

Perplexity as a Measure for Temporal Common Sense in Pre-Trained Language Models

Tai Mai

mai@cl.uni-heidelberg.de

Abstract

Pre-trained language models are thought to learn both linguistic capabilities, as well as factual knowledge and common sense during their pre-training. To measure and interact with the unadulterated intuition obtained during pre-training, the pre-trained language model has to be used as is, without any further training such as fine-tuning. As a case study, this project attempts to capture the temporal common sense of pre-trained language models such as GPT-2, RoBERTa-base and Llama-3.1, using perplexity as an indicator and Allen’s interval calculus as the annotation scheme for representing temporal relations. Two hypotheses are posed which relate back to the assumption that sentences reflecting common sense should be associated with a lower perplexity than sentences that do not. A dataset is constructed automatically and curated manually. The results of experiments conducted in this project did not support the hypotheses, indicating a possible fault in the data design or experimental setup.

1 Introduction

Language models nowadays see an ever increasing amount of data during their pre-training. For example, GPT-2 (Radford et al., 2019) was trained on 40 GB of text scraped from the web. RoBERTa (Liu et al., 2019) was trained on over 160 GB of text, comprised of books, Wikipedia and news articles, as well as text scraped from the web. Llama-3.1 (Dubey et al., 2024) is a very recent model that was trained on over 15 trillion tokens of publicly available online data. While these figures are not directly comparable, they speak to the vast amounts of data that are used to pre-train a language model. Language models are typically pre-trained in an unsupervised manner. For example, causal language models, such as GPT-2 (Radford et al., 2019) and Llama-3.1, (Dubey et al., 2024) generate text by predicting tokens one by one, each conditioned on the preceding tokens. Their objective is to cor-

rectly model their training data by maximizing the likelihood of token sequences that are present in the training data. On the other hand, masked language models, such as RoBERTa (Liu et al., 2019), are trained on another unsupervised task known as masked language modeling. In contrast to causal language modeling, masked language modeling is the process of randomly masking tokens in a training sentence and tasking the language model with reconstructing the original sentence based on the context surrounding the masked token. Pre-training a language model is designed to enable it to form an intuition for the language such that the resulting foundation model’s intuition can be built upon in different downstream tasks, such as classification. This process of specializing a language model for a specific downstream task is called fine-tuning and often entails adding and training a few additional neural layers — commonly called a task head — to the pre-trained model to make it compatible with the desired downstream task. Besides obtaining an intuition for the language, pre-trained language models have also been shown to be able to form a semblance of factual common sense through pre-training (Zhou et al., 2020). However, testing a pre-trained language model for common sense in, say, temporal relations can be tricky. For example, employing a language model to classify the relation between two events in time is, in its most straightforward form, a classification task that would most commonly require fine-tuning a classification head on top of the pre-trained model. Querying a model after fine-tuning, however, makes it impossible to say whether its understanding of time was obtained during pre-training or as a result of fine-tuning. To isolate the effects of pre-training alone, this kind of evaluation has to be done on the foundation model directly, without any further training.

This project attempts an approach to investigate the amount of temporal common sense that different pre-trained foundation language models ob-

tain through pre-training. In lieu of fine-tuning the model for a classification task, the foundation model’s perplexity metric is chosen as a proxy for how natural it judges a particular sentence. The results show that this implementation of perplexity is perhaps not a good measure to correlate with temporal common sense.

2 Methods

2.1 Perplexity metrics

Perplexity in language models can be interpreted as a measure of how likely a language model thinks a token sequence is. The higher the perplexity metric is for a particular sequence, the more “surprised” the language model is to see that sequence. Perplexity is most commonly used for causal language modeling (CLM) and is defined as follows¹,

$$\text{PPL}(X) = \exp \left\{ -\frac{1}{t} \sum_i^t \log p_{\text{CLM}}(x_i | x_{<i}) \right\}, \quad (1)$$

where X denotes the token sequence of tokens x , t is the number of tokens in the sequence, and $p(x_i | x_{<i})$ represents the probability of token x_i at position i , given all the previous tokens $x_{<i}$. Therefore, this definition of perplexity is very specific to causal language modeling and cannot be directly applied to masked language modeling.

To obtain an analogous metric for masked language modeling (MLM), Salazar et al. (2020) propose the following pseudo-perplexity metric.

$$\text{PPPL}(X) := \exp \left\{ -\frac{1}{t} \sum_i^t \log p_{\text{MLM}}(x_i | X_{\setminus x_i}) \right\} \quad (2)$$

The only difference to Equation 1 is in the context on which the token x_i is conditioned. In causal language modeling, the choice of token x_i at position i is conditioned on all tokens to the left of position i . Future tokens cannot be taken into account in causal language modeling, therefore the context is unilateral. In masked language modeling, however, tokens in the sequence are chosen and masked at random. To reconstruct a masked token x_i , masked language models are provided with all other unmasked tokens in the sequence, both before and after position i . Therefore, the context in masked language modeling is bilateral.

These metrics are chosen to gauge how natural a language model thinks a sequence is, specifically in the case of temporal relations described in the following section.

2.2 Allen’s Interval Calculus

Allen (1983) proposes a system of relating two events, e_1 and e_2 , to each other in time via a set of basic temporal relations. This set can be boiled down to the following seven relations.

- *before*(e_1, e_2): e_1 starts and ends before e_2 starts
- *meets*(e_1, e_2): e_1 starts before e_2 and e_2 starts when e_1 ends
- *overlaps*(e_1, e_2): e_1 starts before e_2 and e_1 ends between the start and end of e_2
- *starts*(e_1, e_2): e_1 starts at the same time as e_2 and e_1 ends before e_2
- *during*(e_1, e_2): e_1 starts after e_2 and e_1 ends before e_2
- *finishes*(e_1, e_2): e_1 starts after e_2 and e_1 ends at the same time as e_2
- *equals*(e_1, e_2): e_1 starts and ends at the same time as e_2

The complete set proposed by Allen (1983) also contains inverses of the previous relations. However, these inverses can also be modeled by the seven relations above by simply swapping the positions of the events as arguments to the relations, i.e. *after*(e_1, e_2) is equivalent to *before*(e_2, e_1). Therefore, this project will only consider the seven relations mentioned above. These relations can then also be verbalized into natural language. For example, one possible template verbalization of the relation *before* could be “ e_1 happens before e_2 .” An event tuple such as $\langle \text{birth}, \text{death} \rangle$ can then be connected and verbalized as “*Birth happens before death.*” The objective of this project is to investigate how well the perplexity values of these verbalizations correlate with temporal common sense. Specifically, the first hypothesis investigated in this project is that a verbalization that reflects temporal common sense, such as “*Birth happens before death,*” should receive a lower perplexity score than one that does not reflect temporal common sense, such as “*Birth happens during death.*”

¹<https://huggingface.co/docs/transformers/en/perplexity>

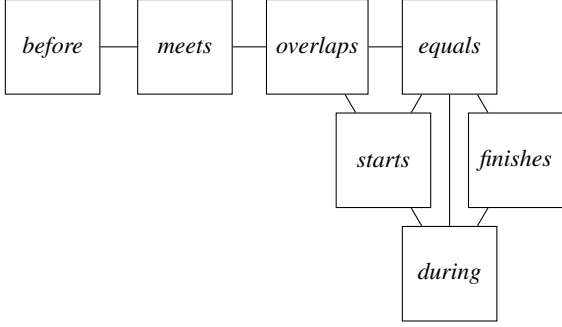


Figure 1: Conceptual neighborhood according to Freksa (1992), excluding the inverse relations.

2.3 Conceptual Neighborhood

Freksa (1992) arranges the relations proposed by Allen (1983) in an undirected graph structure to encode how closely the relations are related to each other. Figure 1 shows this graph after excluding the inverse relations. With this graph structure in mind, this project attempts to investigate whether the perplexity scores correlate with the conceptual distance in graph hops. As an example, the relations *before* and *during* are separated by four node hops in the conceptual neighborhood graph. As mentioned in Section 2.2, we hypothesize that an atypical verbalization, such as “*Birth happens during death*,” should receive a higher perplexity score than a common sense verbalization, such as “*Birth happens before death*.” Therefore, the second hypothesis investigated in this project is that this change in perplexity should correlate with the degrees of separation between the two relations when plugging an event tuple from one relation into a verbalization of another relation.

3 Data

To test these two hypotheses, a dataset containing examples of Allen’s interval calculus relations is needed. Since such a dataset does not exist to the best of our knowledge, a simple dataset has been constructed by utilizing Claude-3.5 Sonnet² and then curated by hand. For each of the 7 temporal relations, a set of 20 event tuples are gathered as examples of the temporal relation in question. For the relation *before*, one of the examples collected is the tuple $\langle \text{birth}, \text{death} \rangle$. Furthermore, for each of the relations, a set of 10 possible template verbalizations is collected, with $\{\text{event1}\}$ and $\{\text{event2}\}$ as placeholders for respective temporal events, for

example “ $\{\text{event1}\}$ happens before $\{\text{event2}\}$.” When placing every event tuple into every template verbalization of every relation, this yields a total dataset size of $20 \cdot 10 \cdot 7 = 1400$ verbalized examples.

Before Claude-3.5 Sonnet was released, GPT-3.5 and GPT-4o³ were also considered for the task of generating the dataset. However, upon manual inspection of the generated data, these models did not prove to be able to come up with (enough) correct and/or diverse examples and were subsequently dropped in favor of Claude-3.5 Sonnet, which required a lot less manual intervention. The prompting steps used to generate this dataset can be seen in Appendix A.1. The dataset and code can be accessed publicly online⁴.

4 Evaluation

4.1 Confusion Matrices

4.1.1 Relation Prediction

To relate back to the classification task, the perplexity values can theoretically be used to predict the correct relation. As explained in Section 3, a set of 20 event tuples are collected as examples for every relation. One such example for the relation *before* is the event tuple $\langle \text{birth}, \text{death} \rangle$. This event tuple is then plugged into all 10 template verbalizations of each of the 7 relations, resulting in verbalizations such as “*Birth happens during death*.” The perplexity values are then aggregated as averages for each verbalized relation across all its 10 template verbalizations. The relation whose template verbalizations achieve the lowest perplexity scores on average is then chosen as the predicted relation. Under the first hypothesis, verbalizing with the correct relation’s templates should receive lower perplexity scores than the other, more atypical relations. This can be represented as a confusion matrix that contains a tally of how often a relation (in the columns) was predicted for an event tuple coming from its true relation (in the rows). Figure 2 shows the resulting confusion matrices.

Assuming the first hypothesis investigated in this project, the perplexity values should be lowest on the matrices’ diagonals and higher for off-diagonal cells. Therefore, one might expect the diagonals in the confusion matrices to contain the highest

³<https://chatgpt.com>

⁴https://github.com/Tai-Mai/temporal-common-sense/blob/main/data/claude_examples.json

²<https://www.anthropic.com/news/claude-3-5-sonnet>

	before	meets	overlaps	starts	during	finishes	equals
before	1	0	13	0	0	0	6
meets	2	0	13	0	0	0	5
overlaps	0	0	16	0	1	0	3
starts	1	0	11	3	3	0	2
during	0	0	11	2	4	0	3
finishes	1	0	11	1	2	0	5
equals	0	0	2	0	2	0	16

(a) GPT-2

	before	meets	overlaps	starts	during	finishes	equals
before	1	0	8	1	1	9	0
meets	1	0	4	1	1	13	0
overlaps	0	0	9	2	2	7	0
starts	0	0	2	0	1	17	0
during	0	0	2	0	4	14	0
finishes	0	0	0	1	1	18	0
equals	0	0	1	0	0	19	0

(b) GPT-2 (normalized)

	before	meets	overlaps	starts	during	finishes	equals
before	0	0	10	0	0	0	10
meets	2	0	9	0	0	0	9
overlaps	1	0	8	1	0	0	10
starts	0	0	14	5	1	0	0
during	1	0	8	1	1	0	9
finishes	0	0	8	4	0	2	6
equals	0	0	2	0	0	0	18

(c) RoBERTa-base

	before	meets	overlaps	starts	during	finishes	equals
before	0	0	2	0	17	0	1
meets	0	0	1	0	19	0	0
overlaps	0	0	1	0	19	0	0
starts	0	0	0	1	19	0	0
during	0	0	0	1	18	1	0
finishes	0	0	0	0	19	1	0
equals	0	0	0	0	20	0	0

(d) RoBERTa-base (normalized)

	before	meets	overlaps	starts	during	finishes	equals
before	0	0	12	0	0	0	8
meets	0	0	13	0	1	0	6
overlaps	0	0	14	1	1	0	4
starts	0	0	5	15	0	0	0
during	0	0	17	1	1	0	1
finishes	0	0	8	4	0	5	3
equals	0	0	0	1	0	0	19

(e) Llama-3.1

	before	meets	overlaps	starts	during	finishes	equals
before	0	1	2	2	14	0	1
meets	0	1	0	0	19	0	0
overlaps	1	1	3	1	14	0	0
starts	0	0	0	0	20	0	0
during	0	0	0	2	17	1	0
finishes	0	0	0	0	18	2	0
equals	0	0	0	1	19	0	0

(f) Llama-3.1 (normalized)

Figure 2: Counts of predicted relations (columns), given the event tuples from the true relation (rows).

prediction counts. However, this does not prove to be the case. Instead, there seems to emerge a similar pattern among all three models where the columns *overlaps* and *equals* tend to contain the highest counts. This seems to indicate that the perplexity values are highly dependent on the template verbalizations. If a template verbalization already contains an unlikely sequence of tokens before plugging in an actual event tuple, it could dominate the perplexity value and obscure any effect that the event tuple has.

As an attempt to mitigate this effect, a strategy to normalize the perplexity values was devised. Instead of choosing the raw perplexity values for the predictions, as has been the case until now, the perplexity values are normalized by using reference verbalizations. A reference verbalization is constructed by substituting the placeholders {event1} and {event2} in the template with “an event” and “another event,” respectively. In the case of the relation *before*, one such reference verbalization would be “An event happens before another event”.

$$\frac{\text{PPL}(\text{before}(\text{“birth”}, \text{“death”}))}{\text{PPL}(\text{before}(\text{“an event”}, \text{“another event”}))}. \quad (3)$$

This is done to cancel out the effects of the verbalization by itself and distill the effect of the *interaction* between the event tuples and the verbalizations in the ratio. In the case of masked language models, the perplexity metric $\text{PPL}(\cdot)$ can be substituted with the pseudo-perplexity $\text{PPPL}(\cdot)$. Choosing the lowest average perplexity ratio for the prediction yields the normalized confusion matrices in Figure 2. Unfortunately, this strategy does not seem to have the desired effect, as an even more concentrated column pattern seems to emerge.

4.1.2 Perplexities

Instead of visualizing the prediction counts, we further opt to visualize the perplexity values themselves. This results in new confusion matrices with the original, true relation in the rows and the verbalized relation in the columns, as can be seen in Figure 3.

Consistent with the previous confusion matrices, verbalizing event tuples with the relation *overlaps* and *equals* tends to result in relatively low perplexity values compared to other relations. This manifests as dark *overlaps* and *finishes* columns in the heat map. Meanwhile, the *finishes* column contains the highest perplexity value in all three models.

Once again, the strategy to normalize the perplexity values was employed and the resulting perplexity value ratios are displayed in the confusion matrices. This time, this strategy seems to be successful in removing the column pattern observed before, as can be seen in Figure 4. Numerically, one can observe that most of these ratios are greater than 1, indicating that the specific verbalizations with instantiated events obtain higher perplexity values than the generic reference verbalizations. Only in the case of Llama-3.1, the *equals* row consistently contains ratios smaller than 1. In fact, while the column pattern has been mitigated, the *equals* row brings to attention another pattern that emerges in the rows of the confusion matrices. In all three models, the *equals* row contains the lowest perplexity ratios in all columns. This suggests that the event tuples from the relation *equals* are associated with lower perplexity scores, no matter which relation they are verbalized with. For both causal language models, GPT-2 and Llama-3.1, a similar pattern can be observed in the *before* and *overlaps* rows which tend to be brighter than the other rows, indicating that the event tuples from these relations tend to be associated with higher perplexity values, regardless of the relation they are verbalized with.

4.2 Correlation Analysis

As discussed in Section 4.1.2, the hypothesized diagonal pattern did not seem to be discernible. To confirm this observation quantitatively, we evaluate the second hypothesis of this project. Under the second hypothesis, when verbalizing an event tuple $\langle e_1, e_2 \rangle$ using another relation than its original relation rel_{og} , the perplexity values $\text{PPL}(\cdot)$ should scale according to the conceptual distance $d_{\text{hops}}(rel, rel_{og})$ between the original relation and the confused relation, measured in graph hops in Freksa’s conceptual neighborhood graph (Figure 1). For example, the hypothesis assumes for the event tuple $\langle \text{birth}, \text{death} \rangle$ that

$$\text{PPL}(\text{meets}(\text{birth}, \text{death})) < \text{PPL}(\text{finishes}(\text{birth}, \text{death})), \quad (4)$$

because *meets* is conceptually closer to the tuple’s original relation *before* ($d_{\text{hops}}(\text{meets}, \text{before}) = 1$) than *finishes* is ($d_{\text{hops}}(\text{finishes}, \text{before}) = 4$). This means that an event tuple $\langle e_1, e_2 \rangle$ should receive higher absolute perplexity scores when verbalized using wrong relations $rel(e_1, e_2)$ that are more conceptually distant from the tuple’s original relation $rel_{og}(e_1, e_2)$, compared to relations that are concep-

	before	meets	overlaps	starts	during	finishes	equals
before	2984.83 ± 15857.07	1513.99 ± 3247.95	414.28 ± 524.48	1890.22 ± 6944.68	4464.15 ± 24339.47	6256.40 ± 21066.53	1172.27 ± 5768.10
meets	1108.21 ± 4948.94	1115.70 ± 4830.70	337.26 ± 521.11	1337.81 ± 6752.78	1600.68 ± 8179.04	3721.45 ± 14662.58	715.09 ± 2781.10
overlaps	3087.77 ± 10379.99	2030.44 ± 4023.18	602.74 ± 807.17	2133.72 ± 5799.02	4431.64 ± 17101.91	6471.13 ± 18915.42	1849.40 ± 7438.36
starts	743.85 ± 1969.34	1225.52 ± 7430.95	411.96 ± 611.01	730.95 ± 2998.13	969.99 ± 3947.74	2190.22 ± 11901.11	929.31 ± 6058.70
during	1073.26 ± 3782.34	924.84 ± 1794.59	391.63 ± 449.59	934.53 ± 3513.53	1693.64 ± 8081.68	2778.34 ± 13470.27	703.36 ± 2658.25
finishes	1273.55 ± 6562.82	987.78 ± 2137.62	425.29 ± 479.68	694.65 ± 2107.22	1217.84 ± 5279.73	1311.82 ± 4217.64	651.85 ± 2398.93
equals	308.30 ± 845.95	334.84 ± 888.09	178.47 ± 234.22	317.70 ± 1344.36	330.64 ± 1481.00	1010.99 ± 7168.72	223.30 ± 1172.12

(a) GPT-2

	before	meets	overlaps	starts	during	finishes	equals
before	115.98 ± 636.18	103.18 ± 703.74	16.45 ± 24.70	47.55 ± 115.24	139.10 ± 665.97	144.21 ± 388.11	20.75 ± 52.22
meets	38.56 ± 84.25	52.45 ± 339.70	19.98 ± 77.31	83.63 ± 701.49	81.59 ± 306.21	191.60 ± 1445.17	32.25 ± 203.56
overlaps	108.47 ± 300.49	64.60 ± 210.86	27.51 ± 51.55	64.14 ± 158.72	93.88 ± 213.26	150.75 ± 530.49	32.16 ± 83.43
starts	74.06 ± 162.86	80.07 ± 245.96	33.15 ± 78.66	49.02 ± 169.66	115.73 ± 388.80	127.38 ± 566.16	71.98 ± 447.15
during	61.96 ± 119.27	59.06 ± 133.99	28.78 ± 66.30	46.47 ± 94.24	93.02 ± 297.71	61.69 ± 118.87	33.20 ± 93.01
finishes	58.98 ± 104.33	55.83 ± 103.87	30.54 ± 53.55	43.24 ± 95.76	77.24 ± 194.51	58.90 ± 147.33	34.93 ± 75.91
equals	26.28 ± 82.48	16.95 ± 27.19	10.03 ± 9.25	18.90 ± 21.41	22.75 ± 62.69	21.08 ± 31.04	7.68 ± 7.51

(b) RoBERTa-base

	before	meets	overlaps	starts	during	finishes	equals
before	4174.32 ± 35118.93	1612.78 ± 5974.53	504.68 ± 1494.81	2786.86 ± 18784.01	4695.45 ± 20236.12	9931.14 ± 61150.82	900.19 ± 5060.28
meets	520.63 ± 897.28	746.01 ± 3713.41	326.62 ± 763.73	715.03 ± 2714.94	1146.80 ± 5574.56	1848.55 ± 7156.71	681.40 ± 3303.30
overlaps	2686.26 ± 11859.85	1824.65 ± 5026.93	539.61 ± 856.20	1339.61 ± 3577.89	3206.07 ± 11905.50	4398.39 ± 17393.41	1063.41 ± 3902.59
starts	599.30 ± 1443.76	767.26 ± 2365.25	384.44 ± 764.10	429.05 ± 1536.46	841.24 ± 3081.23	1889.10 ± 12392.25	529.81 ± 1611.46
during	787.51 ± 2150.14	651.21 ± 1080.01	287.00 ± 305.31	471.53 ± 829.94	965.43 ± 3525.90	773.29 ± 1824.04	369.66 ± 634.70
finishes	657.56 ± 1180.09	660.82 ± 1249.24	417.13 ± 736.64	514.26 ± 973.45	741.58 ± 1606.33	573.48 ± 1294.19	497.69 ± 1349.38
equals	183.81 ± 275.58	182.08 ± 303.41	107.95 ± 96.29	123.76 ± 132.22	159.63 ± 358.99	153.60 ± 204.07	75.46 ± 55.97

(c) Llama-3.1

Figure 3: Average perplexity values when plugging event tuples from the true relation (rows) into verbalizations of all other relations (columns). For the masked language model RoBERTa-base, the pseudo-perplexity is shown instead.

	before	meets	overlaps	starts	during	finishes	equals
before	4.97 ± 14.91	3.16 ± 4.02	1.86 ± 1.59	2.78 ± 4.46	5.52 ± 16.02	3.09 ± 5.53	3.86 ± 12.68
meets	2.48 ± 4.78	2.30 ± 5.81	1.51 ± 1.31	1.99 ± 4.17	2.90 ± 5.76	1.97 ± 3.92	2.68 ± 6.08
overlaps	6.29 ± 10.15	4.19 ± 4.74	2.65 ± 2.33	3.55 ± 4.41	6.83 ± 15.26	3.54 ± 4.72	5.98 ± 16.26
starts	2.41 ± 3.60	2.68 ± 9.02	1.96 ± 2.88	1.58 ± 2.63	2.54 ± 3.67	1.45 ± 3.21	3.52 ± 13.36
during	2.76 ± 4.01	2.30 ± 2.27	1.79 ± 1.30	1.76 ± 2.45	3.02 ± 6.92	1.77 ± 3.52	2.86 ± 5.76
finishes	2.98 ± 6.44	2.34 ± 2.89	1.93 ± 1.60	1.58 ± 1.95	2.57 ± 3.96	1.19 ± 1.44	2.92 ± 5.30
equals	1.06 ± 1.29	1.08 ± 1.28	0.92 ± 0.76	0.87 ± 1.11	1.14 ± 1.47	0.79 ± 2.04	1.12 ± 2.61

(a) GPT-2

	before	meets	overlaps	starts	during	finishes	equals
before	5.00 ± 9.96	4.54 ± 13.51	2.47 ± 2.47	3.00 ± 5.32	3.44 ± 6.31	4.83 ± 9.08	2.56 ± 3.24
meets	3.10 ± 5.75	4.41 ± 23.20	2.89 ± 7.32	4.51 ± 29.92	3.09 ± 5.98	5.57 ± 32.04	3.00 ± 9.09
overlaps	7.91 ± 13.80	4.75 ± 7.84	4.12 ± 5.39	4.42 ± 8.36	5.01 ± 8.84	5.57 ± 13.98	3.61 ± 3.90
starts	6.39 ± 14.93	7.74 ± 20.52	5.27 ± 9.44	3.34 ± 8.86	4.57 ± 15.99	4.54 ± 13.54	6.81 ± 21.04
during	4.45 ± 6.34	5.05 ± 8.56	4.24 ± 6.50	3.07 ± 5.04	2.85 ± 5.23	2.62 ± 4.45	3.86 ± 4.96
finishes	4.38 ± 5.94	4.99 ± 6.90	4.67 ± 5.55	2.87 ± 4.68	2.73 ± 4.22	2.39 ± 3.81	4.39 ± 4.65
equals	2.27 ± 2.39	1.76 ± 1.42	1.80 ± 1.18	1.56 ± 2.03	1.34 ± 2.03	1.08 ± 1.10	1.37 ± 0.91

(b) RoBERTa-base

	before	meets	overlaps	starts	during	finishes	equals
before	5.25 ± 18.90	3.60 ± 8.36	3.04 ± 5.51	4.41 ± 19.68	3.95 ± 6.80	9.63 ± 53.40	3.53 ± 9.25
meets	2.16 ± 3.96	2.03 ± 5.09	2.19 ± 3.11	1.71 ± 3.30	2.22 ± 3.28	2.32 ± 6.15	2.84 ± 6.31
overlaps	6.06 ± 7.56	4.34 ± 6.65	3.45 ± 3.24	3.15 ± 4.59	4.29 ± 5.09	4.92 ± 14.40	4.37 ± 7.02
starts	2.31 ± 3.27	2.28 ± 3.34	2.55 ± 2.83	1.11 ± 1.92	1.90 ± 2.70	2.22 ± 10.48	2.79 ± 3.24
during	2.24 ± 1.90	2.14 ± 1.60	2.14 ± 1.39	1.37 ± 1.36	1.55 ± 1.67	1.26 ± 1.70	2.28 ± 1.58
finishes	2.38 ± 2.44	2.17 ± 2.03	2.79 ± 2.78	1.42 ± 1.53	1.67 ± 1.69	0.95 ± 1.51	2.85 ± 3.16
equals	0.81 ± 0.71	0.91 ± 0.77	0.99 ± 0.76	0.51 ± 0.49	0.61 ± 0.62	0.37 ± 0.28	0.66 ± 0.38

(c) Llama-3.1

Figure 4: Average perplexity ratios *after normalizing* the perplexity values. Perplexity values are normalized by dividing by the perplexity values of reference verbalizations. True relation in the rows, verbalized relation in the columns. For the masked language model RoBERTa-base, the pseudo-perplexity ratio is shown instead.

tually closer to rel_{og} . Formally, we investigate the following correlation for all relations rel and all event tuples $\langle e_1, e_2 \rangle$:

$$\text{Cor}(\text{PPL}(rel(e_1, e_2)), d_{\text{hops}}(rel, rel_{og})) \quad (5)$$

Table 1 lists the correlation coefficients Pearson’s r , Spearman’s ρ , and Kendall’s τ . They confirm that there is no notable correlation between the absolute perplexity scores and conceptual distance in graph hops. One observation that could be made is that the Spearman’s ρ is slightly higher for Llama-3.1 compared to the weaker models GPT-2 and RoBERTa-base.

Instead of correlating the absolute perplexity values with the conceptual distance in graph hops, we try substituting it with the *change* in perplexity $\Delta \text{PPL}(\cdot)$, relative to the original relation’s average perplexity.

$$\text{Cor}(\Delta \text{PPL}(rel(e_1, e_2)), d_{\text{hops}}(rel, rel_{og})) \quad (6)$$

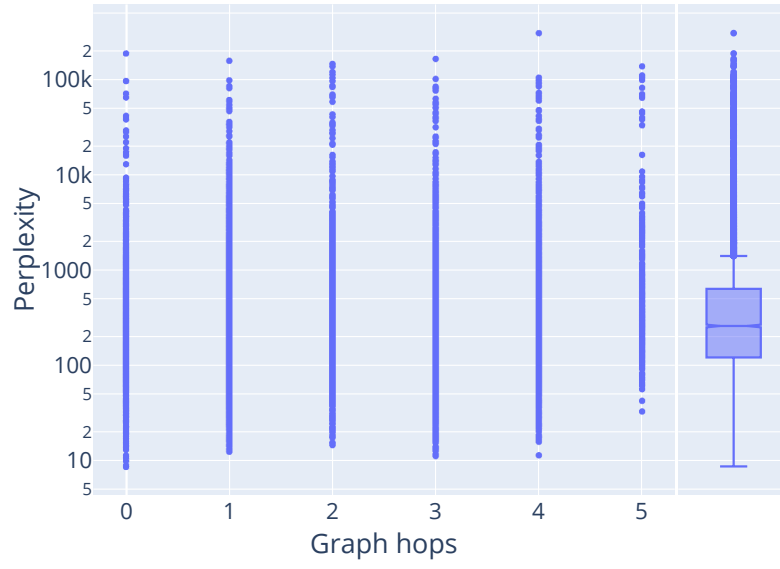
The rationale is the same as before: The change in perplexity should scale with the conceptual distance. Table 2 shows that Spearman’s ρ and Kendall’s τ have now even switched to negative values while being of a similar magnitude as the previous values (near 0). This correlation analysis has therefore confirmed that this experimental setup does not exhibit any correlation between perplexity and conceptual distance.

5 Discussion and Conclusion

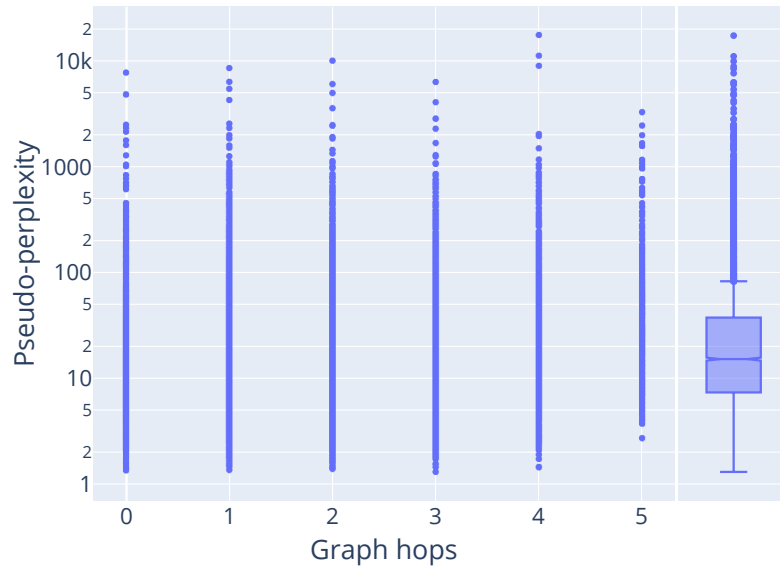
This project aimed to investigate two hypotheses. The first hypothesis is that perplexity is higher for a temporal event tuple $\langle e_1, e_2 \rangle$ if it is verbalized with a wrong relation compared to when it is verbalized with a relation that reflects common sense. The second hypothesis is that the absolute perplexity values or relative perplexity changes are correlated with the conceptual distance between the confused relations, according to the conceptual neighborhood graph by Freksa (1992). The experimental setup of this project was unable to prove these hypotheses. In fact, the results seem to contradict the two hypotheses, showing no notable correlation. However, while these results do not confirm the hypotheses, we cannot assume that they disprove them either. As can be seen in Figures 3 and 4, the standard deviations among the perplexities is often higher than the average itself. This indicates that the perplexity metric is very unstable and often yields very high outlier values that drive up

the standard deviation. These outliers can be seen in the marginal box plots in Figure 5. We have to therefore assume that this experimental setup is invalid and does not adequately measure what it purports to measure. Two possible sources of error are faulty implementations of perplexity or pseudo-perplexity, or, somewhat more likely, a subpar design of the dataset. The dataset is intended to work in a way such that the event tuples can be plugged into all seven relations and still be more or less grammatically correct. The problem likely stems from the fact that, while grammatically correct, the resulting verbalizations are still very unnatural and do not reflect how humans would typically express them. For example, one of the template verbalizations for the relation *before* is “{event1} starts and ends before {event2}”. While this template accurately reflects the definition of the *before* relation according to Allen (1983), it is not how humans would typically phrase it in everyday speech. We suspect that this might be what could be driving up perplexity values for verbalizations, even if they conceptually agree with common sense. While the normalization strategy aimed to mitigate this issue, and seemed to be successful in Figure 4, it did not help in the prediction of the correct relation and also did not exhibit the hypothesized correlation. Further research has to be conducted to examine whether this could be indicating that Freksa’s conceptual neighborhood graph does in fact not accurately reflect how language models intuit temporal relations.

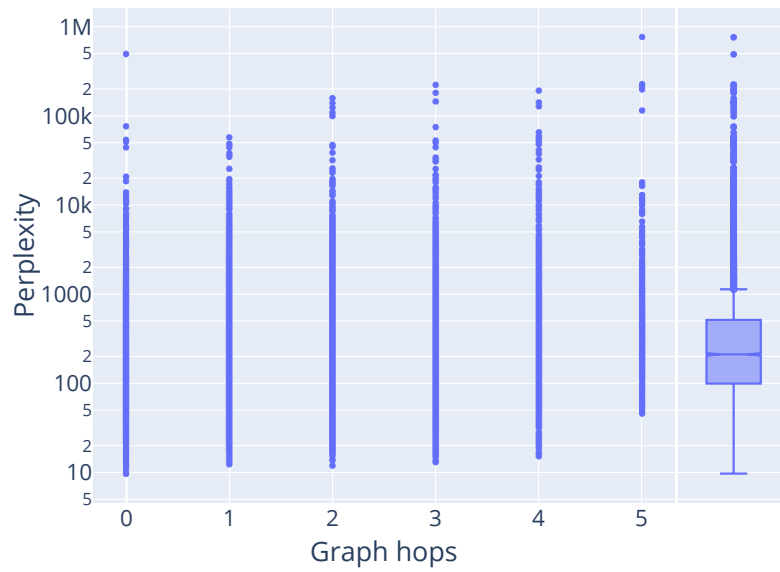
To improve on this experimental setup, a more natural way to verbalize event tuples into relations could be devised. While the strategy employed in this project naively substitutes the placeholder events {event1} and {event2} with their respective examples, this can result in unnatural sentences. Instead, one might try to verbalize by hand or query a capable large language model to formulate it into a more natural sentence. Naturally, this has to be another language model than the ones that are being tested, therefore Llama-3.1 — and arguably also Llama-2 — are out of the question. Using APIs such as ChatGPT/GPT 4o, Gemini (Team et al., 2024), or Claude 3.5 would require $20 \cdot 10 \cdot 7 = 1400$ additional paid API calls, which was out of the scope of this project. However, if budget permits, this might prove to be a viable option for future work to verbalize event tuples with relations more naturally.



(a) GPT-2



(b) RoBERTa-base



(c) Llama-3.1

Figure 5: Absolute perplexity scores. For the masked language model RoBERTa-base, the pseudo-perplexity values are shown instead.

Model	r	ρ	τ
GPT-2	0.0537	0.0654	0.0477
GPT-2 (normalized)	0.0407	0.0608	0.0444
RoBERTa-base	0.0266	0.0906	0.0661
RoBERTa-base (normalized)	0.0231	0.0876	0.0638
Llama-3.1 8B	0.0483	0.1230	0.0901
Llama-3.1 8B (normalized)	0.0571	0.1212	0.0884

Table 1: Pearson’s r , Spearman’s ρ , and Kendall’s τ correlation coefficients between *absolute perplexity values* and graph hops.

Model	r	ρ	τ
GPT-2	0.0319	−0.0968	−0.0698
GPT-2 (normalized)	0.0087	−0.0419	−0.0304
RoBERTa-base	0.0106	−0.0484	−0.0342
RoBERTa-base (normalized)	0.0094	−0.0126	−0.0087
Llama-3.1 8B	0.0242	−0.0665	−0.0469
Llama-3.1 8B (normalized)	0.0278	−0.0224	−0.0148

Table 2: Pearson’s r , Spearman’s ρ , and Kendall’s τ correlation coefficients between *perplexity deltas* and graph hops.

6 Related Works

Previous research has corroborated the idea of language models, especially large language models, learning common sense during pre-training. [Zhao et al. \(2024\)](#) leverage the common sense of large language models for large-scale task planning. [Zhou et al. \(2020\)](#) evaluate the common sense in pre-trained language models similarly to this project, by scoring sentences from various benchmark datasets using the log-likelihoods of sentences, or the equivalent of pseudo-log-likelihoods for masked language models. [Kauf and Ivanova \(2023\)](#) propose an improved way of sequence scoring for masked language models, which changes which tokens in the sequence are masked. This improved strategy showed to be more robust for out-of-vocabulary words. However, this project chooses to keep the simpler pseudo-perplexity metric for ease of implementation. Like this project, [Wenzel and Jatowt \(2023\)](#), too, concentrate on temporal common sense specifically and propose methods to enhance the language model’s capabilities by augmentation in order for them not to default to linguistic heuristics. However, to the best of our knowledge, this project is the first to use Allen’s interval calculus as an annotation scheme to capture temporal common sense.

References

- James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate

Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Roman Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gouget, Virginie Do, Vish Voleti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Golschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Mont-

gomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu

- Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. [The llama 3 herd of models](#).
- Christian Freksa. 1992. [Temporal reasoning based on semi-intervals](#). *Artificial Intelligence*, 54(1):199–227.
- Carina Kauf and Anna Ivanova. 2023. [A better way to do masked language model scoring](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. [Masked language model scoring](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqi, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anas White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomnech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Deendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gimnez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao

Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitaogong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çalar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihla, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiakowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufaret, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakievi, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Jun-

hyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chaitin, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohanane, Jonah Joughin, Egor Filonov, Tomasz Kpa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Ankin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercader, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Ålgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dekhtarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bølle, Dominik Paulus, Khyati Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yoge, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjit Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai

Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Åhdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Brainskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Pawe Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tume, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luwei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikoaj Rybiski, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith

Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebecca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev,

- Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivi re, Alanna Walton, Cl ment Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Pluciska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolichio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshiti Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahr Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atlas, Paulina Lee, V t List k, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul M ller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. [Gemini: A family of highly capable multimodal models](#).
- Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2308.00002*.
- Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9733–9740.

A Appendix

A.1 Claude-3.5 Sonnet Prompting for the Dataset Generation

1. What are the relations in James Allen's interval calculus?

- James Allen's Interval Calculus, introduced in 1983, defines 13 basic relations between time intervals. These relations describe all possible ways two time intervals can be related to each other. Here's a concise list of the 13 relations:

- (a) Before (precedes)
- (b) Meets
- (c) Overlaps
- (d) Starts
- (e) During
- (f) Finishes
- (g) Equals
- (h) Finished-by
- (i) Contains
- (j) Started-by
- (k) Overlapped-by
- (l) Met-by
- (m) After (follows)

These relations are mutually exclusive and exhaustive, meaning any two time intervals must be related by exactly one of these relations. Each relation also has an inverse (e.g., "before" is the inverse of "after"), which is why there are 13 relations instead of 7. Would you like me to elaborate on any specific relation or provide examples of how these relations are used in temporal reasoning?

2. Give me the relations without the inverses

- Certainly. If we consider only the base relations without their inverses, we can reduce James Allen's Interval Calculus to 7 fundamental relations. Here they are:

- (a) Before
- (b) Meets
- (c) Overlaps
- (d) Starts
- (e) During
- (f) Finishes
- (g) Equals

These 7 relations form the core of Allen's Interval Calculus. The other 6 relations in the full set of 13 are inverses of these (except for "Equals," which is its own inverse). Would you like me to explain how these relations are defined or provide examples of their use in temporal reasoning?

3. Please define each of these seven relations in Allen's Interval Calculus.

- Certainly. I'll define each of the seven base relations in Allen's Interval Calculus. Let's consider two time intervals, A and B:
 - (a) Before (A before B): Interval A ends before interval B starts. There is a gap between A and B.
 - (b) Meets (A meets B): The end point of interval A is exactly the start point of interval B. There is no gap between them.
 - (c) Overlaps (A overlaps B): Interval A starts before B, and they overlap for some duration. A ends sometime after B starts but before B ends.

- (d) Starts (A starts B): Both intervals start at the same point, but A ends before B ends.
- (e) During (A during B): Interval A starts after B starts and ends before B ends. A is completely contained within B.
- (f) Finishes (A finishes B): Both intervals end at the same point, but A starts after B starts.
- (g) Equals (A equals B): Both intervals start and end at the same points. They are identical in duration and position.

These definitions assume that intervals are non-zero in duration and have distinct start and end points. Each relation (except “Equals”) has an inverse relation in the full set of 13 relations. Would you like me to provide visual representations or examples of these relations?

4. For each of these relations, give me 10 commonsense examples of temporal events or periods . They should always and universally be true. Don’t use examples that are just “often” or “sometimes” the case. Make sure not to confuse “before” with “meets”. Also, don’t confuse “finishes” with “after”. For “meets”, don’t mention the point in time where they meet. Format the examples in JSON like this: { “before”: [{ “event1”: “breakfast”, “event2”: “lunch”, “example”: “breakfast happens before lunch.” }, { “event1”: ..., ... },] ... }

 - I understand your request. I’ll provide 10 commonsense examples for each of the seven relations in Allen’s Interval Calculus, ensuring they are always and universally true. I’ll format them in JSON as requested, being careful to distinguish between “before” and “meets”, and between “finishes” and “after”. Here’s the JSON with the examples: [...] I’ve created a JSON structure with 10 examples for each of the seven relations in Allen’s Interval Calculus. These examples are designed to be universally true and not just “often” or “sometimes” the case. I’ve been careful to distinguish between “before” and “meets”, as well as between “finishes” and “after”. For the “meets” relation, I’ve avoided mentioning the specific point in time where the events meet. Each relation has its own array of examples, and each example is structured with “event1”, “event2”, and an “example” sentence describing the relation between the events. Is there anything you’d like me to clarify or expand upon regarding these examples or Allen’s Interval Calculus?

5. Please do the same thing again and create 10 new examples for each relation. Do not repeat the ones you’ve created so far. Take care that the “equals” relation needs to be about temporal periods only.

 - Certainly, I’ll create 10 new examples for each relation, ensuring they are distinct from the previous set and that the “equals” relation focuses solely on temporal periods. I’ll maintain the same JSON structure and adhere to the guidelines you’ve provided. [...] I’ve created a new set of 10 examples for each of the seven relations in Allen’s Interval Calculus. These examples are distinct from the previous set and adhere to the guidelines you provided. For the “equals” relation, I’ve ensured that all examples focus solely on temporal periods. Each relation is represented by an array of examples, with each example containing “event1”, “event2”, and an “example” sentence describing the relation between the events. Is there anything you’d like me to clarify or expand upon regarding these new examples or Allen’s Interval Calculus?

6. For each of the relations, generate me a list of possible verbalizations by using placeholders for the two events. Format it as JSON like this: { “before”: [“{event1} happens before {event2}” ...], ... }

 - Certainly! I’ll create a list of possible verbalizations for each of Allen’s Interval Calculus relations, using placeholders for the two events. I’ll format this as JSON as you’ve requested. [...] I’ve created a JSON structure with 10 possible verbalizations for each of the seven relations in Allen’s Interval Calculus. Each verbalization uses placeholders event1 and event2 to represent the two events being related. These verbalizations provide various ways to express each temporal relationship, ranging from simple and direct statements to more elaborate descriptions. They can be useful for natural language processing tasks, generating descriptions of temporal relationships, or for explaining these concepts in different contexts. Is there anything you’d like me to clarify or expand upon regarding these verbalizations or Allen’s Interval Calculus?