The holistic storage of verb+*up* phrases in text-based and audio-based language models

Anonymous ACL submission

Abstract

A crucial aspect of language is trading off between stored representations and abstract knowledge: one must retrieve learned representations, but also generate novel ones by applying productive rules. While recent work has examined abstract knowledge in language models, holistic storage of multi-word units has received far less attention. We probe internal representations in text-based LLMs and an ASR model, testing whether V+*up* phrasal verbs develop distinct representations as a function of frequency and predictability. All models show evidence of holistic storage driven by frequency and predictability, further supporting usage-based theories of language.

1 Introduction

A central debate in linguistics concerns how humans trade off between computation and storage (Houghton, 2025a,b; Houghton and Morgan, 2023, 2024; Kapatsinski et al., 2009; Morgan and Levy, 2015, 2016, 2024; Stemberger and MacWhinney, 2004, 1986). Computation refers to applying abstract knowledge to generate new representations; for example, deriving *wugs* from *wug* via a productive pluralization rule. Storage refers to retrieving a whole representation from memory rather than computing it, such as accessing a holistic representation of a common phrase (e.g., *I don't know*) despite being able to, in principle, generate the phrase compositionally. Both mechanisms are clearly at work in language learning and processing, but the factors that drive what forms are produced via computation and what forms are stored and retrieved holistically remains poorly understood.

1.1 Computation vs Storage in Humans

There is an abundance of evidence for abstract knowledge in language use by humans. Children generalize morphological rules productively to novel words they have never encountered (Berko, 1958), and priming studies show that sentences sharing syntactic structure or semantically related words facilitate each other's processing, implying that something abstract is shared across their representations (Bock, 1986; Meyer and Schvaneveldt, 1971). Similarly, when ordering novel or low-frequency binomials, humans rely on abstract preferences (e.g., preferring the shorter word first) rather than simply producing the more frequent ordering (Morgan and Levy, 2016).

Storage has had a more controversial history. Early accounts held that only irregular forms (e.g., *went*) were stored holistically and that any compositionally derivable form was derived via computation (Pinker and Ullman, 2002). Evidence against this view has since accumulated (Bybee and Scheibman, 1999; Houghton, 2025b; Kapatsinski et al., 2009; Morgan and Levy, 2016; Stemberger and MacWhinney, 2004). Stemberger and MacWhinney (2004) showed that inflection errors are less common for high-frequency words, suggesting holistic storage even for regularly derived forms. Bybee and Scheibman (1999) demonstrated that *don't* is phonetically more reduced in high-frequency phrases like *I don't know* than in lower-frequency phrases like *I don't go*; if *don't* always had the same representation, such context-specific reduction would be difficult to explain.

Processing studies have provided converging evidence. Morgan and Levy (2016) found that while human ordering preferences for low-frequency binomials are driven by abstract preferences, preferences for high-frequency binomials are driven by item-specific preferences, suggesting a frequency-dependent shift from computation to storage. Most directly relevant to the present study, Kapatsinski et al. (2009) found that listeners are slower to detect *up* in high-frequency Verb+*up* (V+*up*) phrases (e.g., *pick up*) than in lower-frequency ones. Specifically, they tasked participants with pressing a button when they heard *up*, which occurred ei-

ther within a word or within a V+*up* phrase. They found that while participants were faster to press the button when responding to medium-frequency phrases relative to low-frequency phrases, they were slower to press the button when the verb+*up* phrases were high frequency. Further, Houghton (2025b) found that this pattern holds for high predictability phrases (V+*up* phrases where *up* is likely to appear given the verb), suggesting that high-frequency and high-predictability V+*up* phrases are represented holistically.

1.2 Computation vs Storage in LMs

Whether language models exhibit analogous computation-storage tradeoffs to humans has become an active area of inquiry. On the computation side, results are mixed: some studies find that models learn abstract generalizations not present in training (Lasri et al., 2022; Misra and Mahowald, 2024; Yao et al., 2025), while others find that models fail to use abstract knowledge where humans do, such as morphological generalization to novel words (Haley, 2020). On the storage side, there is no doubt that models rely heavily on memorization (McCoy et al., 2023), and item-specific frequency effects (i.e., memorization) have been documented in tasks where humans show abstract preferences (Houghton et al., 2025).

Whether language models develop holistic phrasal representations in a similar manner as humans is an open question. If they do, this would suggest that holistic storage is a natural consequence of learning from distributional patterns in the language, requiring no special storage mechanism beyond what is needed for general language learning, and lending support to usage-based accounts of how computation and storage interact. Audio-based models are especially well-suited to address this question because a good deal of the evidence for holistic storage comes from listening paradigms. It is thus important to understand whether LLMs and automatic speech recognition (ASR) models both show similar representations. And yet, holistic phrasal storage has never been examined in any such ASR model. More broadly, holistic phrasal storage in language models has rarely been examined in any modality, and it has not been studied in models trained on human-comparable amounts of data.

1.3 Present Study

The present study addresses this gap. We probe internal representations in text-based language models that were trained on an amount of data comparable to humans (BabyLMs),[1] a large language model (OLMo-3 7B), and an audio-based speech recognition model (Whisper-small). These models were chosen to help illuminate effects of training size, number of parameters, and modality (speech vs text). In order to probe their representations, we trained logistic classifiers to detect the embedding of standalone *up*, then tested the classifier on V+*up* phrases of varying frequency and predictability. If these models develop holistic phrasal representations, the representation of *up* in high-frequency and high-predictability V+*up* phrases should diverge from that of standalone *up* more than in lower-frequency and lower-predictability phrases, resulting in lower logit scores from the classifier. Our specific contributions are:

- We train and release three open-access autoregressive models trained on the BabyLM v3 corpus (Charpentier et al., 2025), checkpointed every 20M tokens, to facilitate future research on human-scale language learning.
- We show that holistic phrasal storage emerges in both text-based LLMs and an ASR model, establishing that frequency- and predictability-driven holistic representations arise even from models trained on an amount of data comparable to humans, and even across modalities.
- We show that frequency effects on phrasal storage are robust across model sizes, but predictability effects strengthen with scale, suggesting that sensitivity to co-occurrence statistics beyond raw frequency requires greater representational capacity.

2 Model Training

In order to examine holistic storage in models that have seen an amount of data comparable to humans, we trained three autoregressive language models on the BabyLM v3 corpus (Charpentier et al., 2025), a 150M-token dataset designed to be more reflective of the scale and quality of the language that

[1] The model was trained on 150 million tokens. The average college-aged human experiences $\tilde{3}$50 million words (Levy et al., 2012) so the model is trained on a little less than half the tokens that an average college-aged human has experienced.

humans receive.[2] All three models follow the OPT decoder-only transformer architecture (Zhang et al., 2022), with 125M, 350M, and 1.3B parameters, respectively.

Prior to training, we fit a byte-pair encoding (BPE) tokenizer directly on the BabyLM training corpus, yielding a vocabulary of 8,192 subword types. This tokenizer was shared across all three model sizes, ensuring that cross-model comparisons are not confounded by differences in tokenization. A full description of the model training is included in the Appendix (Section A).

3 Experiment 1: UP Independently

Experiment 1 tests whether LLMs develop holistic phrasal representations analogous to those proposed by usage-based theories (e.g., Houghton, 2025b; Kapatsinski et al., 2009), using classifiers trained on standalone *up* representations and applied to V+*up* phrases varying in frequency and predictability. We investigate this across OLMo-3 7B and the three BabyLMs (OPT-125M, 350M, and 1.3B).

3.1 Methods

For each model (OLMo-3 7B and BabyLM OPT-125M, OPT-350M, and OPT-1.3B), we extracted the representation of the token *up* from each hidden layer of the models for each sentence. Using these representations, we trained a separate logistic regression classifier for each layer to distinguish standalone *up* (which did not occur in V+*up* phrases in the training of the classifier) from other tokens in the same sentence (these other tokens did not contain the segment *up* in any capacity); the training items are described in the next section. The trained classifier was then tested on a held-out test set comprising V+*up* phrases, and for each sentence the classifier returned a logit score reflecting the predicted probability that the representation of *up* in the V+*up* phrase resembled the standalone class.

Frequency for each V+*up* type is operationalized as its raw corpus count, log-transformed. Letting $c_{vup} = \text{count}(V\text{+}up)$:

$$\text{log-frequency} = \log\big(c_{vup}\big) \tag{1}$$

Predictability is operationalized as the log-odds ratio of V+*up* occurrences to V not followed by up in the corpus. Letting $c_V = \text{count}(V)$:

$$\text{log-predictability} = \log\left(\frac{c_{vup}}{c_V - c_{vup}}\right) \tag{2}$$

Counts were derived from the training corpus of each model: the BabyLM V3 dataset (Charpentier et al., 2025) for the BabyLM models, and Dolma v1.7 (queried via the infini-gram API; Liu et al. 2024) for OLMo-3 7B. Although OLMo-3 7B was trained on Dolma 3, Dolma v1.7 is the most recent snapshot indexed by infini-gram and draws from the same underlying sources, making it a reasonable approximation of OLMo-3 7B's training distribution.[3]

3.1.1 Classifier Training

The classifier was trained to distinguish the language models' representations of the preposition *up* from representations of other tokens in the same sentence. Positive training examples consisted of 1,000 occurrences of *up* which occurred in sentences strictly as a preposition,[4] drawn from sentences in the C4 corpus (Raffel et al., 2020). Negative examples were 1,000 tokens randomly selected from the same sentences, restricted to tokens whose decoded string consists entirely of alphabetic characters (no numbers, punctuation, or special characters), and excluding the preposition *up* itself and any token containing *up* as a substring. The validation set was drawn from the same pool of sentences (a non-overlapping subset), with token positions resolved per model's tokenizer; exact counts fall slightly below 1,000 for some models because the tokenizer does not always produce an isolated *up* token. For example, when *up* appears before punctuation and the tokenizer fuses the two into a single unit (e.g., *up,*), there is no isolable *up* position to extract, so that instance is excluded. The classifiers for all models were trained on the same underlying sentences.

The test set comprised V+*up* phrases (e.g., *pick up*) with at least 20 occurrences in the corpus; up to 20 sentences were sampled per type. Because the BabyLM corpus is smaller than Dolma, fewer

[2] Though it is worth noting, as has been pointed out before (e.g., Houghton, 2025b), that it may be misleading to compare the tokens that LLMs receive to the "tokens" that humans receive, since humans encounter language in a context-rich environment while LLMs see only the raw text.

[3] The relative frequencies and co-occurrence patterns of common English phrasal verbs are unlikely to differ substantially between Dolma versions, as both are large-scale web-text corpora of similar composition.

[4] We used a morphological parser to filter out sentences in which *up* was not tagged as a preposition.

V+*up* types attain a valid (non-zero) predictability estimate; only types with a valid estimate are included in the analyses. Full item-level statistics are reported in the Appendix (Table 3). For OLMo-3 7B, 4,081 unique V+*up* types were included, with a median corpus frequency of 43,608 and a median log-predictability of -5.39. The BabyLM models shared 1,039 items, with a lower median frequency (34) and log-predictability (-3.40), reflecting the smaller vocabulary of the BabyLM training data. Classifier training and validation split sizes are shown in the Appendix (Table 4).

3.1.2 Analyses

In order to examine the effects of frequency and predictability on the representations of *up* in V+*up* phrases, we implemented a Bayesian mixed-effects regression model using the *brms* package (Bürkner, 2017). For each statistical analysis, the outcome variable was the logit score returned by the classifier on each test item. We included fixed-effects for frequency and predictability, both of which were centered and scaled (denoted c_log_freq and c_log_predic below). We also included random intercepts for phrasal verb type (since there were multiple sentences for each phrasal verb). Weak, uninformative priors were included on each fixed-effect. The model syntax is included below:

$$\begin{aligned} \text{logit} \sim\ & c_\text{log_freq} \times c_\text{log_predic} \\ & + (1 \mid \text{verb_up}) \end{aligned} \tag{3}$$

Separate statistical models were fit for each language model at the final hidden layer. For each coefficient we report the posterior mean, standard error, 95% credible interval (CI), and the proportion of posterior samples greater than zero (% > 0). We consider an effect to be meaningful if the 95% CI does not contain zero.[5]

In order to examine the difference in representations across hidden layers, we additionally fit a generalized additive model (GAM, Wood, 2017). As with the Bayesian analyses, the dependent variable was the classifier logit. Specifically, two models were fit separately, one for frequency and one for predictability, each with a tensor product smooth over the predictor and hidden layer index (with a a random intercept for verb):

$$\begin{aligned} \text{logit} \sim\ & \text{te}(\text{log_freq}, \text{layer}) \\ & + s(\text{verb_up}, \text{bs='re'}) \\ \text{logit} \sim\ & \text{te}(\text{log_predic}, \text{layer}) \\ & + s(\text{verb_up}, \text{bs='re'}) \end{aligned} \tag{4}$$

3.2 Results

Full numerical results are reported in the Appendix (Table 5 for the Bayesian model and Table 6 for the GAM); they are visualized in Figure 1 and Figure 2. Figure 8 (Appendix) shows the layer-by-layer difference between high- and low-predictor items more directly.

For all four LLMs, there was a negative effect of both frequency and predictability on the classifier's logit score. In other words, both frequency and predictability result in the representation of *up* in V+*up* phrases being less similar to the standalone representation of *up*.

Additionally, the by-layer analysis demonstrates that this divergence begins to appear early on in the larger models (specifically, OLMo-3 7B and BabyLM 1.3B show this early effect), while the smaller models do not show an effect until later layers. This suggests that for larger models, the representations of high-frequency V+*up* phrases are more distinct from the representations of the individual constituent words (in the case of the present experiment, the representation of standalone *up*) than for smaller models.

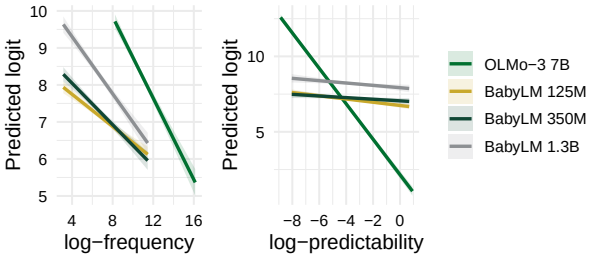

Figure 1: Final-layer brms predicted logit by frequency (left) and predictability (right) for all models (UP independently). Shading indicates 95% CIs.

3.3 Discussion

Experiment 1 found that the logits from a classifier trained on standalone *up* representations decreased for high-frequency and high-predictability V+*up* phrases. Additionally, the effect of predictability appears to be weaker in the smaller models. These results replicate those of Kapatsinski et al. (2009)

[5]Though as Houghton et al. (2024) point out, unlike frequentist statistics, Bayesian analyses don't force us to commit to a binary of significance vs. non-significance, and the percentage of samples greater than zero can be interpreted in a continuous manner.

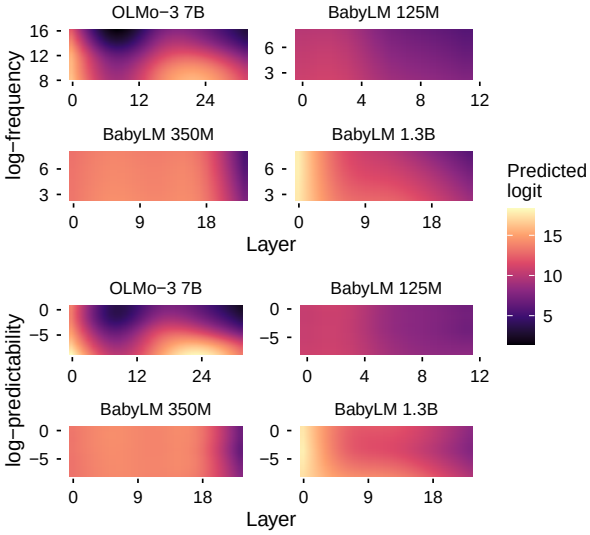

Figure 2: GAM-predicted logit by layer and predictor for all models (UP independently). Top: frequency; bottom: predictability.

and Houghton (2025b), who targeted humans' representations. In Experiment 2, we more closely replicate these studies by also examining representations of *up* as a subword unit.

4 Experiment 2: UP as Subword

One possible criticism of the results of Experiment 1 is that the lower logit score for high-frequency/high-predictability V+*up* phrases could reflect semantic bleaching rather than holistic storage. That is, high-frequency items often undergo semantic bleaching or extension (e.g., Harmon and Kapatsinski, 2017), resulting in a widening of the phrase's meaning. For example, in English the word *literally* has been widened beyond its original meaning (e.g., "I'm literally dying", where *literally* behaves as an emphatic). To address this concern, we replicate the design of Experiment 1 with one key change: rather than training the classifier on the preposition *up* exclusively, we additionally train it on instances of *up* embedded within a larger word (e.g., *update*), requiring the model to identify the phonetic sequence *up* regardless of whether it functions as a preposition, rather than the preposition *up* specifically. This tests whether the frequency and predictability effects observed in Experiment 1 generalize to the *up*-segment more broadly. This design is also arguably more faithful to Kapatsinski et al. (2009) and Houghton (2025b)'s studies, where participants were tasked with recognizing *up* in general, regardless of whether it occurred within a word or as a preposition.

4.1 Methods

The procedure was identical to Experiment 1 (same models; classifiers trained and evaluated at each hidden layer; same test set, outcome variable, and statistical approach) with one difference: the classifier was trained on a broader set of positive examples that included instances of *up* embedded within a larger word (e.g., *update*, *upon*).

4.1.1 Classifier Training

The training set for this experiment combined two types of positive example: 1,000 occurrences of standalone *up* (identical to Experiment 1) and 1,000 occurrences of *up* embedded within a larger word (e.g., *update*, *upon*, etc). Critically, the subword positives were restricted to **unique word types**: each up-containing word contributed exactly one instance, so the classifier could not learn to recognise a high-frequency form like *upon* from repeated exposure. Negative examples (1,000 drawn from each sentence pool) were tokens consisting entirely of alphabetic characters that did not contain *up* as a substring. The validation set was constructed by the same procedure (targeting 1,000 positive, 1,000 negative per sentence pool; exact counts vary by tokenizer). As in Experiment 1, each model used the same underlying sentences with their own tokenizers. The test set was identical to Experiment 1. Classifier training/validation split sizes are shown in the Appendix (Table 7).

4.1.2 Analyses

Analyses were identical to those of Experiment 1, using the same model formula (Equation 3), predictor definitions (Equation 1, Equation 2), priors, and random-effects structure. The brms models were fit separately for OLMo-3 7B and each BabyLM model at the final transformer layer. Across-layer GAMs were fit using the same approach as Experiment 1 (Equation 4).

4.2 Results

Full numerical results are reported in the Appendix (Table 8 for the Bayesian model and Table 9 for the GAMs); they are visualized in Figure 3 and Figure 4. Figure 9 (Appendix) shows the layer-by-layer difference more directly. Frequency effects replicate Experiment 1 across all four models. Predictability effects are more variable: BabyLMs show absent or positive effects, OLMo-3 7B shows progressively stronger negative effects, consistent with predictability sensitivity increasing with scale

(though in the present study, it's unclear if it's due to the scale of data or scale of the model, or both). The by-layer patterns similar show a progressively stronger effect of frequency for later layers relative to earlier layers, while the effect of predictability follows a similar pattern for OLMo, but not for the BabyLMs. For the BabyLMs, the effect of predictability across layers either stays relatively constant across layers or, in the case of the 1.3B model, becomes negative around layer 5 and then slowly grows positive again.

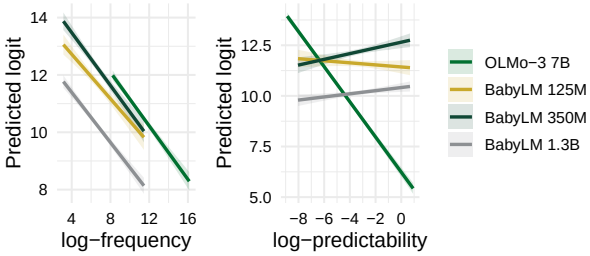

Figure 3: Final-layer brms predicted logit by frequency (left) and predictability (right) for all models (UP as subword). Shading indicates 95% CI.

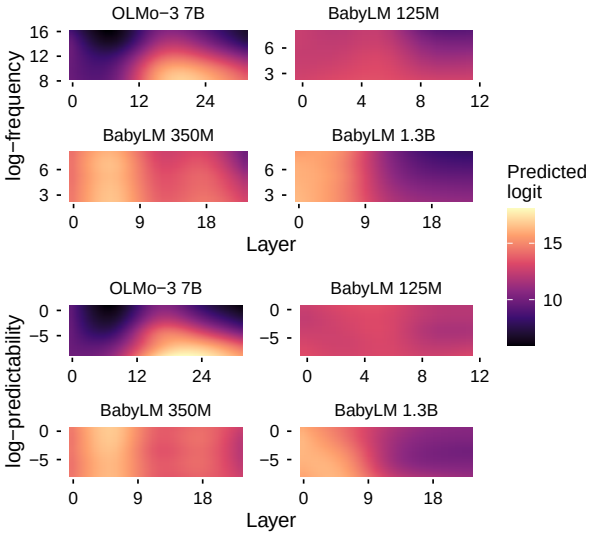

Figure 4: GAM-predicted logit as a function of transformer layer (x-axis) and predictor value (y-axis) for each model (UP as subword). Top row: log-frequency; bottom row: log-predictability. Color encodes the predicted logit score (random effects excluded).

4.3 Discussion

The results of Experiment 2 mostly replicate those of Experiment 1. Frequency shows a strong, consistent effect such that the representation of *up* in high-frequency V+*up* phrases diverge from representations of *up*, even as a subword unit. The effect of predictability seems to emerge in larger models, suggesting that sensitivity to predictability may require more representational capacity than sensitivity to frequency. This is perhaps unsurprising: frequency is directly recoverable from surface co-occurrence counts, while predictability requires tracking the conditional distribution of the particle given the verb — a more relational statistic that may only become reliably encoded at larger scale.

5 Experiment 3: Whisper (ASR Model)

In Experiment 3, we apply the same classifier approach to Whisper-small, an automatic speech recognition (ASR) model trained on spoken audio rather than written text. Whisper differs from the models in Experiments 1 and 2 in both its training modality and its encoder-decoder architecture. Examining Whisper allows us to ask whether the frequency and predictability effects on phrasal representations generalize beyond text-based models to representations learned from speech.

5.1 Methods

The procedure mirrored Experiments 1 and 2, with two key differences: (1) the model was Whisper-small, an ASR model with an encoder-decoder architecture trained on spoken audio rather than text; and (2) the stimuli were spoken audio segments rather than written sentences. For each segment, we ran the audio through Whisper and extracted the hidden-state representation of *up* at each layer of the encoder and decoder separately. A logistic regression classifier was trained independently for each component (encoder, decoder) on spoken instances of standalone *up* (see Classifier Training below), then applied to spoken V+*up* phrasal verb segments. As in Experiments 1 and 2, each segment contributed one observation, with random intercepts by phrasal verb type. For the sake of brevity, we only trained the classifier to differentiate between Whisper's embeddings of *up* as a standalone preposition and non-*up* tokens.

5.1.1 Classifier Training

The Whisper experiment used spoken language data drawn from a subset of the GigaSpeech corpus (Chen et al., 2021) containing 224,118 timestamped speech segments in which *up* occurred. Word-level timestamps within each segment were obtained using WhisperX forced alignment. Each occurrence of *up* was classified by running spaCy

part-of-speech tagging on the full segment transcript: if the token immediately preceding *up* was tagged as a VERB, the instance was labelled *V+up* (e.g., *pick up*, *clean up*). This classification procedure mirrors the one used to identify standalone *up* in Experiment 1. Of the 224,118 segments that contained *up*, 165,367 were classified as *V+up*, spanning 4,161 unique V+*up* types.

The **training and validation sets** used 1,000 standalone *up* instances each as positive examples, paired with 1,000 randomly selected non-*up* words from the same segment as negative examples. This mirrors the standalone-*up* training design of Experiment 1. Negative words were restricted to those whose string consists entirely of alphabetic characters (matching the criteria used in Experiments 1 and 2), with a minimum duration of 20 ms. Training and validation segments were non-overlapping.

The **test set** comprised V+*up* types with at least 5 occurrences in the audio dataset, with up to 20 instances per type. Corpus frequency and predictability were drawn from Dolma v1.7 (via infini-gram), identical to OLMo-3 7B. After filtering to types with a valid predictability estimate, 1,426 unique V+*up* types were retained (median corpus frequency: 243,412; median log-predictability: -4.08). The high median frequency reflects the audio coverage threshold: types requiring at least 5 spoken occurrences are predominantly common V+*up* phrases (e.g., *pick up*, *end up*, *set up*) that are also extremely frequent in Dolma. Despite the high median-frequency, the test items still spanned a wide range of frequencies and thus we still have the statistical power to detect effects of frequency. Split sizes and item-level statistics are included in the Appendix (Table 10 and Table 3).

5.1.2 Analyses

Analyses followed the same procedure as Experiments 1 and 2, using the same model formula (Equation 3) and predictor definitions (Equation 1, Equation 2). The brms models were fit separately for Whisper's encoder and decoder at the final hidden layer. Corpus statistics were drawn from Dolma v1.7 (queried via the infini-gram API) for both components. Across-layer GAMs were fit using the same approach as Experiment 1 (Equation 4).

5.2 Results

Full numerical results are reported in the Appendix (Table 11 for the Bayesian model and Table 12 for the GAM model); they are visualized in Figure 5 and Figure 6 below. Figure 10 (Appendix) shows the layer-by-layer difference between high- and low-predictor items more directly.

The results for Whisper show a similar trend across both the encoder and decoder. Specifically, we see a negative effect for frequency and predictability across both.

The by-layer analysis demonstrates that the negative effects in the encoder are variable across layers, with predictability staying mostly negative throughout, while frequency starts out negative, rises throughout the middle layers, then becomes negative again at later layers. On the other hand, the effects steadily decrease throughout the decoder for both frequency and predictability. It's unclear why the effect of frequency starts negative and then becomes less negative towards the middle layers in the encoder. One possible interpretation is that high-frequency segments are often reduced (e.g., Bybee and Scheibman, 1999): these segments, in terms of acoustics, would be less similar to standalone *up* than low-frequency (un-reduced) forms. The middle layers may be able to abstract across this, reducing the frequency-effect on representation. The decrease in later layers, then, reflects holistic storage (as opposed to phonetic reduction). Future research is needed to confirm this.

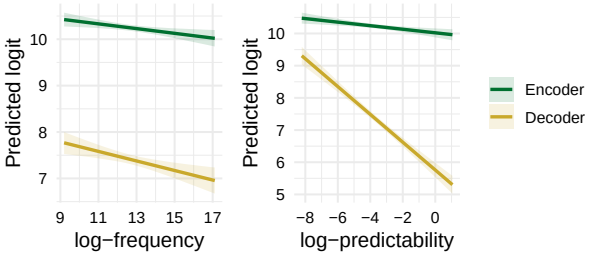

Figure 5: Final-layer brms predicted logit by frequency (left) and predictability (right) for Whisper encoder and decoder. Shading indicates 95% CI.

5.3 Discussion

The results of Whisper demonstrate an interesting similarity between audio speech recognizers and text-based language models: both the encoder and the decoder of Whisper show similar patterns to the LLMs (high-predictability and high-frequency V+*up* phrases showing lower classifier logit scores). This is especially surprising for the encoder, which is traditionally thought to consist of mostly acoustic representations (though see Pluth et al., 2026, for evidence that the encoder may encode semantics

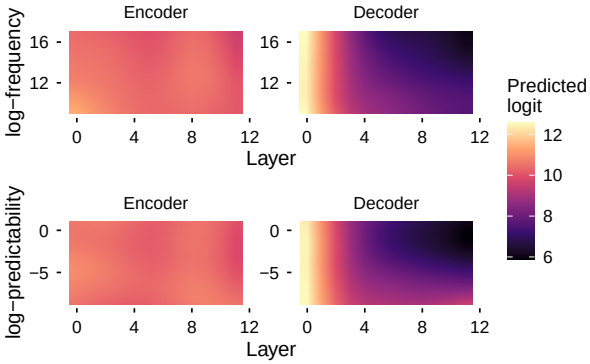

Figure 6: GAM-predicted logit by layer and predictor for Whisper encoder and decoder. Top: frequency; bottom: predictability. Color = logit; darker = lower logit (less similar to standalone *up*); random effects excluded.

and higher-level linguistic information).

6 Conclusion

The present study demonstrates that both text-based LLMs and ASR models represent high-frequency and high-predictability V+*up* phrases holistically, with predictability effects emerging as a function of model size. Notably, this holds even in models trained on an amount of data comparable with humans, suggesting that holistic storage does not require the massive training corpora typical of modern LLMs. Further, the effect of predictability is clearer for larger models, suggesting that increasing the number of parameters facilitates predictability effects. Finally, the results suggest that for larger models, these effects emerge in earlier layers, while for smaller models the effect emerges in later layers. The Whisper results add a further nuance: the effect of holistic storage is similar for both the encoder and decoder, adding to the growing body of literature suggesting the rich representations contained within the encoder (Pluth et al., 2026).

There is further reason to find interest in the similarity between the results for Whisper and the text-based LLMs: in LLMs, the token for *up* is identical across all of our items in testing. On the other hand, since Whisper takes audio as its input, each *up* token is distinct (the audio representations are slightly different). These differences between the input types could in theory result in different representational behavior between the models. Interestingly, however, both models employ similar strategies. It's unclear why exactly this is the case, but one possibility is that the encoder is abstracting away some of the acoustic variability, and representing *up* similarly enough that the model is able to treat these as a single token.

The by-layer results suggest that holistic storage is better understood as a spectrum than a binary. Representations that diverge only in later layers may reflect a form of partial holistic storage, where lower-level representations remain more compositional but higher-level abstractions increasingly integrate item-specific information. Larger models, which have encountered V+*up* phrases more frequently, show more divergent representations even in early hidden layers, suggesting that with sufficient exposure or capacity, storage-like behavior becomes a more pervasive organizational principle.

Holistic storage is itself a form of exemplar-based representation: rather than computing a phrase from its parts each time, the whole phrase is represented as a stored unit. If holistic storage and grammatical abstraction arise from the same process, then the traditional distinction between stored items and productive rules may be a post-hoc description of a continuous representational landscape rather than a reflection of genuinely distinct cognitive processes. The gradient by-layer results support this view: rather than a sharp transition from compositional to holistic representations, we observe a progressive divergence that deepens across layers and strengthens with frequency and predictability. This is more consistent with a view in which storage-like behavior is one end of a continuum than with a view in which some items are stored and others are computed.

Overall, the present study demonstrates another dimension upon which the behavior of transformer models, across modalities, resemble that of humans, providing further evidence for usage-based theories. Specifically, usage-based accounts predict that holistic storage should be a gradient, frequency- and predictability-driven phenomenon (e.g., Bybee and Scheibman, 1999; Houghton, 2025b; Kapatsinski et al., 2009). The present results confirm this: storage-like representations emerge gradually as a function of frequency and predictability, and they arise in models that have no explicit storage mechanism and, in the case of the BabyLMs, have seen no more data than a human. More broadly, the results suggest that transformer architectures may serve as a productive modeling framework for investigating not just whether human-like storage patterns emerge, but *how* and *when* they do so, and at a level of mechanistic detail that behavioral data alone cannot provide.

7 Limitations

The primary limitation is that we examined only one construction (V+*up* phrases) and in one language (English). We look forward to expanding the present study to other languages and other constructions in future work. We also only examined representations at the final checkpoint; however, we look forward to examining the emergence of storage as a function of training dynamics in the future.

References

Jean Berko. 1958. The child's learning of english morphology. WORD, 14(2-3):150–177.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

Joan Bybee and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of don't in english. *Linguistics*, 37(4).

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *arXiv preprint arXiv:2502.10645*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, and 2 others. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*.

Coleman Haley. 2020. This is a bert. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341.

Zara Harmon and Vsevolod Kapatsinski. 2017. Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, 98:22–44.

Zachary Houghton, Misaki Kato, Melissa Baese-Berk, and Charlotte Vaughn. 2024. Task-dependent consequences of disfluency in perception of native and non-native speech. *Applied psycholinguistics*, 45(1):64–80.

Zachary Nicholas Houghton. 2025a. The bolts and nuts of language processing: An investigation into the noisy-channel processing of binomials. Master's thesis.

Zachary Nicholas Houghton. 2025b. *Multi-Word Representations in Minds and Models: Investigating the Storage of Multi-Word Phrases in Humans and Large Language Models*. Ph.D. thesis.

Zachary Nicholas Houghton and Emily Morgan. 2023. Does predictability drive the holistic storage of compound nouns? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Zachary Nicholas Houghton and Emily Morgan. 2024. Frequency-dependent preference extremity arises from a noisy-channel processing model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Zachary Nicholas Houghton, Kenji Sagae, and Emily Morgan. 2025. The role of abstract representations and observed preferences in the ordering of binomials in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 695–702.

Vsevolod Kapatsinski, Joshua Radicke, Roberta Corrigan, Edith A Moravcsik, Hamid Ouali, and Kathleen Wheatley. 2009. Frequency and the emergence of prefabs. *Formulaic language*, 2:499–520.

Karim Lasri, Olga Seminck, Alessandro Lenci, and Thierry Poibeau. 2022. Subject verb agreement error patterns in meaningless sentences: Humans vs. bert. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 37–43.

Roger Levy, Evelina Fedorenko, Mara Breen, and Edward Gibson. 2012. The processing of extraposed structures in english. *Cognition*, 122(1):12–36.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 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.

David E. Meyer and Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 913–929.

Emily Morgan and Roger Levy. 2015. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

Emily Morgan and Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:384–402.

Emily Morgan and Roger Levy. 2024. Productive knowledge and item-specific knowledge trade off as a function of frequency in multiword expression processing. *Language*, 100(4):e195–e224.

Steven Pinker and Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

Dan Pluth, Zachary Nicholas Houghton, Yu Zhou, and Vijay K. Gurbani. 2026. Mechanistic interpretability of asr models using sparse autoencoders. *Preprint*, arXiv:2605.12225.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Joseph P. Stemberger and Brian MacWhinney. 2004. Are inflected forms stored in the lexicon. *Morphology: Critical concepts in linguistics*, 6:107–122.

Joseph Paul Stemberger and Brian MacWhinney. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition*, 14(1):17–26.

Simon N Wood. 2017. *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both direct and indirect evidence contribute to dative alternation preferences in language models. *arXiv preprint arXiv:2503.20850*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A BabyLM Model Training

Models were trained from random initialization for 20 epochs using the AdamW optimizer with fused weight updates and bfloat16 mixed precision. The learning rate was set to 3×10^{-4} for the 125M model and 1×10^{-4} for the 350M and 1.3B models, each preceded by a linear warmup over 10% of total training steps. Training was distributed across two NVIDIA A100 80GB GPUs. Hyperparameter details are given in Table 1.

Parameter	OPT-125M	OPT-350M	OPT-1.3B
Hidden size	768	1,024	2,048
Attention heads	12	16	32
Layers	12	24	24
FFN intermediate size	3,072	4,096	8,192
Vocabulary size	8,192	8,192	8,192
Context length (tokens)	1,024	1,024	1,024
Batch size (per device)	400	200	100
Learning rate	3×10^{-4}	1×10^{-4}	1×10^{-4}
Warmup steps	366	732	1,465
Training epochs	20	20	20
Optimizer	AdamW (fused)	AdamW (fused)	AdamW (fused)
Precision	bfloat16	bfloat16	bfloat16

Table 1: Hyperparameters for the three BabyLM models. All models share a BPE tokenizer (vocabulary size 8,192) trained on the BabyLM corpus. Training used two NVIDIA A100 80GB GPUs with distributed data parallelism.

B BabyLM Model Convergence

All three models trained to completion over 20 epochs, with training loss decreasing monotonically throughout (Figure 7). Final gradient norms were small for all models (0.16, 0.24, and 0.38 for the 125M, 350M, and 1.3B models, respectively), and the learning rate decayed to zero at the end of training, consistent with full convergence under the scheduled warmup-then-decay regime. No signs of loss divergence or instability were observed.

The 1.3B model achieves the lowest training loss (2.55) and validation perplexity (15.18), as expected given its greater capacity (Table 2). The 350M model, however, shows slightly elevated training loss (2.63) and validation perplexity (18.86) relative to the 125M model (training loss 2.55; validation perplexity 16.81). This is likely attributable to the lower learning rate applied to the 350M model (1×10^{-4} vs. 3×10^{-4} for the 125M), which, combined with a relatively modest dataset size of 150M tokens, may have produced slower effective convergence. The effect is small and does not affect our qualitative comparisons, but we note it here for completeness.

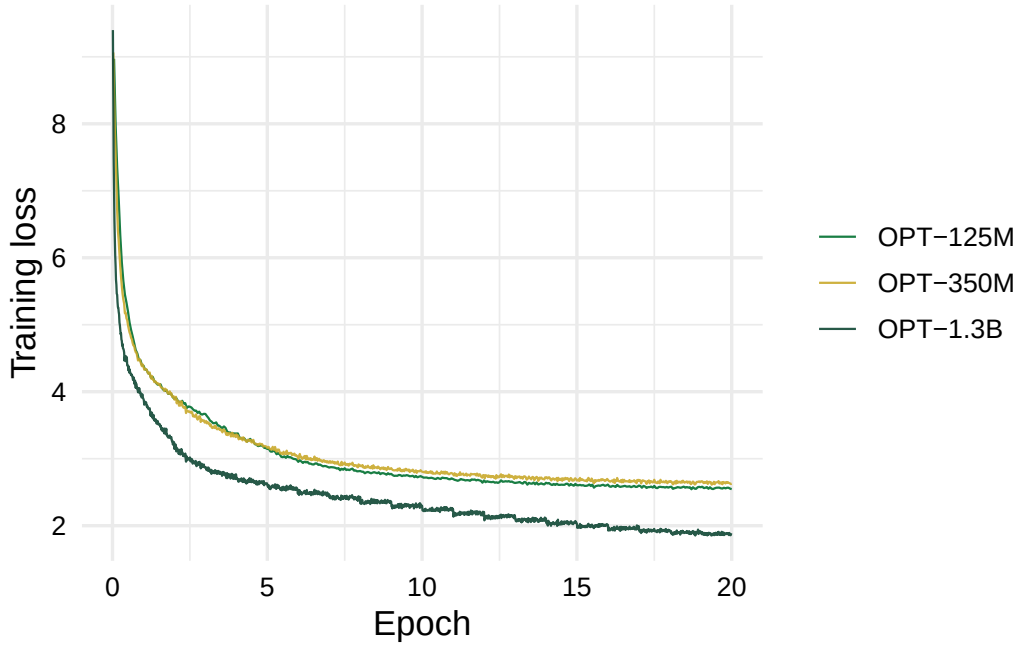

Figure 7: Training loss over 20 epochs for each BabyLM model. Loss is logged every 10 gradient steps.

Model	Val. loss	Perplexity
OPT-125M	2.82	16.81
OPT-350M	2.94	18.86
OPT-1.3B	2.72	15.18

Table 2: Final validation loss and perplexity for each BabyLM model, evaluated on the BabyLM development set after 20 epochs of training.

C Test Set Items

Model	N items	Med. freq	Freq range	Med. predic.	Predic. range
OLMo-3 7B	4081	43608	2–103,842,740	-5.38	-10.98–2.74
BabyLM	1039	34	10–9,963	-3.40	-10.38–1.99
Whisper encoder	1724	169679	960–103,842,740	-4.27	-10.98–2.74
Whisper decoder	1724	169679	960–103,842,740	-4.27	-10.98–2.74

Table 3: Test set statistics for all experiments. Only items with a valid (non-zero) predictability estimate are included. Frequency is the raw corpus count; predictability is the log odds (see Equation 2). BabyLM models (125M, 350M, 1.3B) share identical items and corpus statistics. Whisper encoder and decoder evaluate the same audio items with Dolma-derived frequency counts.

D Experiment 1: UP Independently

Model	Train pos	Train neg	Val pos	Val neg
OLMo-3 7B	1,000	1,000	992	1,000
BabyLM 125M	1,000	1,000	916	995
BabyLM 350M	1,000	1,000	916	995
BabyLM 1.3B	1,000	1,000	916	995

Table 4: Classifier training and validation split sizes for Experiment 1 (UP independently). Positive (pos) = standalone *up*; negative (neg) = other alphabetic token from the same sentence.

Model	Coef	Est.	Err.	95% CI	% > 0
OLMo-3 7B	Intercept	8.16	0.05	[8.06, 8.27]	100.0
OLMo-3 7B	Freq	-0.96	0.06	[-1.07, -0.85]	0.0
OLMo-3 7B	Predic.	-2.66	0.05	[-2.77, -2.56]	0.0
OLMo-3 7B	Freq:Predic.	-0.01	0.05	[-0.12, 0.09]	41.3
BabyLM 125M	Intercept	7.12	0.04	[7.05, 7.19]	100.0
BabyLM 125M	Freq	-0.41	0.04	[-0.48, -0.34]	0.0
BabyLM 125M	Predic.	-0.19	0.04	[-0.26, -0.12]	0.0
BabyLM 125M	Freq:Predic.	-0.12	0.04	[-0.19, -0.05]	0.0
BabyLM 350M	Intercept	7.24	0.04	[7.15, 7.32]	100.0
BabyLM 350M	Freq	-0.53	0.05	[-0.62, -0.44]	0.0
BabyLM 350M	Predic.	-0.10	0.05	[-0.19, -0.01]	1.5
BabyLM 350M	Freq:Predic.	-0.16	0.05	[-0.25, -0.07]	0.0
BabyLM 1.3B	Intercept	8.19	0.05	[8.10, 8.29]	100.0
BabyLM 1.3B	Freq	-0.72	0.05	[-0.83, -0.62]	0.0
BabyLM 1.3B	Predic.	-0.14	0.05	[-0.24, -0.04]	0.4
BabyLM 1.3B	Freq:Predic.	-0.11	0.05	[-0.21, -0.01]	1.8

Table 5: Joint frequency × predictability model at the final layer for UP independently. Estimates are from Bayesian linear mixed-effects models (brms); % > 0 indicates the percentage of posterior samples with a positive estimate.

Model	Predictor	EDF	F	p
OLMo-3 7B	Frequency	23.33	38662.42	< 0.001
BabyLM 125M	Frequency	15.33	16285.99	< 0.001
BabyLM 350M	Frequency	18.71	40985.53	< 0.001
BabyLM 1.3B	Frequency	16.67	29526.15	< 0.001
OLMo-3 7B	Predictability	23.41	43603.71	< 0.001
BabyLM 125M	Predictability	21.92	12518.19	< 0.001
BabyLM 350M	Predictability	22.89	35860.41	< 0.001
BabyLM 1.3B	Predictability	22.38	23924.58	< 0.001

Table 6: GAM tensor-product smooth summary for UP independently. EDF = estimated degrees of freedom; p = p-value for the te(predictor, layer) smooth term.

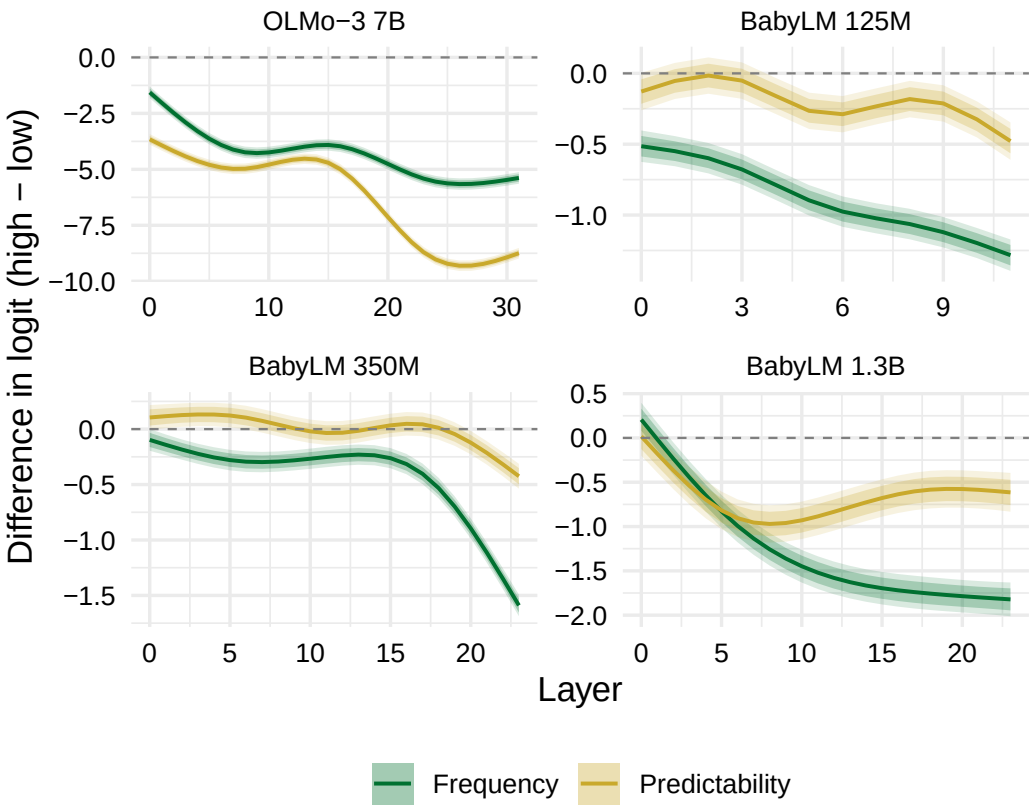

Figure 8: Layer-by-layer difference in GAM-predicted logit between high (90th percentile) and low (10th percentile) predictor values, for UP independently. Green = frequency effect; yellow = predictability effect. Inner and outer shaded ribbons show 80% and 95% confidence intervals derived from the GAM standard error. A negative difference indicates that high-predictor phrases yield a lower classifier logit (i.e., their representation of *up* is less similar to the standalone class).

E Experiment 2: UP as Subword

Model	Train standalone	Train subword	Train neg	Val rows
OLMo-3 7B	1,000	1,000	2,000	3,868
BabyLM 125M	1,000	1,000	2,000	3,343
BabyLM 350M	1,000	1,000	2,000	3,343
BabyLM 1.3B	1,000	1,000	2,000	3,343

Table 7: Classifier training and validation split sizes for Experiment 2 (UP as subword). Positive examples: 1,000 standalone *up* + 1,000 unique up-within-word types. Negative examples: 1,000 from each sentence pool. Val sizes vary by model due to tokenizer coverage.

Model	Coef	Est.	Err.	95% CI	% > 0
OLMo-3 7B	Intercept	10.67	0.05	[10.57, 10.76]	100.0
OLMo-3 7B	Freq	-0.82	0.05	[-0.92, -0.72]	0.0
OLMo-3 7B	Predic.	-1.97	0.05	[-2.06, -1.87]	0.0
OLMo-3 7B	Freq:Predic.	-0.04	0.05	[-0.14, 0.05]	19.1
BabyLM 125M	Intercept	11.60	0.08	[11.45, 11.75]	100.0
BabyLM 125M	Freq	-0.73	0.08	[-0.90, -0.58]	0.0
BabyLM 125M	Predic.	-0.09	0.08	[-0.25, 0.06]	12.2
BabyLM 125M	Freq:Predic.	-0.29	0.08	[-0.44, -0.14]	0.0
BabyLM 350M	Intercept	12.16	0.07	[12.02, 12.29]	100.0
BabyLM 350M	Freq	-0.87	0.07	[-1.01, -0.72]	0.0
BabyLM 350M	Predic.	0.25	0.07	[0.11, 0.39]	100.0
BabyLM 350M	Freq:Predic.	-0.26	0.07	[-0.39, -0.12]	0.0
BabyLM 1.3B	Intercept	10.14	0.05	[10.04, 10.24]	100.0
BabyLM 1.3B	Freq	-0.82	0.05	[-0.92, -0.71]	0.0
BabyLM 1.3B	Predic.	0.14	0.05	[0.04, 0.24]	99.7
BabyLM 1.3B	Freq:Predic.	-0.02	0.05	[-0.12, 0.09]	36.9

Table 8: Joint frequency × predictability model at the final layer for UP as subword. Estimates are from Bayesian linear mixed-effects models (brms); % > 0 indicates the percentage of posterior samples with a positive estimate.

Model	Predictor	EDF	F	p
OLMo-3 7B	Frequency	23.36	43325.50	< 0.001
BabyLM 125M	Frequency	16.97	503.64	< 0.001
BabyLM 350M	Frequency	20.90	5559.92	< 0.001
BabyLM 1.3B	Frequency	19.58	16811.79	< 0.001
OLMo-3 7B	Predictability	23.11	47144.94	< 0.001
BabyLM 125M	Predictability	20.62	349.42	< 0.001
BabyLM 350M	Predictability	21.50	5337.65	< 0.001
BabyLM 1.3B	Predictability	21.82	15592.30	< 0.001

Table 9: GAM tensor-product smooth summary for UP as subword. EDF = estimated degrees of freedom; p = p-value for the te(predictor, layer) smooth term.

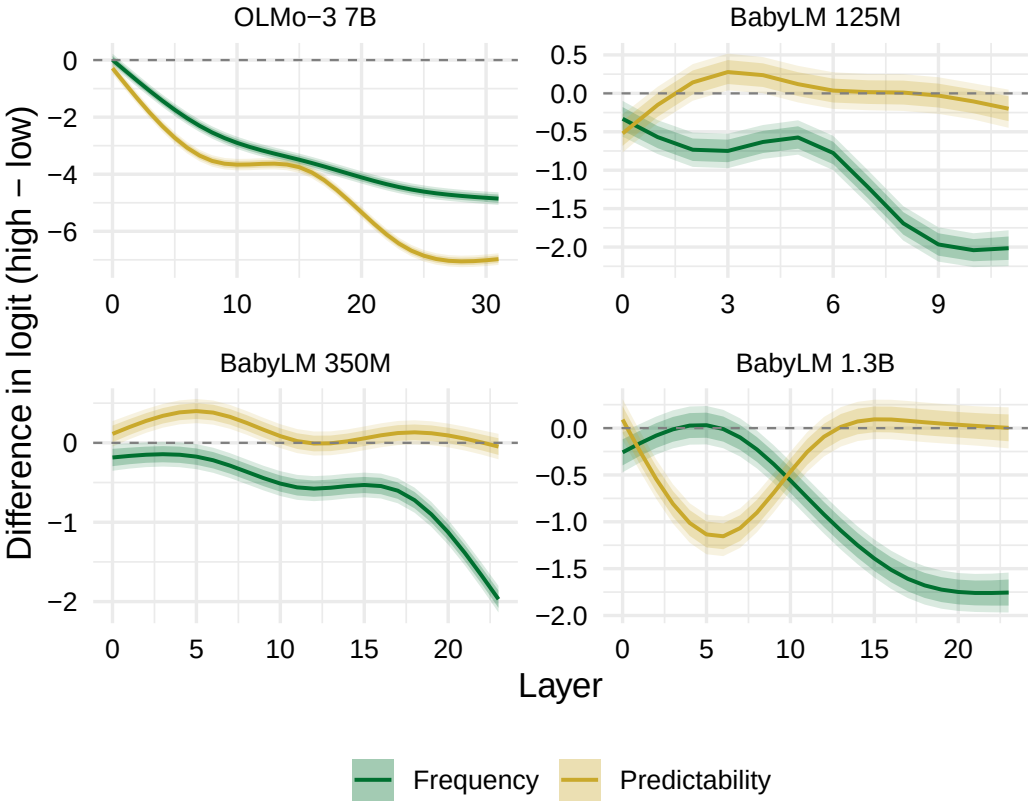

Figure 9: Layer-by-layer difference in GAM-predicted logit between high (90th percentile) and low (10th percentile) predictor values, for UP as subword. Green = frequency effect; yellow = predictability effect. Inner and outer shaded ribbons show 80% and 95% confidence intervals derived from the GAM standard error. A negative difference indicates that high-predictor phrases yield a lower classifier logit (i.e., their representation of *up* is less similar to the standalone class).

F Experiment 3: Whisper

Component	Train pos	Train neg	Val pos	Val neg
Encoder	1,000	1,000	1,000	1,000
Decoder	1,000	1,000	1,000	1,000

Table 10: Classifier training and validation split sizes for Experiment 3 (Whisper). Positive examples are *word_up* instances; negatives are random non-*up* alphabetic tokens from the same segment.

Model	Coef	Est.	Err.	95% CI	% > 0
Encoder	Intercept	10.23	0.03	[10.17, 10.30]	100.0
Encoder	Freq	-0.09	0.03	[-0.15, -0.02]	0.3
Encoder	Predic.	-0.11	0.03	[-0.17, -0.04]	0.1
Encoder	Freq:Predic.	0.02	0.03	[-0.05, 0.08]	68.6
Decoder	Intercept	7.40	0.05	[7.30, 7.51]	100.0
Decoder	Freq	-0.17	0.05	[-0.27, -0.07]	0.1
Decoder	Predic.	-0.83	0.05	[-0.93, -0.72]	0.0
Decoder	Freq:Predic.	-0.07	0.05	[-0.17, 0.03]	9.3

Table 11: Joint frequency × predictability model at the final layer for Whisper encoder and decoder. Estimates are from Bayesian linear mixed-effects models (brms); % > 0 indicates the percentage of posterior samples with a positive estimate.

Model	Predictor	EDF	F	p
Encoder	Frequency	15.95	52.81	< 0.001
Decoder	Frequency	21.24	9161.40	< 0.001
Encoder	Predictability	18.92	45.48	< 0.001
Decoder	Predictability	22.46	9115.45	< 0.001

Table 12: GAM tensor-product smooth summary for Whisper encoder and decoder. EDF = estimated degrees of freedom; p = p-value for the te(predictor, layer) smooth term.

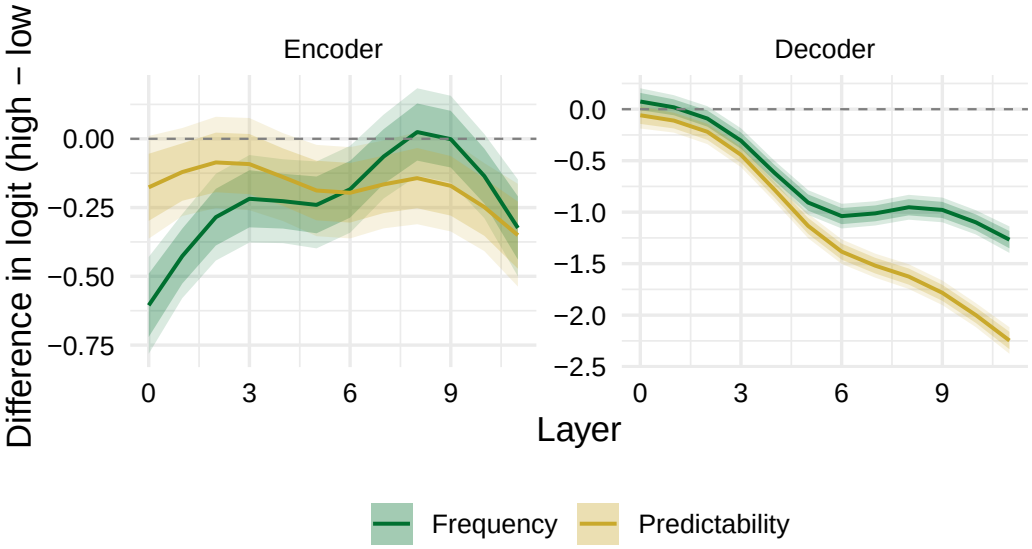

Figure 10: Layer-by-layer difference in GAM-predicted logit between high (90th percentile) and low (10th percentile) predictor values, for Whisper encoder and decoder. Green = frequency effect; yellow = predictability effect. Inner and outer shaded ribbons show 80% and 95% confidence intervals derived from the GAM standard error. A negative difference indicates that high-predictor phrases yield a lower classifier logit (i.e., their representation of *up* is less similar to the standalone class).