Emergent Abstract Ordering Preferences in Large Language Models

Zachary Nicholas Houghton Kenji Sagae Emily Morgan

1 Introduction

It is uncontroversial that large language models (LLMs) have demonstrated better language proficiency than any other computational model of language in history. Their historic rise to fame has brought with it many heated debates regarding whether large language models constitute human-like models of language or whether what they are doing is completely different from humans (Bender et al., 2021; Piantadosi, 2023; Piantadosi & Hill, 2022).

Many of these debates have centered around the tradeoff between computation and storage, that is, how much these models are simply reproducing from their training data vs how much of their productions are novel utterances using learned linguistic patterns. On one hand, there is no doubt that large language models store and reproduce large chunks of language. In fact, OpenAI has even been sued by the *New York Times* for allegedly reproducing entire articles verbatim (*The New York Times Company v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI, LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC*, 2024). This sentiment – that large language models are nothing but glorified copy cats – has been echoed by several other prominent linguists (Bender et al., 2021; Bender & Koller, 2020; c.f., Piantadosi, 2023).

Specifically, proponents of the "LLMs as copy cats" argument have pointed out that large language models are trained on an inconceivably large amount of data. For example, the OLMo models were trained on trillions of tokens (Groeneveld et al., 2024)[1]. As such, it is extremely difficult to determine whether utterances by an LLM are truly novel, or whether they are reproduced from their training data. This is further complicated by the fact that training data for LLMs is typically either not publicly available, or so huge that it's incredibly difficult to work with. On the other hand, it is clear that large language models are learning some linguistic patterns. For example, McCoy et al. (2023) demonstrated that Chat GPT-2 is able to generate well-formed novel words as well as well-formed novel syntactic structures, despite also finding that it still copies extensively.

A similar debate in the field has centered around whether large language models learn any knowledge about the meaning of words. For example, Bender & Koller (2020) have argued that large language models, which are only trained on the form, have no way of learning anything about meaning. However, Piantadosi & Hill (2022) have countered by pointing out that co-occurrence statistics can be extremely informative about a word's meaning. For example, they argued that many words, such as "justice", contain no clear referent and instead have to be learned by humans based on the context that they occur in. It seems plausible that large language models could learn at least some information about meanings similarly.

These debates, however, have been highly theoretical and speculative and very few empirical studies have been done to actually investigate these questions (c.f., LeBrun et al., 2022; McCoy et al., 2023; Wei et al., 2022). Thus in the present paper we address these debates by taking an in-depth look at large language models' abilities to abstract across their training.

Our specific contributions are as follows: We make a 3-grams corpus of Dolma (Soldaini et al., 2024) along with the scripts

[1] This is magnitudes larger than the 350 million words that the average college-aged speaker has seen in their lifetime (Levy et al., 2012)

to reproduce it open-access. We also use this corpus to create novel binomials that the OLMo 7B model (Groeneveld et al., 2024) has never seen and demonstrate that OLMo shows evidence of using learned abstract ordering preferences. Finally, we demonstrate a timescale of these preferences emerging over training that can be used to generate predictions about human learning.

1.1 Abstractions in Large Language Models

The evidence for learned abstractions in large language models is extremely mixed. For example, Haley (2020) demonstrated that many of the BERT models are not able to reliably determine the plurality of novel words. Specifically, they tasked BERT with choosing between two forms of a novel noun: the plural form and the singular form. They tested BERT on 5 different languages: English, German, Dutch, Spanish, and French. Interestingly, they also had a condition which contained a prime. In English, the prime sentence was: "This is a _____", with the singular form of the noun replacing the blank space. Humans are able to use this contextual information to determine the plural form of the noun reliably, and Haley (2020) argued that it is theoretically possible for BERT to learn to do so as well using self-attention. While BERT was able to perform better-than-chance on novel nouns, cross-linguistically it failed to use information about the prime sentence to achieve better performance.

On the other hand, Wei et al. (2022) demonstrated that BERT can generalize well to novel subject-verb pairs. Specifically, they tested BERT's performance on novel sentences along with semantically incoherent but syntactically sensible sentences (e.g., *colorless green ideas sleep furiously*). They found that BERT performs well on items it wasn't trained on but still struggles with low-frequency lexical items. Similarly, as mentioned earlier, McCoy et al. (2023) examined to what extent GPT-2 was simply copying its training data vs producing novel utterances. They found that while GPT-2 copies extensively, it also produces both novel words as well as novel syntactic structures.

Finally, there is evidence that transformer models can learn abstractions from other domains as well. For example, Tartaglini et al. (2023) examined the ability of a transformer model in a same-different task (i.e., determining if two entities, e.g., two shapes, in an image are the same or different). They found that certain models can reach near perfect accuracy on items they have never seen before. They argued that this demonstrates their abilities to learn abstract representations.

1.2 Abstractions In Humans

Abstractions have been a part of just about every linguistic theory out there, including both generativist and non-generativist theories. This is for good reason, too: one of the hallmarks of human language learning is the ability to produce novel, never-heard-before utterances. In order to do so, most theories posit that humans leverage their remarkable ability to learn linguistic patterns beyond simple co-occurrence rates (c.f., Ambridge, 2020). For example, when presented a novel noun, children are able to consistently produce the proper plural form of that noun (Berko, 1958). Similarly, children are able to abstract across different contexts to learn a word's general meaning (Yu & Smith, 2007).

Abstractions are useful because when humans produce a novel utterance that they have never heard before, their novel utterances contain a level of systematicity that allows the interlocutor to understand it with very little difficulty. This is even the case for binomials (e.g., *cat and dog*), whose order does not particularly affect the meaning of the utterance. Binomial ordering preferences are well-documented in the literature. For example, Morgan & Levy (2016) demonstrated that humans show ordering preferences for binomials beyond simply preferring the more frequent ordering (e.g., preferring male-coded words before female-coded words). They coded a list of binomials for a variety of semantic constraints, phonological constraints, and metric constraints that affect human ordering preferences for binomials (Benor & Levy, 2006). They found that humans ordering preferences were driven by abstract ordering preferences, such as a preference to place short words before longer words, even after accounting for the relative frequency (which ordering preference occurs more in corpus data). In other words, human ordering preferences are driven by both the observed preferences in corpus data (i.e., the number of times they've encountered each ordering of the binomial) as well as by abstract ordering preferences (Morgan & Levy, 2016). Specifically, they developed a model to quantify the abstract ordering preferences of humans for a given binomial in English. The model predicts the probability that a binomial expression is realized as *A and B* (the alphabetical form was used as a neutral reference order) as a function of constraints that have been shown to influence binomial ordering preferences in humans (e.g. a preference to place more culturally-central words first, Benor & Levy, 2006).

Morgan & Levy (2016) demonstrated that the model's predicted abstract ordering preferences are not the same as the observed preferences in corpus data (i.e., the model wasn't simply predicting the more frequent ordering). Despite this, however, they showed in both a forced-choice task and a self-paced reading task that abstract ordering preferences drive, to some extent, both novel and attested binomial orderings. In other words, the ordering preference for a specific binomial cannot be predicted purely from the proportion of occurrences in the alphabetical order to the occurrences in nonalphabetical. This suggests that humans are not simply reproducing their input, but learning abstract ordering preferences from the data.

1.3 Present Study

In the present study we examine whether large language models are simply copying their input, or whether they are learning more abstract linguistics patterns. We use binomials as a test case because human ordering preferences deviate from the observed preferences for them (that is, humans don't simply prefer the more frequent ordering).

Specifically, in Experiment 1 we examine whether OLMo's 7B model (Groeneveld et al., 2024) is sensitive to abstract ordering preferences overall. We also examine the individual constraints that go into calculating abstract ordering preference, such as the preference for short words before long words, to determine whether OLMo is sensitive to the same specific constraints as humans. In Experiment 2, we examine the same questions at different learning stages of the model's learning in order to determine how these abstract ordering preferences emerge as a function of the training.

2 Dataset

2.1 Dolma

For both experiments, we use the dataset described in this section. In order to examine whether large language models learn preferences above and beyond simply memorizing co-occurrence rates, we created a 3-grams corpus of Dolma (Soldaini et al., 2024). Specifically, we used Dolma version 1_7, which was used to train OLMo-7B-v1.7 (Groeneveld et al., 2024), a large language model trained on 2.05 trillion tokens. Our corpus contains every 3-gram (ignoring punctuation and capitalization) in the Dolma corpus, as well as the number of times that 3-gram appeared. Similarly, in order to calculate individual word-frequencies we also created a 1-gram corpus of Dolma.

We then created a list of binomials and searched the corpus to find a list of binomials that did not occur in the Dolma corpus. We eliminated binomials which occurred more than zero times in either their alphabetical or nonalphabetical orderings. Thus, OLMo has had no experience with either ordering of any of our binomials. Individual word frequencies were also calculated for these items using the 1-grams corpus of Dolma. Our full list of items is presented in full in the appendix section (Section Section A).

2.2 Abstract Ordering Preferences

In order to examine whether large language models are learning preferences similar to humans, we calculated the abstract ordering preference value for each of our binomials (following Morgan & Levy, 2016). Morgan & Levy (2016) demonstrated that their model's estimated abstract ordering preference value is a significant predictor of human binomial ordering preferences, even after accounting for the frequency of the binomial. abstract ordering preference is calculated from a mix of semantic and phonological properties that human binomial ordering preferences have been shown to be sensitive to (Benor & Levy, 2006). For each of these constraints, a positive value indicates a preference for the alphabetical word to be placed first (following the given constraint). A negative value indicates a preference for the nonalphabetical word to be placed first. For example, a positive value of *Freq* indicates that the alphabetical word is more frequent and thus is predicted to be placed first, while a negative value indicates that the nonalphabetical word is more frequent. The constraints are as follows:

- **Formal Markedness**: The word with more general meaning or broader distribution comes first. For example, in *boards and two-by-fours*, boards are a broader class of which two-by-fours is one member.
- **Perceptual Markedness**: Elements that are more closely connected to the speaker come first. This constraint encompasses Cooper & Ross (1975)'s (1975) 'Me First' constraint and includes numerous subconstraints, e.g.: animates precede inanimates; concrete words precede abstract words. For example, in *deer and trees*, deer are animate while trees are inanimate.
- **Cultural Centrality**: More culturally central or common elements appear first. For example, in *see and hear* seeing is the more salient form of perception.
- **Power**: The more powerful or culturally prioritized word comes first. For example, in *clergymen and parishioners*, clergymen have higher rank within the church.
- **Intensity**: Elements with more intensity appear first.
- **Iconicity**: When two elements are sequential they should appear in the appropriate sequence.
- **Frequency**: The more frequent element should appear first.
- **Length**: The shorter word should appear first.
- **Lapse:** Avoid multiple unstressed syllables.
- **Final Stress**: The final syllable of the second word should not be stressed.

3 Experiment 1

In Experiment 1, we examine the extent to which OLMo-7B's ordering preferences are driven by abstract ordering preferences for novel binomials. In order to do so, we obtain ordering preferences for each of the binomials in our dataset which have been coded for the abstract ordering preference constraints. If OLMo has learned any abstract ordering preferences, it should order binomials non-randomly. If it is just reproducing the binomials in ordering based purely off the frequency of the items in its input, we should see only an effect of frequency (i.e., it should simply place the more frequent word first).

3.1 Methods

3.1.1 Language Model Predictions

For each model, we calculated the ordering preferences of the alphabetical form (a neutral reference point) for each binomial in the dataset. The predicted probability of the alphabetical form was calculated as the product of the model's predicted probability of each word in the binomial. In order to accurately calculate the probability of the first word in the binomial, each binomial was given the prefix "Next item:". Thus the probability of the alphabetical form, *A and B,* is:

$$
\begin{aligned}
P_{alphabetical} &= P(A|Next\ item:) \\
&\times P(and|Next\ item:A) \\
&\times P(B|Next\ item:A\ and)
\end{aligned}
\tag{1}
$$

where *A* is the alphabetically first word in the binomial and *B* is the other word. Similarly, the probability of the nonalphabetical form, *B and A*, is:

$$
\begin{aligned}
P_{nonalphabetical} &= P(B|Next\ item:) \\
&\times P(and|Next\ item:B) \\
&\times P(A|Next\ item:B\ and)
\end{aligned}
\tag{2}
$$

Finally, we calculated the log odds ratio of the probability of the alphabetical form to the probability of the nonalphabetical form to obtain a single numeric value representing the overall ordering preference for a given binomial. A larger positive value represents a preference for the alphabetical form and a larger negative value represents a preference for the nonalphabetical form:

$$
LogOdds(AandB) = log(\frac{P_{alphabetical}}{P_{nonalphabetical}})
$$

3.1.2 Analyses

We present two mixed-effects analyses using Bayesian linear regression models, implemented in *brms* (Bürkner, 2017) with weak, uninformative priors. For each of our models, the intercept represents the grand mean and the coefficient estimates represent the distance from the grand mean. Bayesian statistics don't force us into a binary interpretation of significance or non-significance, however we can consider an estimate to be statistically significant if the credible interval for that estimate excludes zero.

For both analyses, the dependent variable is $LogOdds(AandB)$, which was described above. Our dependent variable in the first analysis is the abstract ordering preference for each binomial ($AbsPref$). Our dependent variables in the second analysis are the individual constraints that are used to calculate $AbsPref$. The model equations are below in Equation 3 and Equation 4. Note that Formal Markedness and Iconicity were dropped from the second model because the constraint values were zero for all of the binomials.

$$
LogOdds(AandB) \sim AbsPref \tag{3}
$$

$$
LogOdds(AandB) \sim Percept + Culture + Power + Intense + Freq + Len + Lapse + FinalStress \tag{4}
$$

3.2 Results

The results for the first analysis are presented below in Table 1. Our results suggest that there is a main-effect (β=2.538, CI-2.5 = 0.273, CI-97.5 = 5.171) of abstract ordering preference for OLMo's 7B model. A visualization of these results can be found below in Figure 1.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	-1.368	0.633	-2.673	-0.223
AbsPref	2.538	1.275	0.273	5.171

Table 1: Model results examining the effect of AbsPref on LogOdds(AandB).

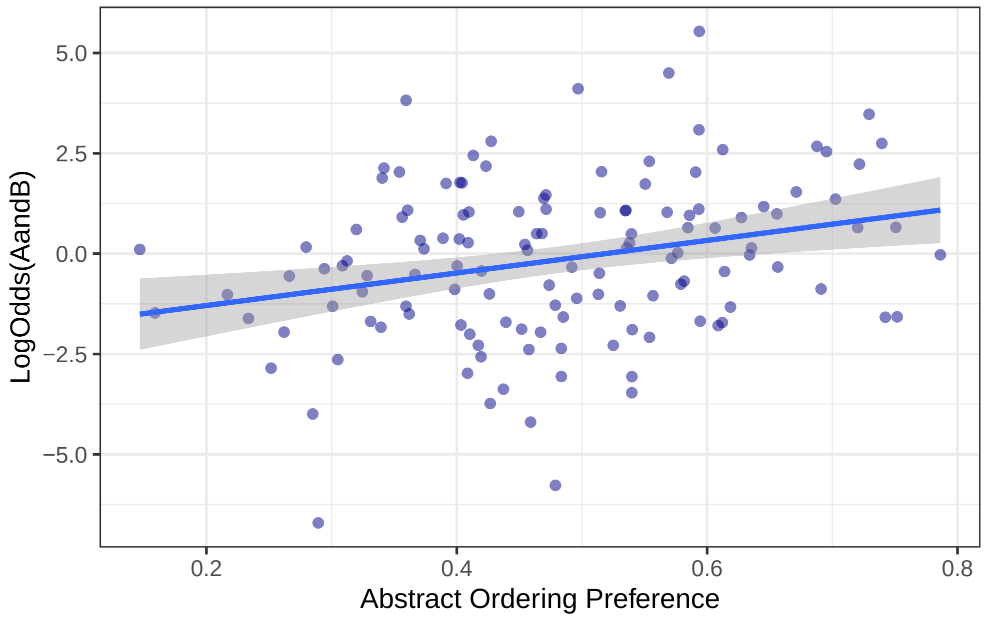

Figure 1: Visualization of the effects of AbsPref on LogOdds(AandB)

While these results suggest that the large language models' ordering preferences are sensitive to similar factors as humans, it's unclear whether this similarity holds on the level of the individual constraints. Thus, in the second analysis we examine which specific constraints the model is sensitive to, and to what extent. For this analysis, following Houghton et al. (2024), we also present the percentage of posterior samples greater than zero. The results of this analysis can be found below in Table 2. Further, a visualization can be found below in Figure 2.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.100	0.172	-0.440	0.235	27.940
Percept	0.185	0.253	-0.314	0.682	76.890
Culture	0.341	0.276	-0.198	0.888	89.260
Power	0.723	0.293	0.155	1.302	99.390
Intense	0.120	0.422	-0.713	0.962	61.170
Freq	0.093	0.091	-0.085	0.268	84.580
Len	-0.204	0.153	-0.502	0.098	9.105
Lapse	-0.110	0.307	-0.721	0.485	36.395

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Final Stress	0.170	0.383	-0.577	0.931	67.175

Table 2: Model results examining the effect of each individual constraint on LogOdds(AandB).

Perhaps surprisingly, the model is most sensitive to the Power constraint (β = 0.723, CI-2.5 = 0.293, CI-97.5 = 0.155) , however there appears to be a weak effect of Culture as well, since 89% of the posterior samples are greater than zero despite the credible interval crossing zero (β = 0.341, CI-2.5 = 0.276, CI-97.5 = -0.198). Surprisingly, there also appears to be a negative effect of length, with a slight preference to place the longer word first, which is the opposite direction from what we see in humans.

3.3 Discussion

Our results demonstrate that OLMo-7B has learned abstract ordering preferences for binomials beyond simply the frequencies of the individual words. Specifically, we find a main-effect of abstract ordering preferences on the model's binomial ordering preferences. Additionally, we find a strong preference to place the more powerful word first, a weak preference to place the more culturally central word first, and a weak preference to place the longer word first.

These results together suggest that the model is learning abstract ordering preferences but these are not identical to humans. For example, while humans also show a preference for placing the more powerful and more culturally central words first, humans also prefer to place the *shorter* word first (Morgan & Levy, 2015, 2016). However, we find the opposite finding: large language models prefer to place the longer word first. One explanation for this is a difference in terms of the input between humans and large language models. The length constraint is determined by the number of syllables. This is salient information in the audio cues that humans receive during learning [**NEED CITATION**], but it's less clear how salient of a cue this is for large language models, who receive sub-word tokens (which vary in their size, from being individual orthographic symbols to being entire words[2]).

4 Experiment 2

In Experiment 1 we demonstrated that large language models are not simply copying their training, but are learning some abstract ordering preferences from their input. However, OLMo makes public various checkpoints during the model's training, thus allowing us the opportunity to examine how these preferences arise as a function of the training. Thus, in Experiment 2 we examine the evolution of these learned abstract ordering preferences as the model learns over time.

4.1 Methods

4.1.1 Language Model Predictions

Our language model predictions in Experiment 2 were obtained using the same procedure as in Experiment 1. However, instead of calculating these metrics only for the main model, we calculated them at various checkpoints. These checkpoints are listed below, in terms of the steps as well as the number of billions of tokens the model had been trained on at that checkpoint:

- Step 0, 0B Tokens

[2] For example, both *ictionary* and *region* are individual tokens in OpenAI's models (https://gist.github.com/s-macke/ae83f6afb89794350f8d9a1ad8a09193).

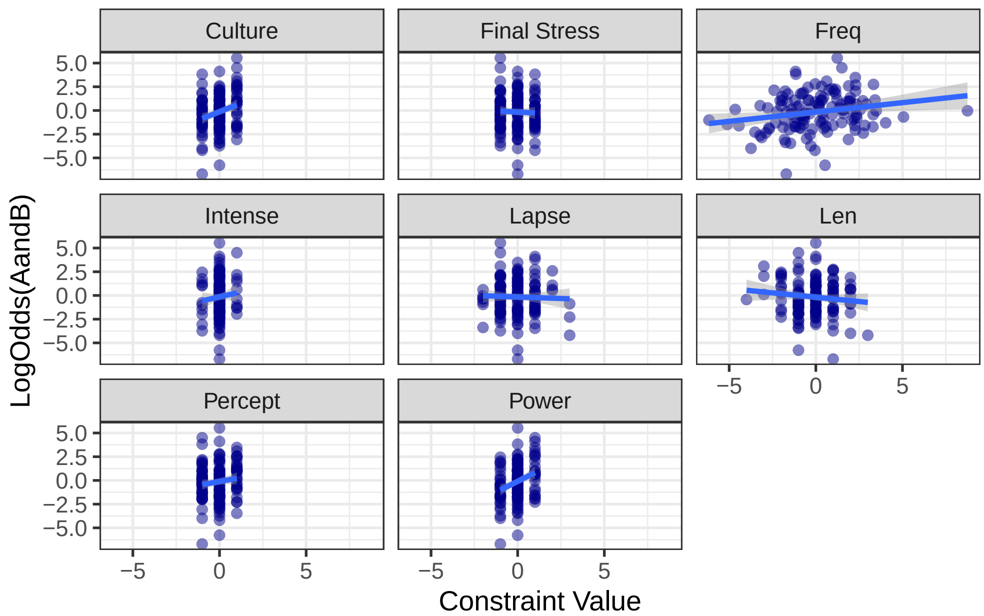

Figure 2: Visualization of the effects of each individual constraint on LogOdds(AandB).

- Step 1000, 2B Tokens
- Step 10000, 41B Tokens
- Step 50000, 209B Tokens
- Step 100000, 419B Tokens
- Step 200000, 838B Tokens
- Step 400000, Tokens 1677B

4.1.2 Analysis

We ran the same two analyses as in Experiment 1, however, we ran these analyses for the each of the checkpoints listed above.

4.2 Results

Our model estimates for the effect of GenPref on LogOdds(AandB) at each checkpoint are presented below in Table 3 and visualized in Figure 3.

Number of Tokens	Estimate	Est.Error	Q2.5	Q97.5
0B	0.396	1.149	-1.912	2.606
2B	1.226	1.375	-0.956	4.569

Number of Tokens	Estimate	Est.Error	Q2.5	Q97.5
41B	0.869	1.157	-1.066	3.482
209B	1.046	1.052	-0.765	3.379
419B	1.699	1.229	-0.313	4.415
838B	1.776	1.248	-0.322	4.505
1677B	3.872	1.520	0.956	6.872

Table 3: Model results examining the effect of AbsPref on LogOdds(AandB) for each checkpoint.

The model results are visualized below in Figure 3.

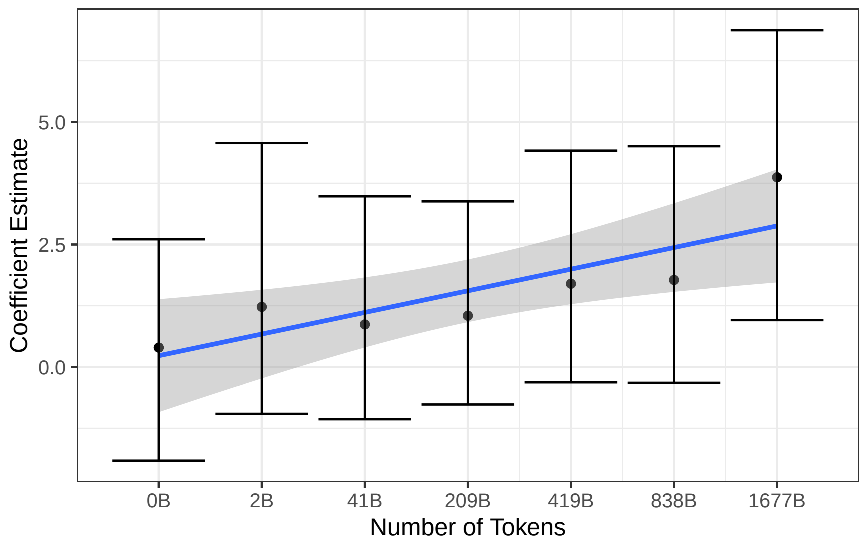

Figure 3: Visualization of the model predictions for the effect of AbsPref on LogOdds(AandB) for each checkpoint.

Our results demonstrate that it takes quite a large number of tokens for the model to learn the abstract ordering preferences. As Figure 3 demonstrates, the effect of abstract ordering preference isn't convincing until the model has experienced 838 billion tokens. However, it does appear that the model develops a slight preference quite rapidly. For example, by 2 billion tokens there appears to be a very slight (though unconvincing) effect of abstract ordering preferences on the ordering of binomials.

Similar to Experiment 1, in our second analysis we present a breakdown of the effects of each individual constraint. In this analysis, however, we demonstrate the effect of each constraint at each checkpoint. The full table results can be found in the Section B, but we present a visualization below in Figure 4.

Interestingly, it appears that early on the model already shows evidence of learning human-like preferences. For example, by 10 billion tokens, the model has learned to place more intense words first, shorter words first, and more powerful words first. However, the model seems slower to learn to place more culturally central words first. Further, as it receives more training the effect of length undergoes a reversal in direction.

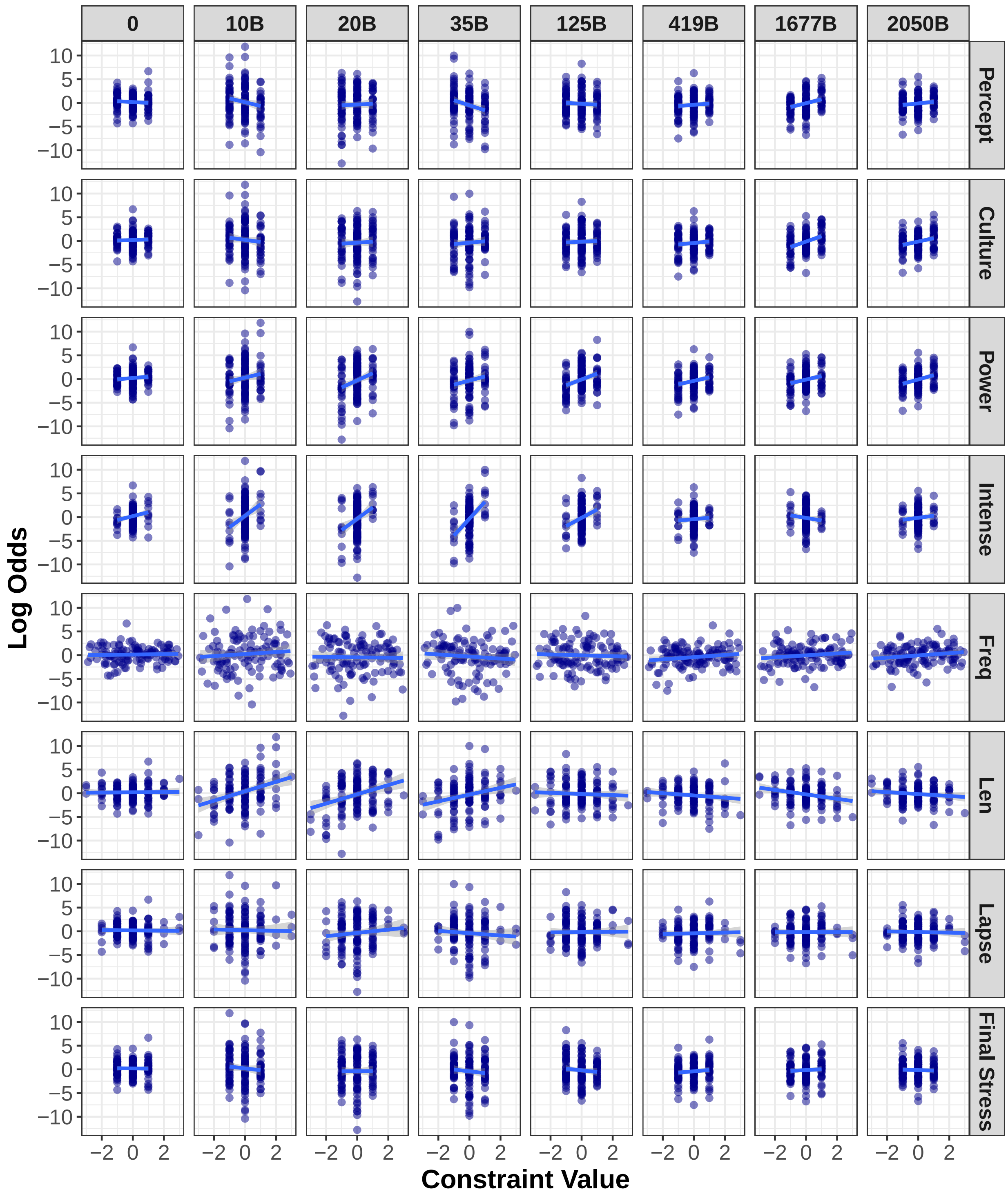

Figure 4: Visualization of the effect of each constraint on the ordering preference at each checkpoint.

4.3 Discussion

Our results demonstrate that OLMo learns human-like ordering preferences early on for most of the constraints, but takes longer to learn human-like ordering preferences for the culture constraint. Further, the model is human-like in its predictions for length early on, but as it receives more training data it learns the opposite length prediction. It is unclear what exactly is causing this reversal, but as we suggested earlier it may be a function of the tokenization differences between human input and large language models' input. We look forward to examining this question in more depth in future studies.

Our results can also be interpreted as predictions for human data. For example, it takes the model longer to learn the Culture constraint than the other constraints. Is the same true for humans?

5 Conclusion

In the present study, we examined the ordering preferences in OLMo 7B's main model as well as the model at various stages in learning. We found that the main model shows human-like ordering preferences, with the exception of a preference for longer words before shorter words. Further, we show that while the effect of abstract ordering preference on a whole takes a great deal of time (over 400 billion tokens to be convincing), the model seems to pick up on individual constraints quite early on, and initially even learns the correct direction of the length constraint.

Our results suggest that large language models are not simply copying their input, but are learning interesting, human-like phenomena from their training. However, they are not learning identically to humans, as demonstrated by the opposite direction of the length preference. This is not surprising given the differences in tokenization methods.

6 Limitations

The main limitation is the number of models tested. We only tested one model in this study, so it's possible that other large language models may demonstrate different ordering preferences. However, we believe that the advantages of demonstrating an in-depth analysis of a single model outweighs a more broad analysis of several models, especially given the lack of easily available open access training data, which is crucial to determining ordering preferences.

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509–559. https://doi.org/10.1177/0142723719869731

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, *82*(2), 233–278. https://doi.org/10.1353/lan.2006.0077

Berko, J. (1958). The Child's Learning of English Morphology. WORD, *14*(2-3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 128. https://www.jstatsoft.org/article/view/v080i01

Cooper, W. E., & Ross, J. R. (1975). World order. *Papers from the Parasession on Functionalism*, *11*, 63–111.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. (2024). Olmo: Accelerating the science of language models. *arXiv Preprint arXiv:2402.00838.*

Haley, C. (2020). This is a BERT. Now there are several of them. Can they generalize to novel words? *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 333–341.

Houghton, Z., Kato, M., Baese-Berk, M., & Vaughn, C. (2024). Task-dependent consequences of disfluency in perception of native and non-native speech. *Applied Psycholinguistics*, 1–17. https://doi.org/10.1017/S0142716423000486

LeBrun, B., Sordoni, A., & O'Donnell, T. J. (2022). Evaluating distributional distortion in neural language modeling. *arXiv Preprint arXiv:2203.12788.*

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in english. *Cognition*, *122*(1), 12–36. https://doi.org/10.1016/j.cognition.2011.07.012

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.

Morgan, E., & Levy, R. (2015). *Modeling idiosyncratic preferences : How generative knowledge and expression frequency jointly determine language structure*. 1649–1654.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv Preprint arXiv:2208.02957.*

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv Preprint arXiv:2402.00159.*

Tartaglini, A. R., Feucht, S., Lepori, M. A., Vong, W. K., Lovering, C., Lake, B. M., & Pavlick, E. (2023). Deep neural networks can learn generalizable same-different visual relations. *arXiv Preprint arXiv:2310.09612.*

The New York Times Company v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI, LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC. (2024). Civil Action No. 1:23-cv-11195-SHS, United States District Court, Southern District of New York, May 31, 2024.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv Preprint arXiv:2206.07682.*

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Appendices

A Full List of Stimuli

Below is a table of our list of binomials as well as the individual constraint values for each.

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
kiwis	wolverines	0	0	0	-1	-1	0	-0.291	1	-1	-1	0.427
kiwis	narwhals	0	1	0	-1	-1	0	2.715	0	-1	-1	0.515
kiwis	ocelots	0	0	0	-1	0	0	2.737	0	-1	-1	0.458
ibex	kiwis	0	0	0	1	0	0	-1.171	0	1	0	0.497
harpies	kiwis	0	-1	0	1	1	0	-1.948	0	0	0	0.471
axolotls	wolverines	0	-1	0	-1	0	0	-2.689	-1	-1	-1	0.262
axolotls	ibex	0	-1	0	-1	0	0	-1.228	-2	-1	0	0.331
axolotls	harpies	0	1	0	-1	-1	0	-0.450	-2	0	0	0.413
axolotls	keas	0	0	1	-1	0	0	1.052	-3	-1	1	0.591
axolotls	bonobos	0	-1	0	-1	0	0	-0.895	-1	0	0	0.328
axolotls	wombats	0	-1	0	-1	0	0	-0.655	-2	-1	-1	0.266
axolotls	lions	0	-1	-1	-1	-1	0	-5.114	-2	0	0	0.159
ocelots	platypuses	0	1	-1	1	0	0	0.665	1	3	1	0.525
ibex	platypuses	0	1	0	1	0	0	2.232	2	3	1	0.691
harpies	platypuses	0	0	0	1	1	0	1.454	2	2	0	0.585
keas	platypuses	0	0	-1	0	0	0	-0.048	3	3	1	0.459
bonobos	platypuses	0	1	0	1	0	0	1.898	1	2	0	0.612
harpies	wolverines	0	-1	0	0	0	0	-2.239	1	-1	1	0.571
keas	wolverines	0	-1	-1	-1	0	0	-3.741	2	0	0	0.285
bonobos	wolverines	0	0	0	0	0	0	-1.795	0	-1	-1	0.426
capybaras	ibex	0	0	0	0	0	0	-1.659	-2	-1	-1	0.356
capybaras	harpies	0	1	0	-1	-1	0	-0.882	-2	0	0	0.404
capybaras	keas	0	1	1	1	0	0	0.621	-3	-1	-1	0.593

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
ibex	narwhals	0	1	0	-1	0	0	1.544	0	0	0	0.536
harpies	narwhals	0	-1	0	0	1	0	0.767	0	-1	-1	0.423
keas	narwhals	0	0	-1	-1	0	0	-0.736	1	0	0	0.362
ibex	ocelots	0	0	0	0	0	0	1.566	1	0	0	0.576
keas	ocelots	0	-1	-1	-1	0	0	-0.714	2	0	0	0.340
bonobos	ocelots	0	0	1	0	0	0	1.233	0	-1	-1	0.594
harpies	ibex	0	-1	0	0	1	0	-0.777	0	-1	-1	0.391
ibex	keas	0	1	1	1	0	0	2.280	-1	0	0	0.729
bonobos	ibex	0	0	0	0	0	0	-0.333	-1	-1	-1	0.419
ibex	koalas	0	0	-1	0	0	0	-0.433	1	1	1	0.471
ibex	sloths	0	0	-1	1	0	0	-0.035	-1	0	0	0.427
aardvarks	ibex	0	0	1	0	0	0	-1.945	0	0	0	0.568
harpies	keas	0	-1	1	1	1	0	1.503	-1	-1	-1	0.569
bonobos	harpies	0	1	0	-1	-1	0	0.444	-1	0	0	0.470
harpies	koalas	0	-1	-1	0	1	0	-1.210	1	0	0	0.359
harpies	wombats	0	-1	0	1	1	0	-0.204	0	-1	-1	0.467
aardvarks	harpies	0	1	0	0	-1	0	-1.167	0	1	1	0.579
bonobos	keas	0	1	1	1	0	0	1.947	-2	-1	-1	0.656
keas	koalas	0	-1	-1	-1	0	0	-2.712	2	1	1	0.339
keas	sloths	0	-1	-1	0	0	0	-2.315	0	0	0	0.301
keas	wombats	0	-1	-1	-1	0	0	-1.707	1	0	0	0.289
keas	lions	0	-1	-1	-1	-1	0	-6.166	0	1	1	0.217
aardvarks	keas	0	1	1	1	0	0	0.335	-1	0	0	0.695
bonobos	wombats	0	0	0	1	0	0	0.240	-1	-1	-1	0.496
aardvarks	bonobos	0	0	0	0	0	0	-1.611	1	1	1	0.551
ocarinas	vibraphones	0	0	1	0	0	0	0.461	-1	-1	-1	0.540
cymbals	ocarinas	0	0	1	1	0	0	3.485	2	0	0	0.786
clarinets	ocarinas	0	0	1	1	0	0	2.245	0	1	1	0.742
cellos	ocarinas	0	0	1	0	0	0	2.344	2	0	0	0.720

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
didgeridoos	vibraphones	0	0	0	0	0	0	0.533	-1	0	0	0.479
lutes	marimbas	0	0	0	0	0	0	1.225	2	1	1	0.645
kalimbas	lutes	0	0	0	0	0	0	-2.407	-2	-1	-1	0.342
clarinets	kalimbas	0	0	1	1	0	0	2.827	0	1	1	0.752
kalimbas	trumpets	0	0	-1	-1	0	0	-4.437	-1	0	0	0.234
cellos	kalimbas	0	0	1	1	0	0	2.926	1	0	0	0.751
kalimbas	saxophones	0	0	-1	-1	0	0	-3.124	0	-1	-1	0.252
lutes	saxophones	0	0	-1	-1	0	0	-0.717	2	0	0	0.398
casserole	eagle	0	-1	0	0	-1	0	-1.526	-1	1	1	0.410
kite	linguist	0	-1	1	0	0	0	1.602	1	1	1	0.656
algorithm	perfume	0	-1	-1	0	0	0	1.992	-2	-1	-1	0.280
forest	screwdriver	0	0	0	0	0	0	3.294	1	0	0	0.612
slipper	volcano	0	0	1	-1	-1	0	-1.806	1	0	0	0.540
harmonica	microscope	0	0	0	0	0	0	-1.745	-1	-2	-1	0.437
cookbook	zenith	0	1	1	0	0	0	1.136	0	1	1	0.722
hammock	hydrogen	0	1	-1	0	0	0	-2.060	1	1	0	0.410
neuron	toaster	0	-1	-1	0	0	0	0.383	0	1	0	0.309
marshmallow	telescope	0	0	0	0	0	0	-1.166	0	-1	-1	0.439
casserole	optics	0	1	0	0	0	0	-0.665	-1	1	1	0.557
encyclopedia	comet	0	1	0	-1	0	0	0.947	-4	-1	0	0.420
nimbus	waffle	0	-1	-1	0	0	0	-1.680	0	0	0	0.312
photon	pumpkin	0	-1	-1	0	0	0	-0.793	0	1	1	0.367
lantern	syntax	0	1	1	0	0	0	-0.941	0	-1	-1	0.609
echo	vineyard	0	-1	0	0	0	0	1.884	0	0	0	0.484
nebula	snowman	0	-1	-1	0	0	0	0.182	-1	-2	-1	0.320
botany	teapot	0	-1	-1	0	0	0	0.438	-1	-2	-1	0.325
chisel	kaleidoscope	0	0	0	1	0	0	0.139	2	-1	-1	0.606
lava	teacup	0	-1	-1	1	0	0	1.969	0	-1	-1	0.405
entropy	orchard	0	-1	-1	0	0	0	0.370	-1	-1	0	0.361
axolotl	vineyard	0	1	0	0	0	0	-3.510	-2	0	0	0.417

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
clockwork	meadow	0	-1	0	0	0	0	-0.987	0	1	1	0.464
algebra	telescope	0	-1	0	0	0	0	8.751	0	-2	-1	0.634
arcade	topaz	0	0	0	0	0	0	2.276	0	0	0	0.554
asteroid	compass	0	-1	0	1	0	0	-0.862	-1	0	0	0.452
bicycle	nebula	0	1	1	0	0	0	1.991	0	0	0	0.703
bungalow	entropy	0	1	0	0	0	0	-1.198	0	2	1	0.535
carnation	gnome	0	0	0	0	0	0	-1.769	-2	-1	-1	0.354
cinnamon	harmonica	0	0	1	0	0	0	2.300	1	0	0	0.688
coral	syntax	0	1	1	0	0	0	-0.020	0	-1	-1	0.627
dandelion	pendulum	0	0	0	0	0	0	-0.533	-1	1	0	0.409
delirium	telescope	0	-1	-1	0	1	0	-1.442	-1	-2	-1	0.294
anchors	sandstorms	0	1	0	0	-1	0	3.416	0	-1	-1	0.593
scissors	volcanoes	0	0	1	-1	0	0	0.577	1	0	0	0.594
equations	lanterns	0	-1	0	0	0	0	2.297	-1	0	0	0.454
satellites	tulips	0	-1	0	0	0	0	1.493	-1	1	1	0.479
compasses	hedgehogs	0	-1	0	0	0	0	-0.612	-1	-2	-1	0.400
comets	neckties	0	-1	0	1	0	0	2.294	0	-1	-1	0.516
castles	headphones	0	0	-1	1	0	0	-1.098	0	-1	-1	0.402
paperclips	pyramids	0	0	1	-1	0	0	-2.884	0	0	0	0.483
constellations	kettles	0	-1	0	0	0	0	0.938	-2	0	0	0.389
kaleidoscopes	whales	0	-1	-1	-1	0	0	-4.654	-3	0	0	0.147
meadows	pianos	0	-1	-1	0	0	0	1.147	1	0	0	0.403
magnets	zebras	0	-1	1	0	0	0	1.821	0	0	0	0.586
parrots	submarines	0	1	0	-1	0	0	-0.346	1	-1	-1	0.492
crayons	jungles	0	1	1	0	0	0	0.286	0	0	0	0.671
harbor	teapot	0	-1	0	0	0	0	2.561	0	-1	-1	0.457
notebook	quicksand	0	0	1	-1	0	0	3.309	0	0	0	0.614
glacier	lantern	0	-1	0	0	0	0	0.285	0	0	0	0.450
microscope	puddle	0	0	0	0	0	0	1.361	-1	1	1	0.538
compass	swan	0	-1	0	0	0	0	0.199	-1	-1	-1	0.371
bonsai	cathedral	0	1	0	-1	0	0	-2.004	1	1	1	0.540

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
honeycomb	violin	0	0	-1	0	0	0	-1.398	0	0	0	0.374
sailboat	stadium	0	0	0	0	0	0	-3.213	1	2	1	0.468
acorns	skyscrapers	0	1	0	0	0	0	-0.399	1	1	1	0.635
bell	trellis	0	0	1	0	0	0	3.327	1	1	1	0.740
inkwell	kite	0	0	-1	0	0	0	-3.215	-1	0	0	0.305
foxglove	trombone	0	0	-1	0	0	0	-1.888	0	1	1	0.403
carousel	quill	0	0	1	0	0	0	0.942	-2	0	0	0.554
lighthouse	onion	0	-1	-1	0	0	0	-1.153	0	1	1	0.359
cactus	chessboard	0	0	0	0	1	0	2.084	0	-1	-1	0.513
gallery	raindrop	0	0	0	0	0	0	5.058	-1	-2	-1	0.582
cricket	plow	0	1	0	-1	0	0	2.370	-1	-1	-1	0.474
gingerbread	fresco	0	0	1	0	0	0	0.345	-1	1	1	0.619
cello	sunflower	0	0	0	0	0	0	-0.417	1	0	0	0.535
archway	quilt	0	0	0	0	0	0	-2.775	-1	0	0	0.409
compass	haystack	0	0	0	0	0	0	2.338	0	-1	-1	0.514
beacon	millipede	0	-1	0	0	1	0	4.030	1	-1	-1	0.531
parchment	windmill	0	0	0	0	0	0	0.990	0	-1	-1	0.485
candlestick	meadow	0	1	0	0	0	0	-1.463	-1	1	1	0.540

Table A.1: Full list of binomials as well as their constraints.

B Individual Constraints at Each Checkpoint

Below is a table of the fixed-effects for each individual constraint at each checkpoint.

Parameter	num_tokens	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0B	0.214	0.160	-0.102	0.526
Percept	0B	-0.099	0.237	-0.564	0.365
Culture	0B	0.212	0.259	-0.292	0.725
Power	0B	0.180	0.273	-0.357	0.713
Intense	0B	0.621	0.415	-0.178	1.440
Icon	0B	-0.013	1.576	-3.169	3.035
Freq	0B	-0.080	0.084	-0.246	0.083
Len	0B	0.035	0.144	-0.250	0.317
Lapse	0B	-0.095	0.293	-0.671	0.480
no_final_stress	0B	0.101	0.365	-0.609	0.821
Intercept	2B	-0.019	0.262	-0.527	0.498
Percept	2B	0.466	0.371	-0.256	1.203
Culture	2B	-0.530	0.407	-1.337	0.253
Power	2B	0.418	0.418	-0.392	1.238
Intense	2B	0.036	0.597	-1.151	1.198
Icon	2B	0.007	1.739	-3.224	3.284
Freq	2B	0.266	0.139	-0.008	0.536
Len	2B	0.427	0.233	-0.031	0.882
Lapse	2B	0.526	0.439	-0.324	1.405
no_final_stress	2B	-0.715	0.552	-1.832	0.346
Intercept	41B	0.165	0.232	-0.292	0.620
Percept	41B	-0.202	0.327	-0.850	0.438
Culture	41B	-0.277	0.363	-0.996	0.446
Power	41B	1.027	0.398	0.260	1.818
Intense	41B	1.258	0.603	0.125	2.473
Icon	41B	0.036	1.695	-3.109	3.226
Freq	41B	-0.027	0.123	-0.267	0.216
Len	41B	0.733	0.209	0.322	1.144
Lapse	41B	0.005	0.387	-0.757	0.765
no_final_stress	41B	0.252	0.483	-0.694	1.218
Intercept	209B	0.191	0.190	-0.180	0.564
Percept	209B	0.223	0.270	-0.308	0.757
Culture	209B	0.214	0.297	-0.366	0.794
Power	209B	0.775	0.323	0.143	1.410
Intense	209B	-0.155	0.458	-1.061	0.754
Icon	209B	0.007	1.644	-3.022	3.081
Freq	209B	-0.057	0.099	-0.251	0.137
Len	209B	-0.061	0.170	-0.395	0.273
Lapse	209B	0.086	0.331	-0.561	0.736
no_final_stress	209B	-0.180	0.414	-1.000	0.635
Intercept	419B	-0.383	0.185	-0.747	-0.023

Parameter	num_tokens	Estimate	Est.Error	Q2.5	Q97.5
Percept	419B	0.242	0.268	-0.280	0.771
Culture	419B	-0.269	0.302	-0.857	0.316
Power	419B	0.513	0.315	-0.101	1.132
Intense	419B	0.220	0.455	-0.667	1.121
Icon	419B	-0.013	1.944	-3.262	3.270
Freq	419B	0.284	0.099	0.090	0.478
Len	419B	-0.371	0.166	-0.697	-0.045
Lapse	419B	-0.197	0.336	-0.877	0.459
no_final_stress	419B	0.767	0.421	-0.057	1.619
Intercept	838B	-0.029	0.191	-0.406	0.349
Percept	838B	0.196	0.277	-0.345	0.742
Culture	838B	-0.158	0.303	-0.741	0.446
Power	838B	0.695	0.322	0.064	1.327
Intense	838B	0.364	0.473	-0.547	1.302
Icon	838B	-0.002	1.641	-3.124	3.178
Freq	838B	0.141	0.100	-0.057	0.336
Len	838B	0.161	0.169	-0.168	0.498
Lapse	838B	0.418	0.336	-0.234	1.082
no_final_stress	838B	-0.426	0.421	-1.263	0.389
Intercept	1677B	-0.084	0.177	-0.433	0.261
Percept	1677B	0.488	0.260	-0.018	0.998
Culture	1677B	0.619	0.281	0.061	1.170
Power	1677B	0.708	0.298	0.130	1.291
Intense	1677B	-0.372	0.437	-1.237	0.480
Icon	1677B	-0.006	1.647	-3.031	3.047
Freq	1677B	0.081	0.093	-0.101	0.265
Len	1677B	-0.445	0.158	-0.758	-0.134
Lapse	1677B	-0.227	0.319	-0.854	0.393
no_final_stress	1677B	0.568	0.400	-0.212	1.366

Table B.1: Model results examining the effect of each individual constraint on LogOdds(AandB).