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 7 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Anatol Stefanowitsch and Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Alba Táboas García and Leo Wanner. 2025. Assessing the agreement competence of large language models. In *Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, pages 36–53, Ljubljana, Slovenia. Association for Computational Linguistics.

Kathleen Tan, Tong Ming Lim, Chi Wee Tan, and Wei Wei Chew. 2021. Automatic identification of light verb constructions: A review. *IEM Journal, Special Edition: International Conference on Digital Transformation and Applications*.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 49–56, Sydney, Australia. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2011. Learning english light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions*, pages 31–39. ACL.

Verb	Collocational Objects	Concrete Objects	Total Objects
make	20	20	40
take	20	20	40
receive	20	20	40
have	40	40	80
give	40	40	80

Table 3: Verb inventory used for sentence generation.

Ashwini Vaidya, Sumeet Agarwal, and Martha Palmer. 2016. Syntax-based identification of light-verb constructions. In *Proceedings of the International Conference on Computational Linguistics*, pages 1320–1329, Osaka, Japan.

Veronika Vincze, István Nagy, and János Zsibrita. 2013. Learning to detect english and hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2).

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 939–952. Association for Computational Linguistics.

Beilei Xiang and 1 others. 2021. CLiMP: A benchmark for chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Examples of light and full verb uses

B Dataset specifications

The sentence database were built around the verbs *make, take, give, have, and receive*. Each verb occurred with the same number of collocation and non-collocation objects, like in the table:

B.1 Set used in the presented experiments

We release, as ready-to-use resources, the datasets used to perform the analyses described above. In all generated sentences, the target verb–object construction occurs in sentence-final position, facilitating evaluation with autoregressive language models.

For each verb, collocational and concrete object nouns were paired with temporal sentence contexts and optional adverbial modifiers. Sentences

were generated through rule-based templates implementing grammatical constraints on tense (past simple, future simple, past perfect), voice (active, passive), determiner insertion (articles or possessives, where grammatical), adverb placement, and number agreement.

The dataset includes 170 temporal specifications, 35 adverbial modifications, 3 tense specifications (past simple, future simple, past perfect), and 2 voice specifications (active, passive).

The passive database contains sentences in all three tense specifications, each released in adverb and non-adverb versions. Passive forms were generated using manually specified past participles and singular/plural auxiliary agreement.

The active pronoun database released as ready-to-use contains sentences with the pronoun *they* as subject. Sentences were automatically inflected for tense and combined with temporal and adverbial contexts.

B.2 Full dataset and generation code

In addition to the datasets used in the experiments reported above, we release the full sentence-generation framework and accompanying materials, designed to support large-scale controlled experimentation. The generation pipeline is rule-based and allows users to systematically manipulate a range of grammatical and lexical properties while preserving controlled sentence structure.

The current release includes the mentioned and 280 object nouns distributed across collocational and concrete uses, 170 temporal specifications, 35 adverbial modifiers, 3 tense specifications (past simple, future simple, past perfect): these occur with 100 subject nouns (e.g., *doctor, guest, teacher*), inflected in 2 morphological number values and paired with either a determinate article or a possessive pronoun, where grammatical. The current nominal-subject release includes only active voice constructions.

The approximate size of the full generation combinations corresponds to over 2,000,000,000 potential sentences.