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

Tai Mai
mai@cl.uni-heidelberg.de

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style file for ACL 2023. The document itself conforms to its own specifications, and is, therefore, an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

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 aren't directly comparable, they speak to the vast amounts 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 correctly 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 obtain 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 perplexity is likely 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 follows[1]

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

where X denotes the token sequence of tokens x, t is the number of tokens in the sequence, and $p_\theta(x_i|x_{<i})$ 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 a pseudo-perplexity metric, defined as follows:

$$\mathrm{PPPL}(X) := \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_{\mathrm{MLM}}(x_i|X_{\setminus x_i})\right\}. \tag{2}$$

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, 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 Algebra

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 verbalization of the relation *before* could be "e_1 *happens before* e_2." An event tuple such as $\langle$birth, 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."*

2.3 Conceptual neighborhood

Freksa (1992) arranges the relations proposed by Allen (1983) in an undirected graph structure to encode how closely two 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

[1] https://huggingface.co/docs/transformers/en/perplexity

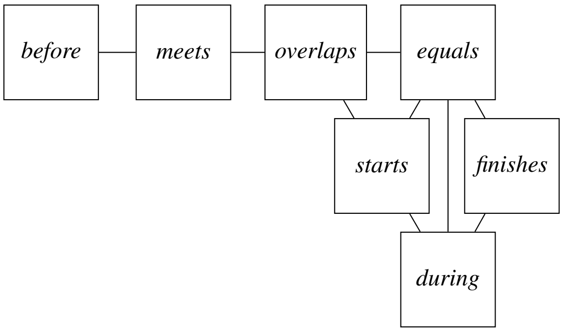

Figure 1: Conceptual neighborhood according to Freksa (1992), excluding the inverse relations.

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 correlates with the degrees of separation between the two relations.

3 Data

To test these two hypotheses, a simple dataset has been constructed by utilizing Claude 3.5 Sonnet[2] and then been curated by hand. For each of the seven temporal relations, a set of 20 event tuples are gathered which are examples of the temporal relation at hand. 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 ten possible verbalizations is gathered, with `{event1}` and `{event2}` as placeholders for respective temporal events, for example "`{event1}` *happens before* `{event2}`." Every event tuple is then placed into every verbalization of every relation, yielding a total dataset size of $20 \cdot 10 \cdot 7 = 140$ verbalized examples.

Before Claude 3.5 Sonnet was released, GPT 3.5 and GPT4o[3] were also considered for this task. 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 conversation prompts used to generate this dataset can be seen in Appendix A.

[2] https://www.anthropic.com/news/claude-3-5-sonnet

[3] https://chatgpt.com

4 Evaluation

4.1 Correlation analysis

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 Allonsius, 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-Yeary, 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, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly,

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 *raw 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.

Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, 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 Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, 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, Changhan 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, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, 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, Xiaojian 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.

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.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9733–9740.

A 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?