Emergent Abstract Word Order Preferences in Large Language Models

Anonymous ACL submission

Abstract

The rise in usage of large language models has brought with it many questions regarding how human-like their abilities are. One question is to what extent they're actually learning anything aside from surface level probabilities of their training data. In the present study we examine their abilities to learn abstract ordering preferences of binomials in English (e.g., *cats and dogs*). In Experiment 1, we demonstrate that large language models learn and use abstract preferences when producing novel binomials. In Experiment 2, we compare LLM preferences to human preferences. Finally, in Experiment 3, we demonstrate a timescale for this knowledge in LLMs across training.

1 Introduction

Large language models have stormed the media in the last few years. Their rise to fame has brought with them heated debates regarding whether large language models constitute human-like models of language learning and processing (Piantadosi, 2023; Piantadosi and Hill, 2022; Bender et al., 2021; Bender and Koller, 2020).

Many of these debates have centered around the tradeoff between computation and storage: how much are these models 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 is even being sued by *The New York Times* for allegedly reproducing entire articles verbatim.[1] This sentiment – that large language models are nothing but glorified copy cats – has been echoed by several other prominent linguists (Bender and Koller, 2020; Bender et al., 2021, c.f., Piantadosi and Hill, 2022).

Specifically, proponents of the "LLMs as copy cats" argument have pointed out that large language models are trained on magnitudes more data than humans.[2] As such, it is difficult to determine how much of the text generated by an LLM is truly novel, and how much is simply reproduced from its 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 difficult to work with. On the other hand, it is clear that large language models are learning at least some linguistic patterns. For example, McCoy et al. (2023) demonstrated that GPT-2 is able to generate well-formed novel words as well as well-formed novel syntactic structures, despite copying extensively.

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; Lasri et al., 2022; Pan and Bergen, 2025; Houghton et al., 2025). Thus in the present paper we address these debates by taking an in-depth look at large language models' abilities to learn abstract knowledge. Specifically we examine the ordering of novel binomials in English (Noun *and* Noun constructions, e.g., *cats and dogs*). The nouns in binomials can be ordered without affecting the meaning of the phrase much (e.g., *cats and dogs* vs *dogs and cats*). Despite this, humans have varying preferences for the ordering of the nouns in binomials. For example, humans prefer *bread and butter* much more than *butter and bread*. On the other hand, humans tend not to have a strong preference for *computers and monitors* over *monitors and computers* (or vice versa). Binomials are a useful test case because there is evidence demon-

[1] https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[2] Modern models are trained on trillions of tokens (Groeneveld et al., 2024), magnitudes larger than the 350 million words that the average college-aged speaker has seen in their lifetime (Levy et al., 2012).

strating that human ordering preferences are driven by abstract preferences, such as a preference for short words first (Morgan and Levy, 2016). Thus in the present study we examine ordering preferences for binomials that the model has never seen before. Through examining novel binomials, we gain insight into whether large language models learn abstract preferences similar to humans.

Our specific contributions are as follows: We make a unigram, bigram, and trigram corpus of Dolma (Soldaini et al., 2024), which is the data that was used to train OLMo-7B (Groeneveld et al., 2024). We also make the scripts to reproduce it open-access. We then use this corpus to find novel binomials that OLMo-7B (Groeneveld et al., 2024) has never seen. In Experiment 1, we examine whether OLMo-7B learns abstract preferences for binomials that it has never seen before. In Experiment 2 we examine human ordering preferences for these same binomials. Finally, in Experiment 3, we demonstrate a timescale of these preferences emerging over training.

2 Background and Related Work

2.1 Abstractions in Large Language Models

The evidence for learned abstractions in large language models is extremely mixed (Haley, 2020; Li and Wisniewski, 2021; Lasri et al., 2022; Li et al., 2023). For example, Haley (2020) demonstrated that many of the BERT models are not able to reliably determine the plurality of novel words. Additionally, Li and Wisniewski (2021) demonstrated that when tasked with producing the correct tense for a word, BERT tends to rely on memorization from its training data as opposed to learning the more general linguistic pattern. Similarly, Houghton et al. (2025) demonstrated that LLMs' ordering preferences for binomials they've experienced before are driven exclusively by their experience as opposed to abstract preferences.

On the other hand, Lasri et al. (2022) demonstrated that BERT can generalize well to novel subject-verb pairs. They tested BERT's performance on novel sentences along with semantically incoherent but syntactically sensible sentences (e.g., *colorless green ideas sleep furiously*) and found that when masking the verb for these sentences BERT still applies higher probability to the correct inflection of the verb. Additionally, Li et al. (2023) demonstrated that BERT is able to use abstract knowledge to correctly predict subject-verb and object-past participle agreements in French.

Research using other language models has also yielded similar results. For example, as mentioned earlier, McCoy et al. (2023) found that while GPT-2 copies extensively, it also produces both novel words as well as novel syntactic structures. Additionally, Misra and Mahowald (2024) examined whether language models trained on an amount of data comparable to humans are able to learn the Article Adjective Numeral Noun construction (AANN, e.g., *a beautiful five days*). They found that language models are able to learn this construction, even without experiencing them in their training, by generalizing across similar constructions such as *a few days*.

Finally, there's evidence that inducing abstractions in LLMs enhances large language models' reasoning abilities (Zheng et al., 2023) and helps them learn language more easily (McCoy et al., 2020). For example, Zheng et al. (2023) used a novel prompting technique to enable LLMs to use abstractions when reasoning. They found that LLMs hallucinate less when they implement abstractions in their reasoning. Similarly, McCoy et al. (2020) demonstrated that large language models can use abstractions, such as an abstract preference for certain syllable structures, to learn language more easily. Their results suggest that inducing abstractions may help reduce the amount of training that large language models require.

2.2 Abstractions in Humans

Most theories of language learning posit that humans leverage a 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 leverage similarities across different contexts to learn a word's general meaning (Yu and Smith, 2007).

There have been several studies showing human ordering preferences for binomials are driven, at least in part, by abstract ordering preferences (Morgan and Levy, 2016, 2015, 2024). For example, Morgan and Levy (2016) demonstrated that humans show ordering preferences for binomials beyond simply preferring the more frequent ordering. In order to demonstrate this, they coded a list of binomials for a variety of semantic, phonological, and metric constraints that affect human ordering preferences for binomials (Benor and Levy, 2006).

They found that human 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 of the binomial (the number of times that the binomial occurs in each ordering).

2.3 Present Study

In the present study we examine whether large language models are simply copying their input, or whether they are behaving more similarly to humans and learning abstract linguistics patterns. We use binomials as a test case because human ordering preferences deviate from the observed preferences (the frequency each order appears in corpus data) for them. While binomials are a single linguistic construction, they are well-studied in the linguistics literature and thus we have strong human baselines that we can compare the LLM's performance to. Additionally, we use novel binomials that OLMo has never seen before. Therefore any preference the model has cannot be driven by experience with the specific item.

In Experiment 1 we examine whether OLMo-7B (Groeneveld et al., 2024) is sensitive to abstract ordering preferences for novel binomials that the model has never seen before. We also examine the individual constraints that drive abstract ordering preferences in humans, such as the preference for short words before long words, to determine whether OLMo is sensitive to the same constraints in the same way as humans. In Experiment 2, we validate the results by eliciting human judgments for the same materials. Finally, in Experiment 3, we examine the same questions at different stages of the model's training in order to determine how these abstract ordering preferences emerge as a function of the training.

3 Datasets

3.1 Dolma

For all three 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 1-grams, 2-grams, and 3-grams corpus of Dolma (an open-source dataset released by the Allen Institute for AI under the Apache License 2.0; Soldaini et al., 2024). Specifically, we used Dolma version 1_7 (2.05 trillion tokens), which was used to train OLMo-7B-v1.7 (similarly released by the Allen Institute for AI under the Apache License 2.0; Groeneveld et al., 2024). Our corpus contains every n-gram (ignoring punctuation and capitalization) in the Dolma corpus, as well as the number of times that n-gram appeared.

We then created a list of binomials and searched the corpus to find those that did not occur in the Dolma corpus. We eliminated binomials which occurred more than zero times in either possible ordering. Thus, OLMo has had no experience with either ordering of any of the binomials in this corpus. We also calculated their unigram and bigram frequencies.[3] Our full list of items comprises 131 binomials and is reproduced in the appendix section (Section F).

3.2 Abstract Ordering Preferences Model

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 and Levy, 2015). Morgan and Levy (2015) 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 each ordering. Abstract ordering preferences are calculated from a mix of semantic, lexical, and phonological properties that human binomial ordering preferences have been shown to be sensitive to (Benor and Levy, 2006). For each of these constraints, a positive value indicates a preference for the alphabetically first word to be placed first (a neutral reference order). A negative value indicates a preference for the nonalphabetical word to be placed first. For example, a positive value of frequency indicates that the alphabetical word is more frequent. The constraints along with the estimated weights in humans (taken from Morgan and Levy, 2015) are as follows: length, no final stress, lapse, frequency, formal markedness, perceptual markedness, power, iconic/scalar sequencing, cultural centrality, and intensity. A detailed explanation of each of these constraints is included in the appendix section (Section B). These constraints are combined into a single abstract preference constraint using a logistic model (following Morgan and Levy, 2015)

[3] Our code and datasets for the analyses run in this paper can also be found at: `github.com/[anonymous]`.

4 Experiment 1: Novel Binomial Preferences

In Experiment 1, we examine whether OLMo-7B's ordering preferences are driven by abstract ordering preferences for novel binomials. In order to do so, we created a list of binomials and searched the Dolma corpus we created to confirm that they did not occur in either alphabetical or nonalphabetical ordering. Two of the authors then coded the binomials for each of the constraints mentioned earlier and disagreements were resolved by discussion. We then examined whether OLMo-7B shows any preference for one ordering over the other for each binomial. If OLMo has developed any abstract ordering preferences, it should show a systematic preference for one ordering over the other. If it has not, then it should show no preference for one ordering over the other.

4.1 Methods

4.1.1 Language Model Predictions

We calculated the ordering preferences of the alphabetical form (a neutral reference point) for each of our 131 binomials 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: ". In addition to this prefix, we also replicated the analyses with three other prefixes in order to assess how much variability there is across different prefixes. These prefixes were "example: ", "instance: ", and "try this: ". Thus the probability of the alphabetical form, *A and B*, is:

$$\begin{aligned} P_{alphabetical} &= P(\text{A}|\text{Prefix}) \\ &\times P(\text{and}|\text{Prefix: A}) \\ &\times P(\text{B}|\text{Prefix: A and}) \end{aligned} \quad (1)$$

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:

$$\begin{aligned} P_{nonalphabetical} &= P(\text{B}|\text{Prefix:}) \\ &\times P(\text{and}|\text{Prefix: B}) \\ &\times P(\text{A}|\text{Prefix: B and}) \end{aligned} \quad (2)$$

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:

$$LogOdds(\text{AandB}) = log(\frac{p_{alphabetical}}{p_{nonalphabetical}}) \quad (3)$$

4.1.2 Analyses

We present three Bayesian linear mixed-effects 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 of the effect from the grand mean.

For all three analyses, the dependent variable is LogOdds(AandB). Further, for the analyses presented here, we used the "Next item: " prefix. Our independent variable in the first analysis is the abstract ordering preference for each binomial (AbsPref). In order to rule out bigram probabilities as driving the model's preferences, our second analysis contains the binomial's bigram probabilities (the odds ratio of the product of bigram probabilities of the alphabetical order to the product of bigram probability of the nonalphabetical order) as well. Finally, our independent variables in the third analysis are the individual constraints used to calculate the abstract preferences. The model equations are below in Equations 4, 5, and 6. Note that Formal Markedness and Iconicity were dropped from the third model because the constraint values were zero for all of the binomials. Further, our constraints demonstrated a level of co-linearity which can result in poor model estimates and inflated credible intervals. In order to address this, we dropped the constraint with the highest variance inflation factor (which turned out to be the lapse constraint). All other constraints had a variance inflation factor value below 1.5. We then performed backward model selection and dropped the predictors whose credible intervals were most centered around zero until the remaining predictors' credible intervals were all at least 75% greater than or less than zero. This resulted in dropping the no final stress, intense, and percept constraints. We acknowledge that this approach is quite exploratory and thus interpretations at the level of the individual constraints must be taken with a grain of salt.

$$\text{LogOdds(AandB)} \sim \text{AbsPref} \quad (4)$$

$$\text{LogOdds(AandB)} \sim \text{AbsPref} \cdot \text{BigramProbs} \quad (5)$$

$$\text{LogOdds(AandB)} \sim \text{Culture} + \text{Power} + \text{Freq} + \text{Len} \quad (6)$$

4.2 Results

The results for the first analysis are presented in Table 1. Our results suggest that there is a main-effect of abstract ordering preference for OLMo-7B.

	Estimate	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).

The results of our second model suggest that the model's ordering preferences were not driven by bigram probabilities (see Table 2).

	Est.	Err.	Q2.5	Q97.5
Intercept	-1.26	0.64	-2.59	-0.11
AbsPref	2.45	1.28	0.20	5.11
Bigrams	0.41	0.49	-0.56	1.41
AbsPref:Bigrams	0.16	0.96	-1.77	2.11

Table 2: Model results examining the effect of AbsPref and bigram probabilities on LogOdds(AandB).

While these results suggest that the large language models' ordering preferences are sensitive abstract ordering preferences and not bigram probabilities, it's unclear whether their behavior is similar to humans on the level of the individual constraints. Thus, in the third analysis we examined 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 in Table 3.

The model is most sensitive to the Power constraint, however there appears to be a marginal effect of Culture as well, since nearly 95% of the posterior samples are greater than zero despite the credible interval crossing zero. 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 typically see in humans. We ran a follow-up analysis with just length and no other constraints to confirm this was not a result of co-linearity, however the effect remained negative.

	Est.	Err.	Q2.5	Q97.5	% > 0
Intercept	-0.13	0.16	-0.45	0.19	20.75
Culture	0.41	0.25	-0.08	0.92	94.95
Power	0.72	0.26	0.20	1.24	99.67
Freq	0.09	0.09	-0.08	0.26	85.17
Len	-0.21	0.13	-0.48	0.05	5.87

Table 3: Model results examining the effect of each individual constraint on LogOdds(AandB).

We replicated each of these analyses with three different prefixes as well. Specifically, we used the prefixes "example: ", "instance: ", and "try this: ". The results for these analyses are included in the appendix (Section D), however we describe the results briefly here. For every replication there was a meaningful effect of abstract preference. However, there was some variability with respect to the sensitivity to the individual constraints. While the effects of Frequency, Culture, and Length were relatively robust (with statistically meaningful effects across 3 of the 4 prefixes), the effect of power was less robust and only statistically meaningful with 2 of the four prefixes. This suggests that the effect of the prefix does influence the ordering preferences, however we leave a more thorough examination of different prefixes to future work.

4.3 Discussion

The results in Experiment 1 demonstrate that OLMo-7B has learned abstract ordering preferences for novel binomials that it has never seen before. Further, these ordering preferences aren't simply based on the individual word frequencies or bigram frequencies. 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.

Previous results have suggested that LLMs do not use abstract preferences when ordering binomials that they've experienced before, even if they are low-frequency (Houghton et al., 2025). The results of Experiment 1 thus suggest that LLMs are learning abstract preferences, but unlike humans

they are not using them for binomials they've experienced before.

In Experiment 2, we examine human ordering preferences for each of these binomials.

5 Experiment 2: Human Judgments

In Experiment 1 we found that large language models are learning abstract ordering preferences for binomials. In Experiment 2 we elicit human judgments for the same set of binomials in order to compare the results from Experiment 1 with them.

5.1 Methods

5.1.1 Procedure

In order to elicit human judgments, we used a 2 alternative forced choice task. Participants were given both ordering of the binomials (in a random order) and asked to choose the order the they thought was most natural. The order of the items as well as the order of the choices were randomized.

100 undergraduate students were recruited from a research university in the United States and given course credit in exchange for their participation.

5.1.2 Analyses

We ran three Bayesian logistic regression analyses, one examining the effects of abstract preferences, one examining the effects of each of the individual constraints, and one examining whether the LLM ordering preferences predict the human preferences. For all three analyses, the dependent variable was whether the participant chose the alphabetical order of the binomial or not.

For the first analysis, the independent variable was the abstract preference value. A random slope was included for abstract preference by participant, and random intercepts were included for both participant and item. For the second analysis, the independent variables were each of the individual constraints (culture, power, frequency, and length). Random slopes were included for each of the individual constraints by participant, and random intercepts were included for both participant and item. For the third model, we included a main-effect of $LogOdds(\text{AandB})$ with a random slope for $LogOdds(\text{AandB})$ by participant, and random intercepts for participant and item. The model equations are included below in Equations 7, 8, and 9.

$$\begin{aligned}\text{HumanPref} \sim\ & \text{AbsPref} \\ & + (\text{AbsPref}|\text{Participant}) \\ & + (1|\text{Item}) \qquad (7)\end{aligned}$$

$$\begin{aligned}\text{HumanPref} \sim\ & \text{Culture} + \text{Power} \\ & + \text{Freq} + \text{Len} \qquad (8)\end{aligned}$$

$$\begin{aligned}\text{HumanPref} \sim\ & \text{LogOdds}(\text{AandB}) \\ & + (\text{LogOdds}(\text{AandB})|\text{Participant}) \\ & + (1|\text{Item}) \qquad (9)\end{aligned}$$

6 Results

The results are presented in Tables 4 through 6.

Overall we find a meaningful effect of abstract preference. Further, we find that humans are sensitive to the culture, frequency, and length constraints, but not the power constraint. In other words, humans prefer the more culturally central word first, the more frequent word first, and the shorter word first, but have no preference with respect to whether the more powerful word is placed first or not.[4] Finally, we find that despite learning some preferences, the LLM's preferences are not a good predictor of the human data (Figure 1).

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.10	0.04	0.02	0.17
AbsPref	0.88	0.21	0.48	1.29

Table 4: Model results for the effects of abstract preferences on participants' ordering preferences. The results demonstrate that human ordering preferences are sensitive to abstract preferences.

	Est.	Err.	Q2.5	Q97.5	% > 0
Intercept	0.09	0.04	0.01	0.18	98.51
Culture	0.06	0.04	-0.02	0.14	94.12
Power	0.00	0.04	-0.07	0.07	53.75
Freq	0.04	0.01	0.02	0.06	99.94
Len	0.08	0.03	0.02	0.14	99.75

Table 5: Model results for the effects of the individual constraints on participants' ordering preferences. The results demonstrate that human ordering preferences are sensitive to culture, frequency, and length, but not power.

[4] Note that this is in contrast to the results we see for most binomials, where previous studies suggest that humans are in fact sensitive to the power constraint (Morgan and Levy, 2016).

	Est.	Err.	Q2.5	Q97.5
Intercept	0.08	0.04	0.00	0.16
LogOdds	0.00	0.01	-0.03	0.02

Table 6: Results for the model examining whether the LLM ordering preferences predict the human data.

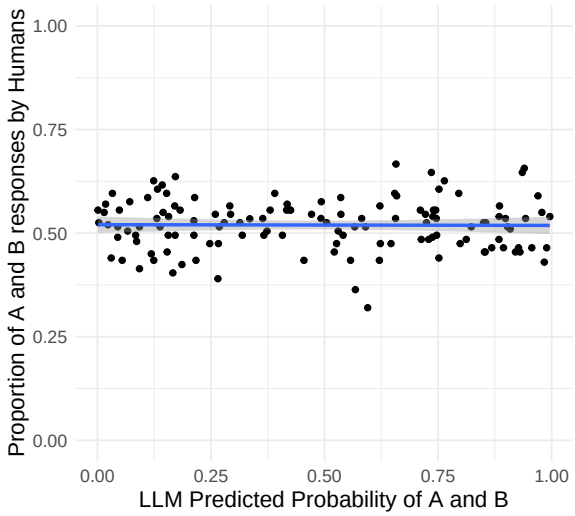

Figure 1: Human results plotted against model predictions. Probabilities for humans are collapsed across participants for each item. The results suggest no correlation between the human predictions and the LLM predictions.

7 Discussion

The results in Experiment 2 suggest that while large language models learn abstract preferences, they are not identical to those that humans apply in the case of these novel binomials. The most obvious discrepancy is in the length constraint: While OLMo showed a preference to place the longer word first, humans prefer to place the shorter word first. It is possible that this is due to a difference in how the input is represented in LLMs and humans. The length constraint is determined by the number of syllables. While syllables are salient cues in the audio that humans receive during learning, it's less clear how salient of a cue they are for LLMs.

For all other constraints, however, OLMo learns a constraint that operates in the same direction as humans. Nonetheless, OLMo and humans seem to weight these constraints quite differently for these novel binomials, leading to the lack of correlation between human and LLM preferences on these items. To further understand why these constraint weightings are so different between OLMo and humans, we can also compare the constraint weightings we see here (Table 3 for OLMo and Table 5 for humans) to the constraint weights reported by Morgan and Levy (2015, Appendix A). The Morgan and Levy (2015) constraint weightings represent human constraint weightings trained on a larger and more representative set of binomials, compared to the novel binomials we study here that are infrequent combinations of infrequent words. The human constraint weightings that we discover here are extremely small compared to the Morgan and Levy (2015) weightings (despite being on the same scale), indicating that even though these predictors are statistically significant, they are producing small effects in the human data, likely due to humans not having strong intuitions about these odd binomials. For example, they might not have strong intuitions about whether an ocelot versus a narwhal is more powerful or more culturally central. In contrast, the OLMo constraint weights actually look somewhat more like the Morgan and Levy (2015) weights, particularly when averaged over the different prefixes (Appendix D). In other words, OLMo when tested on novel binomials looks more like humans when tested on a representative sample of binomials (as indicated by Morgan and Levy), compared with humans when tested on novel binomials. This may be because, again, humans may not have strong intuitions about these infrequent words, whereas OLMo has enough training data that even if the binomials themselves are novel, OLMo may have stronger knowledge of the individual words.

8 Experiment 3: Novel Binomial Preferences Throughout Training

In Experiment 3, we examine the timescale of the model's preferences over the model's training. Specifically, OLMo makes public various checkpoints during the model's training, allowing us the opportunity to examine how these preferences arise as a function of the training. Thus, in Experiment 3 we examine the evolution of these learned abstract ordering preferences as the model learns over time.

8.1 Methods

8.1.1 Language Model Predictions

The language model predictions in Experiment 3 were obtained using the same procedure as in Experiment 1. However, instead of calculating these only for the main model, we calculated them at various checkpoints. Specifically, the number of tokens that each model was trained on is as follows:

0B, 2B, 41B, 209B, 419B, 838B, and 1677B.

8.1.2 Analysis

The analyses in Experiment 3 were identical to those of Equations 4 and 6 from Experiment 1, however we ran these analyses for each of the checkpoints listed above.

8.2 Results

Our model estimates at each checkpoint are presented in Table 7.

# Tokens	Estimate	Error	Q2.5	Q97.5
0B	0.396	1.149	-1.912	2.606
2B	1.23	1.38	-0.96	4.57
41B	0.87	1.16	-1.07	3.48
209B	1.05	1.05	-0.77	3.38
419B	1.70	1.23	-0.31	4.42
838B	1.78	1.25	-0.32	4.51
1677B	3.87	1.52	0.96	6.87

Table 7: Model results examining the effect of AbsPref on LogOdds(AandB) for each checkpoint.

Our results demonstrate that it takes a large number of tokens for the model to learn abstract preferences. The effect of abstract ordering preference isn't convincing until the model has experienced 1677B tokens.

Similar to Experiment 1, we also present an analysis at the level of the individual constraints. The full results are in the appendix section (Section E), but we present a visualization in Figure 2.

Interestingly, it appears that early on the model already shows evidence of learning preferences. For example, by 10B tokens the model has learned to place more powerful words first and shorter words first. However, the model is slower to learn the culture constraint. Further, as it receives more training the effect of length reverses in direction.

8.3 Discussion

Our results demonstrate that OLMo learns preferences early on for the power, frequency, and length constraints, but is slower to learn 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 it may be due to tokenization.

It is interesting that the model is quick to learn the power constraint, but slower to learn the culture constraint. While both constraints require a level of world-knowledge, the model learns which entities are more powerful relatively more quickly.

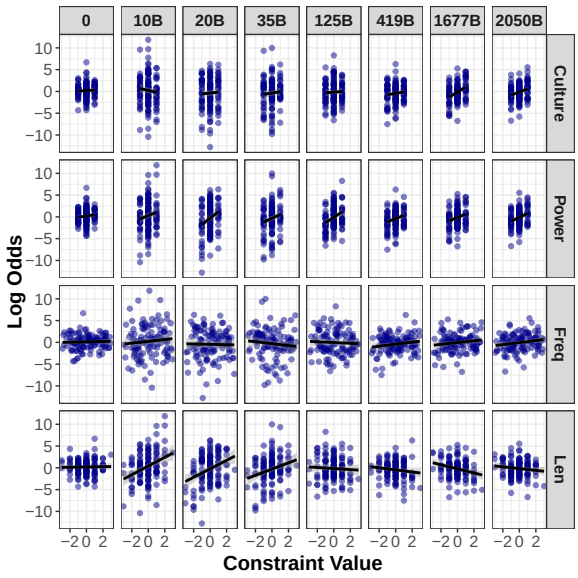

Figure 2: Visualization of the effect of each constraint on the ordering preference at each checkpoint.

9 Conclusion

In the present study, we examined the ordering preferences for binomials in OLMo-7B and in humans. We found that while OLMo does learn ordering preferences, these are not identical to humans. Mainly, OLMo prefers to place longer words first, which is the opposite from the human preferences. For the other constraints, OLMo learns similarly to humans but differs in magnitude.

Additionally, we find that while OLMo is quick to learn some constraints, it takes longer to learn the culture constraint. Moreover, the direction of the learned length constraint undergoes a reversal in direction as OLMo receives more training.

Overall, our results suggest that LLMs are not simply copying their input, but are learning human-like phenomena from their training. However, they are not learning identically to humans. Specifically, LLMs are more confident about the ordering preferences for some novel binomials and use the entirety of the probability scale. On the other hand, humans show much smaller preferences for one ordering over the other for novel binomials. Further, while humans rely on abstract preferences even for binomials that they have encountered before (Morgan and Levy, 2024), LLMs rely on abstract ordering preferences only for items that they have not encountered before (see Houghton et al., 2025).

10 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 outweigh a more broad analysis of several models, especially given the lack of easily available open access training data, which is crucial to guaranteeing that the model has not encountered our items before (Houghton et al., 2025).

Additionally, we only test one construction in this paper (binomials) and in only one language (English). While it is possible that abstract ordering preferences for binomials are different than other constructions, binomials are well understood in the human linguistics literature thus making them a good test case for our analyses.

Finally, we do see some variability as a function of the prefix, but we did not set out to systematically test this. Thus it is unclear how much exactly the prefix affects the results.

References

Ben Ambridge. 2020. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Sarah Bunin Benor and Roger Levy. 2006. The chicken or the egg? a probabilistic analysis of english binomials. *Language*, pages 233–278.

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

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

William E Cooper and John R Ross. 1975. World order. *Papers from the parasession on functionalism*, 11:63–111.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

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.

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*, pages 1–17. Publisher: Cambridge University Press.

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.

Karim Lasri, Olga Seminck, Alessandro Lenci, and Thierry Poibeau. 2022. Subject verb agreement error patterns in meaningless sentences: Humans vs. bert. *arXiv preprint arXiv:2209.10538*.

Benjamin LeBrun, Alessandro Sordoni, and Timothy J O'Donnell. 2022. Evaluating distributional distortion in neural language modeling. *arXiv preprint arXiv:2203.12788*.

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

Bingzhi Li and Guillaume Wisniewski. 2021. Are neural networks extracting linguistic properties or memorizing training data? an observation with a multilingual probe for predicting tense. In *EACL 2021*.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023. Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

R Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*.

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.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *arXiv preprint arXiv:2403.19827*.

Emily Morgan and Roger Levy. 2015. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *CogSci*. Citeseer.

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.

Dingyi Pan and Ben Bergen. 2025. Are explicit belief representations necessary? a comparison between large language models and Bayesian probabilistic models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11483–11498, Albuquerque, New Mexico. Association for Computational Linguistics.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Chen Yu and Linda B Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

A Correlation between each of the constraints in our binomial dataset

The pairwise correlations between each of the constraints in our dataset is included below.

	Percept	Culture	Power	Intense	Freq	Len	Lapse	*BStress
Percept	1.000	0.346	0.034	-0.286	0.102	-0.010	0.225	0.096
Culture	0.346	1.000	0.186	-0.001	0.408	-0.069	-0.080	-0.054
Power	0.034	0.186	1.000	0.356	0.311	0.090	0.133	-0.026
Intense	-0.286	-0.001	0.356	1.000	0.134	0.122	-0.112	-0.182
Freq	0.102	0.408	0.311	0.134	1.000	0.127	-0.145	-0.183
Len	-0.010	-0.069	0.090	0.122	0.127	1.000	0.401	0.253
Lapse	0.225	-0.080	0.133	-0.112	-0.145	0.401	1.000	0.835
*BStress	0.096	-0.054	-0.026	-0.182	-0.183	0.253	0.835	1.000

Table 8: Pairwise correlations between each constraint.

B Explanation of Each of the Individual Constraints

An explanation of each of the individual constraints is provided below. These explanations are taken from Morgan and Levy (2015).

- **Length**: The shorter word should appear first, e.g. *abused and neglected*. In human data, the weight of this constraint was estimated to be 0.15.
- **No Final Stress**: The final syllable of the second word should not be stressed, e.g. *abused and neglected*. In human data, the weight of this constraint was estimated to be 0.36.
- **Lapse**: Avoid unstressed syllables in a row, e.g. *FARMS and HAY-fields* vs *HAY-fields and FARMS*. In human data, the weight of this constraint was estimated to be 0.19.
- **Frequency**: The more frequent word comes first, e.g. *bride and groom*. In human data, the weight of this constraint was estimated to be 0.09.
- **Formal Markedness**: The word with more general meaning or broader distribution comes first, e.g. *boards and two-by-fours*. In human data, the weight of this constraint was estimated to be 0.24.
- **Perceptual Markedness**: Elements that are more closely connected to the speaker come first. This constraint encompasses Cooper and Ross (1975)'s 'Me First' constraint and includes numerous subconstraints, e.g.: animates precede inanimates; concrete words precede abstract words; e.g. *deer and trees*. In human data, the weight of this constraint was estimated to be 0.25.
- **Power**: The more powerful or culturally prioritized word comes first, e.g. *clergymen and parishioners*. In human data, the weight of this constraint was estimated to be 0.26.
- **Iconic/scalar sequencing**: Elements that exist in sequence should be ordered in sequence, e.g. *achieved and maintained*. In human data, the weight of this constraint was estimated to be 1.30.
- **Cultural Centrality**: The more culturally central or common element should come first, e.g. *oranges and grapefruits*. In human data, the weight of this constraint was estimated to be 0.42.
- **Intensity**: The element with more intensity appear first, e.g. *war and peace*. In human data, the weight of this constraint was estimated to be 0.02.

C Constraint Weights from Morgan and Levy (2015)

The value of each of the individual constraints from Morgan and Levy (2015).

Constraint Name	Constraint Weight
Length	0.15
No Final Stress	0.36
Lapse	0.19
Frequency	0.09
Formal Markedness	0.24
Perceptual Markedness	0.25
Power	0.26
Iconic/scalar sequencing	1.30
Cultural Centrality	0.42
Intensity	0.02

Table 9: Constraint weights from Morgan and Levy (2015).

D Results for Different Prefixes

Below are the results for the effects of each of the individual constraints for different prefixes.

Prefix	Coefficient	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Next item:	Intercept	-0.132	0.163	-0.451	0.185	20.75
Next item:	Culture	0.414	0.254	-0.078	0.916	94.95
Next item:	Power	0.720	0.263	0.204	1.236	99.67
Next item:	Freq	0.091	0.087	-0.079	0.263	85.17
Next item:	Len	-0.209	0.134	-0.476	0.052	5.87
example:	Intercept	-0.364	0.315	-0.977	0.248	12.26
example:	Culture	0.840	0.483	-0.076	1.814	96.33
example:	Power	0.020	0.472	-0.909	0.938	51.78
example:	Freq	0.329	0.169	-0.004	0.661	97.38
example:	Len	-0.469	0.264	-0.990	0.040	3.55
instance:	Intercept	-0.388	0.405	-1.184	0.412	16.70
instance:	Culture	1.266	0.654	0.039	2.614	97.88
instance:	Power	0.101	0.590	-1.059	1.273	56.98
instance:	Freq	0.349	0.222	-0.085	0.787	94.41
instance:	Len	-0.597	0.344	-1.274	0.073	3.94
try this:	Intercept	-0.117	0.235	-0.572	0.346	30.80
try this:	Culture	1.423	0.378	0.680	2.162	100.00
try this:	Power	0.609	0.363	-0.097	1.329	95.41
try this:	Freq	-0.020	0.124	-0.264	0.226	43.66
try this:	Len	-0.205	0.190	-0.582	0.168	13.87

E Full Model Results

Table 10 presents the results for the effect of each of the different constraints at each checkpoint in the model.

Parameter	Number of Tokens	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0B	0.223	0.159	-0.087	0.539
Culture	0B	0.149	0.244	-0.327	0.623
Power	0B	0.286	0.249	-0.207	0.777
Freq	0B	-0.070	0.082	-0.232	0.091
Len	0B	0.030	0.127	-0.220	0.278
Intercept	2B	-0.027	0.256	-0.529	0.478

Culture	2B	-0.399	0.390	-1.161	0.361
Power	2B	0.531	0.404	-0.250	1.334
Freq	2B	0.258	0.135	-0.008	0.524
Len	2B	0.492	0.212	0.076	0.909
Intercept	41B	0.179	0.229	-0.268	0.628
Culture	41B	-0.377	0.347	-1.065	0.305
Power	41B	1.290	0.373	0.568	2.037
Freq	41B	-0.035	0.121	-0.274	0.202
Len	41B	0.807	0.188	0.438	1.179
Intercept	209B	0.176	0.182	-0.186	0.537
Culture	209B	0.290	0.283	-0.268	0.847
Power	209B	0.760	0.289	0.194	1.327
Freq	209B	-0.056	0.096	-0.244	0.132
Len	209B	-0.063	0.150	-0.358	0.234
Intercept	419B	-0.458	0.183	-0.816	-0.099
Culture	419B	-0.125	0.282	-0.679	0.437
Power	419B	0.508	0.289	-0.056	1.073
Freq	419B	0.240	0.096	0.053	0.431
Len	419B	-0.298	0.150	-0.591	-0.005
Intercept	838B	-0.022	0.184	-0.381	0.335
Culture	838B	-0.111	0.284	-0.661	0.446
Power	838B	0.865	0.297	0.283	1.456
Freq	838B	0.127	0.099	-0.068	0.319
Len	838B	0.247	0.154	-0.055	0.552
Intercept	1677B	-0.181	0.176	-0.527	0.159
Culture	1677B	0.861	0.273	0.326	1.394
Power	1677B	0.562	0.275	0.031	1.108
Freq	1677B	0.052	0.091	-0.125	0.230
Len	1677B	-0.431	0.142	-0.708	-0.156

Table 10: Model results examining the effect of each individual constraint on LogOdds(AandB).

F Full Stimuli List

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Stress	AbsPref
kiwis	wolverines	0.00	0.00	0.00	-1.00	-1.00	0.00	-0.29	1.00	-1.00	-1.00	0.43
kiwis	narwhals	0.00	1.00	0.00	-1.00	-1.00	0.00	2.71	0.00	-1.00	-1.00	0.52
kiwis	ocelots	0.00	0.00	0.00	-1.00	0.00	0.00	2.74	0.00	-1.00	-1.00	0.46
ibex	kiwis	0.00	0.00	0.00	1.00	0.00	0.00	-1.17	0.00	1.00	0.00	0.50
harpies	kiwis	0.00	-1.00	0.00	1.00	1.00	0.00	-1.95	0.00	0.00	0.00	0.47
axolotls	wolverines	0.00	-1.00	0.00	-1.00	0.00	0.00	-2.69	-1.00	-1.00	-1.00	0.26
axolotls	ibex	0.00	-1.00	0.00	-1.00	0.00	0.00	-1.23	-2.00	-1.00	0.00	0.33
axolotls	harpies	0.00	1.00	0.00	-1.00	-1.00	0.00	-0.45	-2.00	0.00	0.00	0.41
axolotls	keas	0.00	0.00	1.00	-1.00	0.00	0.00	1.05	-3.00	-1.00	1.00	0.59
axolotls	bonobos	0.00	-1.00	0.00	-1.00	0.00	0.00	-0.90	-1.00	0.00	0.00	0.33
axolotls	wombats	0.00	-1.00	0.00	-1.00	0.00	0.00	-0.66	-2.00	-1.00	-1.00	0.27
axolotls	lions	0.00	-1.00	-1.00	-1.00	-1.00	0.00	-5.11	-2.00	0.00	0.00	0.16
ocelots	platypuses	0.00	1.00	-1.00	1.00	0.00	0.00	0.66	1.00	3.00	1.00	0.52
ibex	platypuses	0.00	1.00	0.00	1.00	0.00	0.00	2.23	2.00	3.00	1.00	0.69
harpies	platypuses	0.00	0.00	0.00	1.00	1.00	0.00	1.45	2.00	2.00	0.00	0.58

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Stress	AbsPref
keas	platypuses	0.00	0.00	-1.00	0.00	0.00	0.00	-0.05	3.00	3.00	1.00	0.46
bonobos	platypuses	0.00	1.00	0.00	1.00	0.00	0.00	1.90	1.00	2.00	0.00	0.61
harpies	wolverines	0.00	-1.00	0.00	0.00	0.00	0.00	-2.24	1.00	-1.00	1.00	0.57
keas	wolverines	0.00	-1.00	-1.00	-1.00	0.00	0.00	-3.74	2.00	0.00	0.00	0.28
bonobos	wolverines	0.00	0.00	0.00	0.00	0.00	0.00	-1.79	0.00	-1.00	-1.00	0.43
capybaras	ibex	0.00	0.00	0.00	0.00	0.00	0.00	-1.66	-2.00	-1.00	-1.00	0.36
capybaras	harpies	0.00	1.00	0.00	-1.00	-1.00	0.00	-0.88	-2.00	0.00	0.00	0.40
capybaras	keas	0.00	1.00	1.00	1.00	0.00	0.00	0.62	-3.00	-1.00	-1.00	0.59
ibex	narwhals	0.00	1.00	0.00	-1.00	0.00	0.00	1.54	0.00	0.00	0.00	0.54
harpies	narwhals	0.00	-1.00	0.00	0.00	1.00	0.00	0.77	0.00	-1.00	-1.00	0.42
keas	narwhals	0.00	0.00	-1.00	-1.00	0.00	0.00	-0.74	1.00	0.00	0.00	0.36
ibex	ocelots	0.00	0.00	0.00	0.00	0.00	0.00	1.57	1.00	0.00	0.00	0.58
keas	ocelots	0.00	-1.00	-1.00	-1.00	0.00	0.00	-0.71	2.00	0.00	0.00	0.34
bonobos	ocelots	0.00	0.00	1.00	0.00	0.00	0.00	1.23	0.00	-1.00	-1.00	0.59
harpies	ibex	0.00	-1.00	0.00	0.00	1.00	0.00	-0.78	0.00	-1.00	-1.00	0.39
ibex	keas	0.00	1.00	1.00	1.00	0.00	0.00	2.28	-1.00	0.00	0.00	0.73
bonobos	ibex	0.00	0.00	0.00	0.00	0.00	0.00	-0.33	-1.00	-1.00	-1.00	0.42
ibex	koalas	0.00	0.00	-1.00	0.00	0.00	0.00	-0.43	1.00	1.00	1.00	0.47
ibex	sloths	0.00	0.00	-1.00	1.00	0.00	0.00	-0.04	-1.00	0.00	0.00	0.43
aardvarks	ibex	0.00	0.00	1.00	0.00	0.00	0.00	-1.95	0.00	0.00	0.00	0.57
harpies	keas	0.00	-1.00	1.00	1.00	1.00	0.00	1.50	-1.00	-1.00	-1.00	0.57
bonobos	harpies	0.00	1.00	0.00	-1.00	-1.00	0.00	0.44	-1.00	0.00	0.00	0.47
harpies	koalas	0.00	-1.00	-1.00	0.00	1.00	0.00	-1.21	1.00	0.00	0.00	0.36
harpies	wombats	0.00	-1.00	0.00	1.00	1.00	0.00	-0.20	0.00	-1.00	-1.00	0.47
aardvarks	harpies	0.00	1.00	0.00	0.00	-1.00	0.00	-1.17	0.00	1.00	1.00	0.58
bonobos	keas	0.00	1.00	1.00	1.00	0.00	0.00	1.95	-2.00	-1.00	-1.00	0.66
keas	koalas	0.00	-1.00	-1.00	-1.00	0.00	0.00	-2.71	2.00	1.00	1.00	0.34
keas	sloths	0.00	-1.00	-1.00	0.00	0.00	0.00	-2.31	0.00	0.00	0.00	0.30
keas	wombats	0.00	-1.00	-1.00	-1.00	0.00	0.00	-1.71	1.00	0.00	0.00	0.29
keas	lions	0.00	-1.00	-1.00	-1.00	-1.00	0.00	-6.17	0.00	1.00	1.00	0.22
aardvarks	keas	0.00	1.00	1.00	1.00	0.00	0.00	0.34	-1.00	0.00	0.00	0.69
bonobos	wombats	0.00	0.00	0.00	1.00	0.00	0.00	0.24	-1.00	-1.00	-1.00	0.50
aardvarks	bonobos	0.00	0.00	0.00	0.00	0.00	0.00	-1.61	1.00	1.00	1.00	0.55
ocarinas	vibraphones	0.00	0.00	1.00	0.00	0.00	0.00	0.46	-1.00	-1.00	-1.00	0.54
cymbals	ocarinas	0.00	0.00	1.00	1.00	0.00	0.00	3.48	2.00	0.00	0.00	0.79
clarinets	ocarinas	0.00	0.00	1.00	1.00	0.00	0.00	2.25	0.00	1.00	1.00	0.74
cellos	ocarinas	0.00	0.00	1.00	0.00	0.00	0.00	2.34	2.00	0.00	0.00	0.72
didgeridoos	vibraphones	0.00	0.00	0.00	0.00	0.00	0.00	0.53	-1.00	0.00	0.00	0.48
lutes	marimbas	0.00	0.00	0.00	0.00	0.00	0.00	1.23	2.00	1.00	1.00	0.64
kalimbas	lutes	0.00	0.00	0.00	0.00	0.00	0.00	-2.41	-2.00	-1.00	-1.00	0.34
clarinets	kalimbas	0.00	0.00	1.00	1.00	0.00	0.00	2.83	0.00	1.00	1.00	0.75
kalimbas	trumpets	0.00	0.00	-1.00	-1.00	0.00	0.00	-4.44	-1.00	0.00	0.00	0.23
cellos	kalimbas	0.00	0.00	1.00	1.00	0.00	0.00	2.93	1.00	0.00	0.00	0.75
kalimbas	saxophones	0.00	0.00	-1.00	-1.00	0.00	0.00	-3.12	0.00	-1.00	-1.00	0.25
lutes	saxophones	0.00	0.00	-1.00	-1.00	0.00	0.00	-0.72	2.00	0.00	0.00	0.40
casserole	eagle	0.00	-1.00	0.00	0.00	-1.00	0.00	-1.53	-1.00	1.00	1.00	0.41
kite	linguist	0.00	-1.00	1.00	0.00	0.00	0.00	1.60	1.00	1.00	1.00	0.66
algorithm	perfume	0.00	-1.00	-1.00	0.00	0.00	0.00	1.99	-2.00	-1.00	-1.00	0.28
forest	screwdriver	0.00	0.00	0.00	0.00	0.00	0.00	3.29	1.00	0.00	0.00	0.61
slipper	volcano	0.00	0.00	1.00	-1.00	-1.00	0.00	-1.81	1.00	0.00	0.00	0.54

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Stress	AbsPref
harmonica	microscope	0.00	0.00	0.00	0.00	0.00	0.00	-1.75	-1.00	-2.00	-1.00	0.44
cookbook	zenith	0.00	1.00	1.00	0.00	0.00	0.00	1.14	0.00	1.00	1.00	0.72
hammock	hydrogen	0.00	1.00	-1.00	0.00	0.00	0.00	-2.06	1.00	1.00	0.00	0.41
neuron	toaster	0.00	-1.00	-1.00	0.00	0.00	0.00	0.38	0.00	1.00	0.00	0.31
marshmallow	telescope	0.00	0.00	0.00	0.00	0.00	0.00	-1.17	0.00	-1.00	-1.00	0.44
casserole	optics	0.00	1.00	0.00	0.00	0.00	0.00	-0.66	-1.00	1.00	1.00	0.56
encyclopedia	comet	0.00	1.00	0.00	-1.00	0.00	0.00	0.95	-4.00	-1.00	0.00	0.42
nimbus	waffle	0.00	-1.00	-1.00	0.00	0.00	0.00	-1.68	0.00	0.00	0.00	0.31
photon	pumpkin	0.00	-1.00	-1.00	0.00	0.00	0.00	-0.79	0.00	1.00	1.00	0.37
lantern	syntax	0.00	1.00	1.00	0.00	0.00	0.00	-0.94	0.00	-1.00	-1.00	0.61
echo	vineyard	0.00	-1.00	0.00	0.00	0.00	0.00	1.88	0.00	0.00	0.00	0.48
nebula	snowman	0.00	-1.00	-1.00	0.00	0.00	0.00	0.18	-1.00	-2.00	-1.00	0.32
botany	teapot	0.00	-1.00	-1.00	0.00	0.00	0.00	0.44	-1.00	-2.00	-1.00	0.32
chisel	kaleidoscope	0.00	0.00	0.00	1.00	0.00	0.00	0.14	2.00	-1.00	-1.00	0.61
lava	teacup	0.00	-1.00	-1.00	1.00	0.00	0.00	1.97	0.00	-1.00	-1.00	0.40
entropy	orchard	0.00	-1.00	-1.00	0.00	0.00	0.00	0.37	-1.00	-1.00	0.00	0.36
axolotl	vineyard	0.00	1.00	0.00	0.00	0.00	0.00	-3.51	-2.00	0.00	0.00	0.42
clockwork	meadow	0.00	-1.00	0.00	0.00	0.00	0.00	-0.99	0.00	1.00	1.00	0.46
algebra	telescope	0.00	-1.00	0.00	0.00	0.00	0.00	8.75	0.00	-2.00	-1.00	0.63
arcade	topaz	0.00	0.00	0.00	0.00	0.00	0.00	2.28	0.00	0.00	0.00	0.55
asteroid	compass	0.00	-1.00	0.00	1.00	0.00	0.00	-0.86	-1.00	0.00	0.00	0.45
bicycle	nebula	0.00	1.00	1.00	0.00	0.00	0.00	1.99	0.00	0.00	0.00	0.70
bungalow	entropy	0.00	1.00	0.00	0.00	0.00	0.00	-1.20	0.00	2.00	1.00	0.54
carnation	gnome	0.00	0.00	0.00	0.00	0.00	0.00	-1.77	-2.00	-1.00	-1.00	0.35
cinnamon	harmonica	0.00	0.00	1.00	0.00	0.00	0.00	2.30	1.00	0.00	0.00	0.69
coral	syntax	0.00	1.00	1.00	0.00	0.00	0.00	-0.02	0.00	-1.00	-1.00	0.63
dandelion	pendulum	0.00	0.00	0.00	0.00	0.00	0.00	-0.53	-1.00	1.00	0.00	0.41
delirium	telescope	0.00	-1.00	-1.00	0.00	1.00	0.00	-1.44	-1.00	-2.00	-1.00	0.29
anchors	sandstorms	0.00	1.00	0.00	0.00	-1.00	0.00	3.42	0.00	-1.00	-1.00	0.59
scissors	volcanoes	0.00	0.00	1.00	-1.00	0.00	0.00	0.58	1.00	0.00	0.00	0.59
equations	lanterns	0.00	-1.00	0.00	0.00	0.00	0.00	2.30	-1.00	0.00	0.00	0.45
satellites	tulips	0.00	-1.00	0.00	0.00	0.00	0.00	1.49	-1.00	1.00	1.00	0.48
compasses	hedgehogs	0.00	-1.00	0.00	0.00	0.00	0.00	-0.61	-1.00	-2.00	-1.00	0.40
comets	neckties	0.00	-1.00	0.00	1.00	0.00	0.00	2.29	0.00	-1.00	-1.00	0.52
castles	headphones	0.00	0.00	-1.00	1.00	0.00	0.00	-1.10	0.00	-1.00	-1.00	0.40
paperclips	pyramids	0.00	0.00	1.00	-1.00	0.00	0.00	-2.88	0.00	0.00	0.00	0.48
constellations	kettles	0.00	-1.00	0.00	0.00	0.00	0.00	0.94	-2.00	0.00	0.00	0.39
kaleidoscopes	whales	0.00	-1.00	-1.00	-1.00	0.00	0.00	-4.65	-3.00	0.00	0.00	0.15
meadows	pianos	0.00	-1.00	-1.00	0.00	0.00	0.00	1.15	1.00	0.00	0.00	0.40
magnets	zebras	0.00	-1.00	1.00	0.00	0.00	0.00	1.82	0.00	0.00	0.00	0.59
parrots	submarines	0.00	1.00	0.00	-1.00	0.00	0.00	-0.35	1.00	-1.00	-1.00	0.49
crayons	jungles	0.00	1.00	1.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.67
harbor	teapot	0.00	-1.00	0.00	0.00	0.00	0.00	2.56	0.00	-1.00	-1.00	0.46
notebook	quicksand	0.00	0.00	1.00	-1.00	0.00	0.00	3.31	0.00	0.00	0.00	0.61
glacier	lantern	0.00	-1.00	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.45
microscope	puddle	0.00	0.00	0.00	0.00	0.00	0.00	1.36	-1.00	1.00	1.00	0.54
compass	swan	0.00	-1.00	0.00	0.00	0.00	0.00	0.20	-1.00	-1.00	-1.00	0.37
bonsai	cathedral	0.00	1.00	0.00	-1.00	0.00	0.00	-2.00	1.00	1.00	1.00	0.54
honeycomb	violin	0.00	0.00	-1.00	0.00	0.00	0.00	-1.40	0.00	0.00	0.00	0.37
sailboat	stadium	0.00	0.00	0.00	0.00	0.00	0.00	-3.21	1.00	2.00	1.00	0.47

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Stress	AbsPref
acorns	skyscrapers	0.00	1.00	0.00	0.00	0.00	0.00	-0.40	1.00	1.00	1.00	0.64
bell	trellis	0.00	0.00	1.00	0.00	0.00	0.00	3.33	1.00	1.00	1.00	0.74
inkwell	kite	0.00	0.00	-1.00	0.00	0.00	0.00	-3.21	-1.00	0.00	0.00	0.30
foxglove	trombone	0.00	0.00	-1.00	0.00	0.00	0.00	-1.89	0.00	1.00	1.00	0.40
carousel	quill	0.00	0.00	1.00	0.00	0.00	0.00	0.94	-2.00	0.00	0.00	0.55
lighthouse	onion	0.00	-1.00	-1.00	0.00	0.00	0.00	-1.15	0.00	1.00	1.00	0.36
cactus	chessboard	0.00	0.00	0.00	0.00	1.00	0.00	2.08	0.00	-1.00	-1.00	0.51
gallery	raindrop	0.00	0.00	0.00	0.00	0.00	0.00	5.06	-1.00	-2.00	-1.00	0.58
cricket	plow	0.00	1.00	0.00	-1.00	0.00	0.00	2.37	-1.00	-1.00	-1.00	0.47
gingerbread	fresco	0.00	0.00	1.00	0.00	0.00	0.00	0.34	-1.00	1.00	1.00	0.62
cello	sunflower	0.00	0.00	0.00	0.00	0.00	0.00	-0.42	1.00	0.00	0.00	0.54
archway	quilt	0.00	0.00	0.00	0.00	0.00	0.00	-2.77	-1.00	0.00	0.00	0.41
compass	haystack	0.00	0.00	0.00	0.00	0.00	0.00	2.34	0.00	-1.00	-1.00	0.51
beacon	millipede	0.00	-1.00	0.00	0.00	1.00	0.00	4.03	1.00	-1.00	-1.00	0.53
parchment	windmill	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	-1.00	-1.00	0.48
candlestick	meadow	0.00	1.00	0.00	0.00	0.00	0.00	-1.46	-1.00	1.00	1.00	0.54

Table 11: Full list of binomials as well as their constraints.