Are Human Conversations Special? A Large Language Model Perspective

Anonymous ACL submission

Abstract

In this paper, we study the changes in the attention behavior of large language models (LLMs) when used to understand natural conversations between humans (human-human conversations). By analyzing metrics such as attention distance, dispersion, and interdependency across these domains, we highlight the unique challenges posed to LLMs by conversational data. Our findings reveal that while language models exhibit domain-specific attention behaviors, there is a significant gap in their ability to specialize in human conversations. Through detailed attention entropy analysis and t-SNE visualizations, we demonstrate the need for models trained with diverse, high-quality conversational data to enhance understanding and generation of human-like dialogue.

1 Introduction

Understanding natural language is a cornerstone of artificial intelligence, with transformer-based large language models (LLMs) representing a significant leap forward in this effort (Vaswani et al., 2017; Minaee et al., 2024). These models have shown remarkable proficiency across a range of linguistic tasks, yet their performance varies widely across different types of data. Domain-specialized LLMs have shown greater effectiveness than general LLMs in various specialized settings such as code (Rozière et al., 2024; Li et al., 2023), math (Azerbayev et al., 2024), finance (Wu et al., 2023), and medicine (Labrak et al., 2024; Nazi and Peng, 2023). However, there has been less focus on natural human-human conversations, which embody a rich collection of nuances, contexts, and unspoken cues (Tur and Hakkani-Tür, 2011). We perform a comprehensive analysis of how transformer-based LLMs – embodied in this work by the LLaMa-2 13b (Touvron et al., 2023b) model – process and interpret human conversations in relation to other data such as web content, code, and mathematical texts.

Formal "textbook" conversations – such as those taught in classroom settings to analyze conversational structures – do not typically exhibit the same characteristics as speakers engaged in speaking and communicating naturally (Rings, 1986). Spoken conversations are spontaneous; and to operate effectively in conversations, the knowledge of the participating entity has to stretch far beyond mere awareness of sounds and words. As a result of years of evolution and social environments where the use of language in conversation is practiced daily, humans can structure and build conversations appropriate to any situation without much formal training, and adapt to changing norms with time (Pridham, 2013). These emergent traits are not prevalent or immediately apparent in documents or articles which constitute a large portion of web data; or in other domain-specific corpora like code.

In this work, we begin by analyzing the proportion of authentic human-human conversations in the web data used to (pre)train state-of-the-art LLMs. Our analysis finds that authentic human conversations are rare in occurrence on the web, and the vast majority of “conversation data” merely refers to textbook conversations. Our investigation centers on three key aspects: attention distance, dispersion, and interdependency within different data domains. Through quantitative analysis of attention entropy and qualitative inspections of attention patterns, we seek to understand the intricacies of model behavior across domains. We also employ t-SNE visualizations to compare the hidden state representations of language models when exposed to different types of data, allowing us to visually assess how domain-specific characteristics are encoded within models, offering insights into their ability to distinguish and adapt to varied linguistic challenges.

2 Human-Human Conversations

The majority of human-human conversations are conducted in spoken language rather than via written texts. Natural human conversation is an interactive exchange between two or more people, with a format that can be one-on-one or between multiple people. Examples of such interactions include chats between family or friends, at work, or in the public domain; and can be conducted either face-to-face or virtually. Conversations, however, are far more than just the words that they are made up of (Pridham, 2013). The textual representation of a spoken conversation misses significant information from the speech and visual channels/modalities. Speech contains information about the speaker in terms of their emotions, intelligence, age, psychological traits, etc. (Spirina et al., 2016). The combination of information in visual and speech channels is manifested through body language and gestures and their intensities; and prosodic features such as speed, intonation, speed, amplitude, silence, and laughter. However, the textual representation of spoken language contributes primarily to the meaning and knowledge of the thought in the exchange, while indirectly modeling subtle cues from the speech and visual modalities. Understanding conversation in its complete sense requires understanding the purpose behind the words and the situational, emotional, social, and contextual understanding established in the conversation and their evolution until a specific point in the conversation (Pridham, 2013).

2.1 Characteristics of Human Conversations

Human-human conversations are distinguished by several key characteristics:

Interactivity: Unlike static web data, human conversations are highly interactive, with participants actively responding to and building upon each other's contributions. This interactivity involves turn-taking, feedback signals (e.g., nodding, "uh-huh"), and adjustments in discourse based on the other participants' responses.

Contextuality: Conversations are deeply embedded in specific contexts, which include physical surroundings, social relationships, cultural backgrounds, and the participants' shared history. This context influences not only the content but also the form of the conversation, including language choice, tone, and register. In contrast, domains like code or mathematics are characterized by a high level of abstraction, process, and standardization, where context plays a minimal role in the interpretation of the data.

Adaptability: Participants in a conversation continually adjust their speech based on immediate feedback from their interlocutors. This adaptability covers a wide range of aspects, from changing topics smoothly to modifying speech patterns for clarity or emphasis. Such dynamic adjustments are specific to human interactions and are not found in structured data domains like code, where the syntax and semantics follow rigid, predefined rules.

Emotional and Psychological Dimensions: Conversations convey not just factual information but also emotional and psychological states. Through tone, pace, volume, and choice of words, speakers can express a wide range of emotions and attitudes. These nuanced emotional layers add depth to human conversations that are typically absent in other data domains, where emotional expressiveness is either irrelevant or vastly simplified.

2.2 Human Conversation Data on the Web

The majority of the content on the internet is in the form of articles, documents, blogs, and forums where information is structured. Authentic human conversations are drastically less in proportion to written content in the web data. It has been challenging to find authentic human-human conversation data publicly that can be used for training models due to copyright, privacy, and intellectual property concerns. We analyze the web data from CommonCrawl (Common Crawl, 2023) dumps for human conversation data and the types of conversations in Table 1. We randomly sample a subset of the dump and deduplicate it so that it can be used to approximate the data distribution between human conversations versus the rest of the data. We fine-tune a BERT (Devlin et al., 2018) model for document classification using $\sim$194K samples containing human conversations and non-conversational documents in equal amounts. We find that natural human conversations are rare in the web domain: even accounting for the upper-bound $+0.0043\%$ error from Table 1, such conversations only account for a maximum of $\approx 0.0128\%$ of the total data.

3 Related Work

Recent studies have shown the intricate ways in which various models, including Transformers and

Type	Percentage	Err.
Written/Documents	99.9915%	$\pm 0.0043\%$
Human Conversations	0.00849%	$\pm 0.0043\%$

Table 1: Distribution between Human Conversations and Written/Document data in CommonCrawl.

recurrent neural networks, encode dependency relations within texts (Hewitt and Manning, 2019; Raganato and Tiedemann, 2018). Transformer models have been found to most effectively capture dependency relations within their middle layers (Liu et al., 2019).

Analysis of the attention distance within decoder-only transformer models (Vig and Belinkov, 2019) has provided evidence supporting the hypothesis that deeper layers capture longer-distance relationships. This is a measurement of the mean distance spanned by attention for each head; and is calculated as the average distance between token pairs in all samples in the dataset, weighted by attention between the tokens:

$$\overline{D}_\alpha = \frac{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{i} \alpha_{i,j}(x) \cdot (i-j)}{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{i} \alpha_{i,j}(x)} \quad (1)$$

The exploration of attention dispersion and entropy as measures of how attention is distributed across tokens offers additional insights into the mechanisms through which models understand and process patterns in language:

$$\text{Entropy}_\alpha(x_i) = -\sum_{j=1}^{i} \alpha_{i,j}(x) \log(\alpha_{i,j}(x)) \quad (2)$$

This body of work sets a context for our investigation into the unique characteristics of human-human conversations, comparing these dynamics against the backdrop of general web corpora, including articles, blogs, forums, and specialized domains such as mathematics and programming. Understanding the nuances of how models encode dependency relations and manage attention across different types of text is crucial in distinguishing the specifics of human conversational patterns.

Finally, the analysis conducted in this paper is similar in spirit to the work in the mutlilingual (large) language model (MLLM) community on the effect of using models trained on higher resource languages and datasets with data from lower resource settings. In one effort (Joshi et al., 2020), the authors identify the lack of linguistic diversity when training models – similar to the lack of diversity in data type when training LLMs, which is the focus of our present study; while in another (Rust et al., 2021), a detailed empirical analysis is provided to show the differences between different languages. We take inspiration from these efforts for our attention-centric study of language models and the content used to train them.

4 Analysis

4.1 Attention Distance Difference

Analyzing the difference between the attention distances as defined in Equation (1) can provide a better way to gain insights into how language models form relationships, especially longer-distance relationships in deeper layers by focusing on the difference between attention distances.

Given two domains, D_1 and D_2, with their respective sets of texts X^{D_1} and X^{D_2}, the attention distance for each domain is calculated as:

$$\overline{D}_\alpha^{D_k} = \frac{\sum_{x \in X^{D_k}} \sum_{i=1}^{|x|} \sum_{j=1}^{i} \alpha_{i,j}(x) \cdot (i-j)}{\sum_{x \in X^{D_k}} \sum_{i=1}^{|x|} \sum_{j=1}^{i} \alpha_{i,j}(x)} \quad (3)$$

for $k = 1, 2$, where $\alpha_{i,j}(x)$ is the attention weight from token i to token j in text x, and $|x|$ is the length of text x.

The difference in attention distance between the two domains can then be defined as:

$$\Delta\overline{D}_\alpha = \overline{D}_\alpha^{D_1} - \overline{D}_\alpha^{D_2} \quad (4)$$

This measure, $\Delta\overline{D}_\alpha$, quantifies the difference in how attention spans across tokens vary between the two domains, providing insights into the structural differences in how information is processed and dependencies are captured in texts from D_1 compared to D_2.

By analyzing $\Delta\overline{D}_\alpha$ we can find insights into domain specificity in transformer models by understanding how transformer models adapt their attention mechanism to structural and contextual differences between various domains. We can also identify if models tend to focus on closer or more distant token relationships when dealing with human-human conversations as opposed to more structured and document-oriented content. Positive values of difference in attention distance indicate that the attention distance in the second domain is longer than the first domain, whereas negative values indicate that it is shorter. Positive differences in the middle and end layers indicate more complex relationships requiring longer dependencies, and positive differences in the initial layers indicate longer syntactic and semantic relationships in the sequence tokens.

4.2 Attention Dispersion

We also calculate the entropy of the attention distribution based on Equation (2) to measure the attention dispersion. This provides insights into how domain-specific characteristics and the model's training influence its learning and processing strategies. High entropy is not always desirable, as it is indicative of a lack of focus or understanding. Similarly, very low entropy might suggest overfitting to specific tokens or phrases, potentially reducing the model's ability to generalize across varied inputs within the domain. We perform a comparison of attention dispersion between domains to understand the robustness of the model's understanding of a domain.

High Entropy Domain: A higher entropy in the attention distribution of a domain means that the model finds the information in that domain more uniformly informative or relevant, without specific tokens or phrases standing out as significantly more important than others. This suggests that the domain is more complex or less familiar to the model, leading it to distribute its attention more evenly rather than clearly identifying key information. This would also indicate more variety and ambiguity in how information is presented, requiring broader focus to capture the necessary context for understanding.

Low Entropy Domain: Domains with lower entropy in the attention distribution indicate that the model is focusing its attention more narrowly on specific tokens. This suggests that the model has learned to identify key tokens or phrases that are particularly informative or relevant for understanding or performing tasks in such a domain. The domain is more structured or contains clearer cues that the model can exploit to make predictions or understand content. It also reflects a higher level of familiarity or specialization of the model in this domain, allowing it to more effectively pinpoint the most relevant information.

4.3 Interdependency Analysis

Analyzing interdependencies between various aspects of text from different domains provides insights into underlying structures, patterns, and dynamics of information in domains such as communication, written documents, and code. We devise a novel metric – the *Interdependency Factor* (IF) – to quantify the degree of interdependency between various *aspects* of the data, as long as they can be modeled in a graph as nodes along with directed edges between them. This is intended to indicate the overall complexity of the domain along specific aspects. *Aspect* in this context refers to any derived or absolute representation in a sequence. In this analysis, we use a tokenized representation of text in the domain. However, it can be useful to use higher-level segmentation such as themes that can be common across different domains. When this is not possible or the dependency is modeled at more granular levels by systems such as language models, a lower-level representation (such as tokens) can be used, where the weights on the edges are attention values (Vig and Belinkov, 2019) between the tokens modeled by the transformer (Vaswani et al., 2017) language model layers.

The Interdependency Factor (IF) is defined as follows: given a dataset of text samples, a set N represents all identified node candidate labels in the graph, where each node $n_i \in N$ represents a distinct *aspect* value. To analyze the interdependencies among these nodes, we construct a directed graph $G = (V, E)$, where V corresponds to the set of vertices, with each vertex representing a node in N, and E represents the set of directed edges between these vertices. Each edge $(n_i, n_j) \in E$ is associated with a weight w_{ij}, quantifying the strength or frequency of the transition or relationship from node n_i to node n_j.

The adjacency matrix A of graph G is defined such that each element a_{ij} within A corresponds to the weight w_{ij} of the edge from n_i to n_j. The Interdependency Factor IF is then defined as follows:

$$IF = \frac{1}{|N|^2 - |N|} \sum_{i=1}^{|N|} \sum_{j=1, j \neq i}^{|N|} a_{ij} \qquad (5)$$

This calculates the IF by averaging the weights of all directed edges in the graph, excluding self-transitions (where $i = j$). The normalization factor, $|N|^2 - |N|$, accounts for the total number of possible directed transitions between different nodes, ensuring that the IF remains a relevant measure of interdependency across datasets of varying sizes and complexities. In cases where the weights are not available or not computable, 0 and 1 should be used to indicate the absence and existence of a dependency between two aspects.

5 Experimental Setup

5.1 Domain Datasets

In our analysis, we focus primarily on English data across the domains of human-human conversation, web, and math. Code data is randomly sampled across a variety of programming languages. We use 1000 samples from each domain in our analysis.

Human-Human Conversations: In our study, a wide range of real-life natural conversations between humans across various business and casual settings is used. Scripted conversations such as movie scripts, and single-person presentations or talks are not included. Key aspects such as context dependencies, emotional expressiveness, idiomatic usage, and integration of general and localized or private knowledge are the focus. The data used for human conversations is a set of real conversations between people, processed using a conversation intelligence platform (omitted for blind review), and anonymized by replacing PII and PCI information with synthetic data.

Web Data: Data from the internet containing various types of content such as blog posts, news articles, forums, social media content, etc. is generated using a randomly sampled subset from the CommonCrawl (Common Crawl, 2023). We perform a preliminary data cleanup to remove unnecessary HTML tags, and deduplicate to ensure the entries are unique.

Code: Source code from various programming languages, each with unique syntax and semantics, is used. The focus is on the structure and logic expressed in code, which contrasts with the unstructured and mostly informal nature of human-human conversations. The code data is curated from the GitHub dataset (Codeparrot, 2022).

Mathematics: Mathematical expressions, problems, and proofs across different fields of mathematics make up this domain – derived from the Proof Pile 2 corpus (Azerbayev et al., 2023) – highlighting the abstract, precise, and symbolic characteristics of mathematical communication.

5.2 Language Model

We use a pretrained decoder-only transformer language model – LLaMa-2 13b (Touvron et al., 2023b) – for analyzing attention patterns and embeddings at various layers and heads. This model's architecture contains 40 layers and 40 attention heads. Although exact details of the LLaMa-2 model are not indicated in the accompanying technical report (Touvron et al., 2023b), the model was trained on data that is similar to the LLaMa-1 models (Touvron et al., 2023a). This enables our assumption that the model was trained on approximately 82% of data from web dumps from CommonCrawl and C4, 4.5% of data from the code domain from GitHub, and 2.5% of data from ArXiv, which consists of scientific data with some overlap with math. Apart from the data in the web corpus, the rest of the data is distributed between Wikipedia, books, and StackExchange corpora. After adjusting for the web corpus distribution based on our earlier analysis (c.f. Section 2.2), the pretraining data of the model is expected to have between $\approx 0.0069618\%$ and $\approx 0.010496\%$ of human-human conversations.

6 Results

6.1 Attention Distance Difference Analysis

We calculate the mean difference in attention distances $\Delta\overline{D}_\alpha$ (c.f. Equation (4)) for each of the human-human conversations, code, and math domains; and compare each of these in turn against general web data.

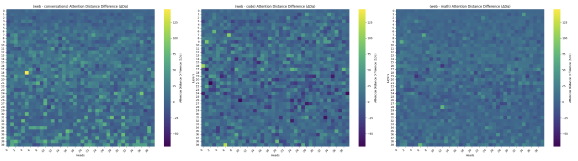


Figure 1: Heatmap of the Attention Distance Difference matrix ($\Delta\overline{D}_\alpha$) calculated for web data against human-human conversations, code, and math.

Figure 1 shows heatmaps of the Attention Distance Difference by layer (Y-axis) and head (X-axis), with one domain fixed as human-human conversations, code, and math respectively; and the other domain as general data from the web. We find significant differences in attention distances in deeper layers when comparing human-human conversations to web data. Higher values in these layers indicate that human conversations necessitate more robust modeling of long-term contextual relationships than general web corpora. This is consistent with the nature of human dialogue, where the flow of information often spans across several exchanges, requiring the model to maintain context over extended sequences. The comparison with code displays a distinctive pattern where higher attention distances are observed in the initial half of the layers, suggesting that models capture structural dependencies effectively in these stages. However, as we progress into deeper layers, there is a reduction in these values, which suggests that attention becomes more localized, focusing on closer contextual relationships. This reflects the structural and syntactic rigidity inherent in programming languages, where local context is often sufficient for understanding many dependencies. Mathematical texts exhibit a relatively even distribution of attention distances across layers when compared to web data. This implies that mathematical texts, with their symbolic and formulaic nature, require a balanced approach where both local and long-distance relationships are equally pertinent across all layers of the model.

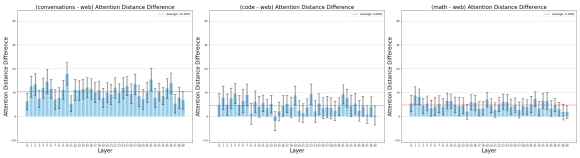


Figure 2: Attention Distance Difference by Layer across all heads calculated for web data against human-human conversations (left), code (middle), and math (right).

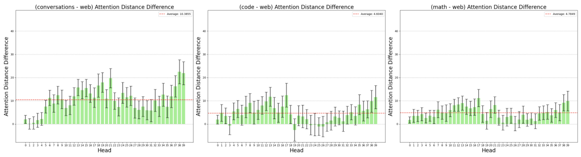


Figure 3: Attention Difference by Head across all layers calculated for web data against human-human conversations (left), code (middle), and math (right).

D_1	D_2	$\Delta\overline{D}_\alpha$
Human Conversations	Web	10.3855
Code	Web	4.6040
Math	Web	4.7849

Table 2: Average attention distance difference between human-human conversations, code, and math domains with web data. A higher value indicates longer contextual dependencies.

Almost all the layers exhibit approximately equal differences in attention, with lower differences manifesting in the final layers (which are typically optimized for generation) as seen in Figure 2. Differences in the initial layers are typical across all domains, as syntactical and semantic modeling representations of the model are different across domains. However, more complex relationships are modeled in the middle layers, where we see significantly higher differences for conversation-web, as compared against code-web and math-web pairs. When compared by individual head in Figure 3, the initial heads show very little difference from the web domain; but the middle heads and heads towards the end exhibit significant deviation from the web domain. These differences in the middle heads are less pronounced in the code and math domains, which indicates that human-human conversation domain modeling tends to have higher attention distances across most heads when compared to code and math domains.

6.2 Attention Dispersion Analysis

To study the dispersion of attention across domain data, we calculate the mean attention entropy (c.f. Equation (2)) and analyze it by layer/head (Figure 4), as well as by layer alone (Figure 5) and head alone (Figure 6) across all four domains considered in this study: general web data, human-human conversations, code, and math.

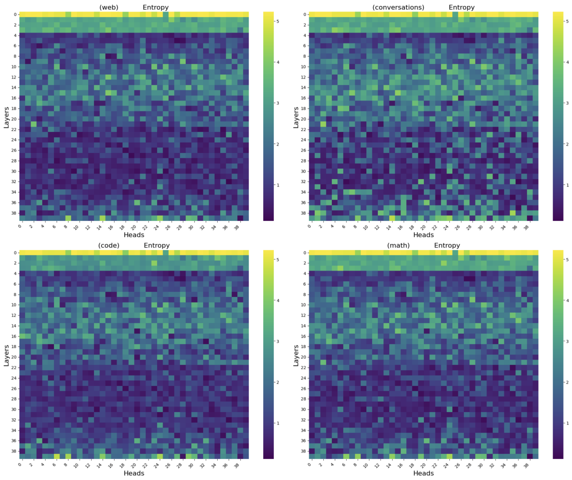


Figure 4: Heatmap of mean attention entropy for web, human-human conversations, code, and math domains respectively.

In Figure 4, the heatmaps represent the entropy by layer/head for web, human-human conversations, code, and math domains. Attention dispersion is highest in the human-human conversations domain. This is consistent with the attention distance difference plot in Figure 1. From layer 22 to layer 36, entropy is typically lower for web, code, and math domains; however, the entropy is high in multiple heads in these layers for human-human conversations. In the conversation domain, for each token, the model has to attend strongly to more tokens than in the rest of the domains – this indicates higher complexity for the domain, which leads to higher attention dispersion in the model while understanding that domain. It also suggests that the model is less familiar with the human-human conversation domain, which can be explained by the scarcity of training data in the domain distribution (c.f. Section 2.2). For web, code, and math domains in comparison, the entropy is noticeably lower. This indicates that the model has a more robust understanding of these domains, and the model can find an optimal attention strategy, reducing attention dispersion, which can be explained by the considerable amount of data from these domains that is reflected in the model's pretraining data (c.f. Section 5.2).

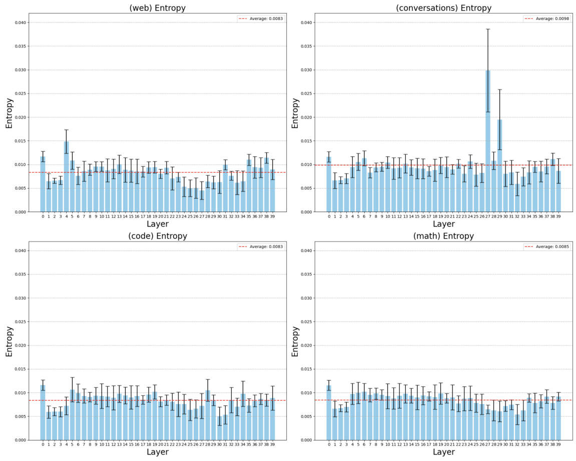


Figure 5: Mean attention entropy by layer across all heads with first token attention removed for web, human-human conversations, code, and math domains respectively. Higher values indicate more attention diffusion in the layer.

We also plot the entropy by removing attention to the first token in sequence, by layers and heads separately, shown in Figure 5 and Figure 6 respectively. We remove the first token's entropy because we find that the model adds redundant attention to the first token which leads to high entropy, especially in the first layer. As shown in Figure 5, certain layers – specifically layers 27 and 29, which are mid-layers of the model – have significantly higher mean attention entropy compared to the rest of the layers for the human-human conversation domain. A similar pattern can be seen in Figure 6, where heads 13, 25, and 38 have high entropy. For web, code, and math domains, such high entropy is not exhibited by the model, indicating that the model has less familiarity with complex relationships in the human-human conversation domain as compared to others.

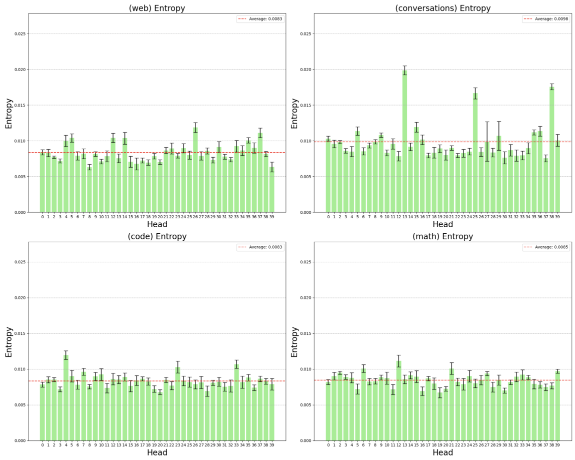


Figure 6: Mean attention entropy by the head across all layers with first token attention removed for web, human-human conversations, code, and math domains respectively. Higher values indicate more attention diffusion in the head.

6.3 Attention Interdependency Analysis

We perform interdependency analysis between human-human conversations and other domains to gain deeper insight into the underlying structures that a model needs across these domains. For human-human conversations and web data, we perform analysis using theme segmentation and dependencies between themes within a conversation or document, as well as a token-level analysis. We perform only token-level interdependency analysis for code and math data, as thematic analysis on these domains does not provide much insight due to their logical and rule-driven nature. We calculate the average attention matrix across all samples in the domain averaged across all attention heads from the middle layer. We calculate IF for these domains to understand the overall interdependency between tokens, with 512 tokens in each sequence to get a quantitative evaluation of the interdependency by domain. This is shown in Table 3.

To further analyze the overall attention at each token, the individual token's weights can be calculated by aggregating the attention weights of the token towards the rest of the tokens as shown in Figure 7. This provides us insights into how attention patterns change across the text sequence by domain. Fluctuations in weight by token index signify frequent changes in overall attention strength, and

Domain	OAE	IF
Web	0.0083	100.207
Human Conversation	0.0098	141.869
Code	0.0083	106.466
Math	0.0085	110.848

Table 3: Column 2: Overall attention entropy (OAE) averaged across heads and layers by domain. Column 3: Interdependency Factor (IF) by domain, calculated for $N = 512$ across all samples.

the value of weight indicates the overall strength of attention, which is an aggregation of attention values of the token attending to all other tokens. Note that due to aggregation, the plot in Figure 7 no longer provides us with information about the global or local interdependencies. Human-human conversations require higher and longer attention resulting in higher average weight as compared to the rest of the domains.

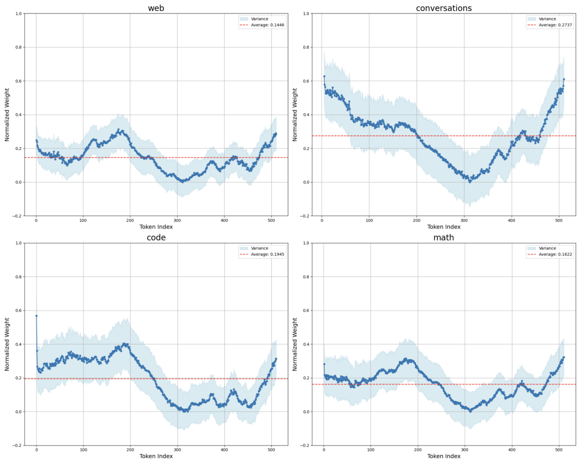


Figure 7: Normalized attention weights of interdependencies aggregated by the token for web, human-human conversations, code, and math domains respectively.

6.4 Language Model Representation

To understand the representation of language models by domains, we use t-SNE (Van der Maaten and Hinton, 2008) to visualize and compare the hidden state representations of the first, middle, and last layers of the LLaMa-2 13b model (Touvron et al., 2023a). Early layers in language models learn the syntactic and semantic relationships in the sequences, and in deeper layers complex relationships are modeled capturing abstract and higher level understanding (Jawahar et al., 2019; Hao et al., 2021). In Figure 8, the representation of the first layer across domains is relatively close, with clear boundaries in the clusters, given that the semantic features of most of the language-based domains are mostly similar. The middle layer representation shows some overlap in web and code data, but a clear and quite significant distance between human-human conversations and math data. We continue to see a similar pattern in the last layer, with slightly better separation in web and code while conversations and math data continue to be distant. This shows that for a language model trained on a general corpus, containing data from various domains, the domain-specific learnings converge differently.

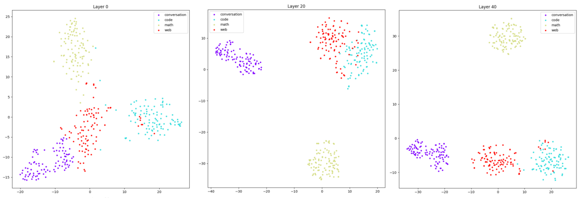


Figure 8: t-SNE plot of the first, middle, and last layers of the LLaMa-2 13b model by domains.

7 Conclusion

In this study, we examined how transformer-based language models (using LLaMa-2 13b as a representative) process natural human conversations in contrast to web content, code, and mathematical texts. We found a general lack of sufficient representation of human-human conversations in web data, which is the largest constituent of pretraining data in most current large language models. Our findings highlight that human-human conversational data challenges a(ny) model into managing long-term contextual relationships and dependencies across layers. Our analysis motivates the importance of domain specialization in language models to enhance their understanding and handling of human conversations; and indicates that training language models with a vast amount of high-quality authentic human conversations is an essential requirement in bridging the gap in model performance.

8 Limitations

The major limitation of the work presented here revolves around the number of models used in our analysis. We used only on a single model – LLaMa-2 13b (Touvron et al., 2023b). There are a number of potential shortcomings that arise from this. First, while we expect that this model is representative of most popularly used large language models today and shares the same fundamental architecture, we have no way of being completely certain that this is the case. Second, since there is no reliable public information about the exact data that was used to train this open source model, we were forced to make certain assumptions around this (see Section 5.2).

Another limitation centers around the size of the datasets that were used to evaluate attention in the model chosen. In Section 5.1, we report the usage of 1000 data samples per domain. Which these data samples were chosen in as representative a manner as possible – taking care to choose from various programming languages for code, picking across different kinds of internet content for web, etc. – there is no clear way of ensuring that these are indeed representative of these domains at large.

Finally, we acknowledge the limitation that while much of our initial motivation is centered around the rich, multimodal nature of human-human conversations (which can include text, audio/speech, and visual modalities), in this particular paper, we are only able to analyse the text based facets of this domain. Some of this can be explained by the relative paucity of widely available multimodal large language models, particularly on the audio side. However, there have been rapid advancements in this area as of the time of submitting this paper, and it presents a very promising and achievable avenue for future work.

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *Preprint*, arXiv:2310.10631.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*.

Codeparrot. 2022. Codeparrot - github code dataset. https://huggingface.co/datasets/codeparrot/github-code.

Common Crawl. 2023. Commoncrawl - get started. https://commoncrawl.org/get-started.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you! *Preprint*, arXiv:2305.06161.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint 2402.06196*.

Zabir Al Nazi and Wei Peng. 2023. Large language models in healthcare and medical domain: A review. *Preprint*, arXiv:2401.06775.

Francesca Pridham. 2013. *The language of conversation*. Routledge.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Lana Rings. 1986. Authentic language and authentic conversational texts. *Foreign Language Annals*, 19(3):203–208.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

A. V. Spirina, M. Yu. Sidorov, R. B. Sergienko, E. S. Semenkin, and W. Minker. 2016. Human-human task-oriented conversations corpus for interaction quality modeling. *Siberian Aerospace Journal*, 17(1):84–90.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Gokhan Tur and Dilek Hakkani-Tür. 2011. Human/human conversation understanding. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 225–255.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

A Appendix

A.1 Attention Including First Token

As accompaniments to Figure 5 and Figure 6 in the main body of the paper, we also provide the corresponding plots without removing attention to the first token. These are are shown in Figure 9 and Figure 10 respectively.

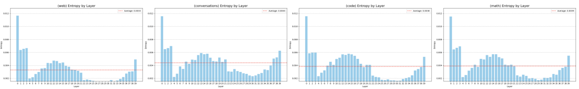


Figure 9: Mean attention entropy by layer without removing first token attention for web, human-human conversations, code, and math domains respectively.

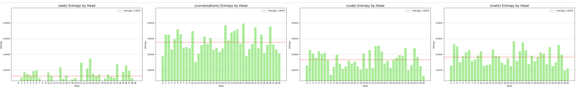
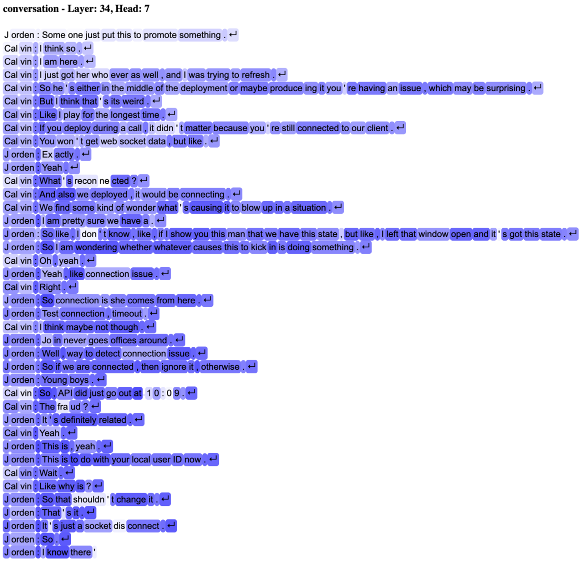


Figure 10: Mean attention entropy by head without removing first token attention for web, human-human conversations, code, and math domains respectively.

A.2 Qualitative Analysis

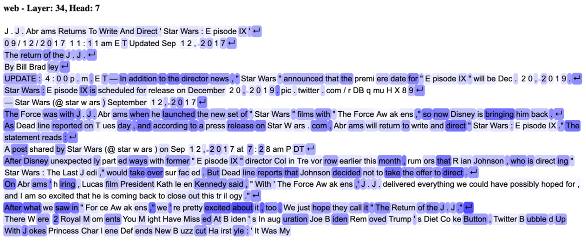
To get an intuitive sense of patterns exhibited by specific layers and attention heads, we used a few examples from each domain to visually inspect the attention dispersion at each token in the example. Examples for each domain showing the attention entropy at each token in the example are shown in Figures 11, 12, 13, and 14. Several heads across layers show similar attention dispersion across all domains. However, certain heads show higher attention dispersion in human-human conversations as compared to other domains. We find that the



```
conversation - Layer: 34, Head: 7

J orden : Some one just put this to promote something . ↵
Cal vin : I think so . ↵
Cal vin : I am here . ↵
Cal vin : I just got her who ever as well , and I was trying to refresh . ↵
Cal vin : So he ' s either in the middle of the deployment or maybe produce ing it you ' re having an issue , which may be surprising . ↵
Cal vin : But I think that ' s its weird . ↵
Cal vin : Like I play for the longest time . ↵
Cal vin : If you deploy during a call , it didn ' t matter because you ' re still connected to our client . ↵
Cal vin : You won ' t get web socket data , but like . ↵
J orden : Ex actly . ↵
J orden : Yeah . ↵
Cal vin : What ' s recon ne cted ? ↵
Cal vin : And also we deployed , it would be connecting . ↵
Cal vin : We find some kind of wonder what ' s causing it to blow up in a situation . ↵
J orden : I am pretty sure we have a . ↵
J orden : So like , I don ' t know , like , if I show you this man that we have this state , but like , I left that window open and it ' s got this state . ↵
J orden : So I am wondering whether whatever causes this to kick in is doing something . ↵
Cal vin : Oh , yeah . ↵
J orden : Yeah , like connection issue . ↵
Cal vin : Right . ↵
J orden : So connection is she comes from here . ↵
J orden : Test connection , timeout . ↵
Cal vin : I think maybe not though . ↵
J orden : Jo in never goes offices around . ↵
J orden : Well , way to detect connection issue . ↵
J orden : So if we are connected , then ignore it , otherwise . ↵
J orden : Young boys . ↵
Cal vin : So , API did just go out at 1 0 : 0 9 . ↵
Cal vin : The fra ud ? ↵
J orden : It ' s definitely related . ↵
Cal vin : Yeah . ↵
J orden : This is , yeah . ↵
J orden : This is to do with your local user ID now . ↵
Cal vin : Wait . ↵
Cal vin : Like why is ? ↵
J orden : So that shouldn ' t change it . ↵
J orden : That ' s it . ↵
J orden : It ' s just a socket dis connect . ↵
J orden : So . ↵
J orden : I know there '
```

Figure 11: Human-Human Conversation example highlighted for mean attention entropy at each token for layer 34, head 7, showing the high attention diffusion compared to web, code, and math examples.

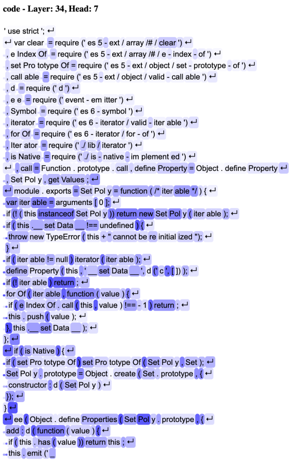


```
web - Layer: 34, Head: 7

J . J . Abr ams Returns To Write And Direct ' Star Wars : E pisode IX ' ↵
0 9 / 1 2 / 2 0 1 7 1 1 : 1 1 am E T Updated Sep 1 2 , 2 0 1 7 ↵
The return of the J . J . ↵
By Bill Brad ley ↵
UPDATE : 4 : 0 0 p . m . E T — In addition to the director news , " Star Wars " announced that the premi ere date for " E pisode IX " will be Dec . 2 0 , 2 0 1 9 . ↵
Star Wars : E pisode IX is scheduled for release on December 2 0 , 2 0 1 9 . pic . twitter . com / r DB q mu H X 8 9 ↵
— Star Wars (@ star w ars ) September 1 2 , 2 0 1 7 ↵
The Force was with J . J . Abr ams when he launched the new set of " Star Wars " films with " The Force Aw ak ens ," so now Disney is bringing him back . ↵
As Dead line reported on T ues day , and according to a press release on Star W ars . com , Abr ams will return to write and direct " Star Wars : E pisode IX ." The
statement reads : ↵
A post shared by Star Wars (@ star w ars ) on Sep 1 2 , 2 0 1 7 at 7 : 2 8 am P DT ↵
After Disney unexpected ly part ed ways with former " E pisode IX " director Col in Tre vor row earlier this month , rum ors that R ian Johnson , who is direct ing "
Star Wars : The Last J edi ," would take over sur fac ed . But Dead line reports that Johnson decided not to take the offer to direct . ↵
On Abr ams ' h iring , Lucas film President Kath le en Kennedy said , " With ' The Force Aw ak ens ,' J . J . delivered everything we could have possibly hoped for ,
and I am so excited that he is coming back to close out this tr il ogy ." ↵
After what we saw in " For ce Aw ak ens ," we ' re pretty excited about it , too . We just hope they call it " The Return of the J . J ." ↵
There W ere 2 Royal M om ents You M ight Have Miss ed At B iden ' s In aug uration Joe B iden Rem oved Trump ' s Diet Co ke Button , Twitter B ubble d Up
With J okes Princess Char l ene Def ends New B uzz cut Ha irst yle : ' It Was My
```

Figure 12: Web data example highlighted for mean attention entropy at each token for layer 34, head 7. Attention diffusion is significantly lower as compared to the human-human conversation as shown in Figure 11.

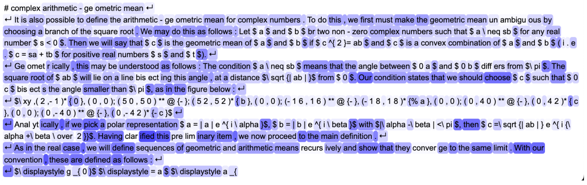
initial layers have high entropy across all domains, whereas the middle and last layers have relatively high entropy in human-human conversations, as compared to web, code, and math domains also indicated by the attention dispersion analysis results.



```
code - Layer: 34, Head: 7

' use strict '; ↵
↵ var clear = require (' es 5 - ext / array /# / clear ') ↵
, e Index Of = require (' es 5 - ext / array /# / e - index - of ') ↵
, set Pro totype Of = require (' es 5 - ext / object / set - prototype - of ') ↵
, call able = require (' es 5 - ext / object / valid - call able ') ↵
, d = require (' d ') ↵
, e e = require (' event - em itter ') ↵
, Symbol = require (' es 6 - symbol ') ↵
, iterator = require (' es 6 - iterator / valid - iter able ') ↵
, for Of = require (' es 6 - iterator / for - of ') ↵
, Iter ator = require (' ./ lib / iterator ') ↵
, is Native = require (' ./ is - native - im plement ed ') ↵
↵ , call = Function . prototype . call , define Property = Object . define Property ↵
, Set Pol y , get Values ; ↵
↵ module . exports = Set Pol y = function ( /* iter able */ ) { ↵
var iter able = arguments [ 0 ]; ↵
if (! ( this instanceof Set Pol y )) return new Set Pol y ( iter able ); ↵
if ( this .__ set Data __ !== undefined ) { ↵
throw new TypeError ( this + " cannot be re initial ized "); ↵
} ↵
if ( iter able != null ) iterator ( iter able ); ↵
define Property ( this , ' __ set Data __ ', d (' c ', [ ]) ); ↵
if (! iter able ) return ; ↵
for Of ( iter able , function ( value ) { ↵
if ( e Index Of . call ( this , value ) !== - 1 ) return ; ↵
this . push ( value ); ↵
}, this .__ set Data __ ); ↵
}; ↵
↵ if ( is Native ) { ↵
if ( set Pro totype Of ) set Pro totype Of ( Set Pol y , Set ); ↵
Set Pol y . prototype = Object . create ( Set . prototype , { ↵
constructor : d ( Set Pol y ) ↵
}); ↵
} ↵
↵ ee ( Object . define Properties ( Set Pol y . prototype , { ↵
add : d ( function ( value ) { ↵
if ( this . has ( value )) return this ; ↵
this . emit (' _
```

Figure 13: Code example highlighted for mean attention entropy at each token for layer 34, head 7. Attention diffusion is significantly lower as compared to the human-human conversation as shown in Figure 11 and equivalent to the web example in Figure 12.



```
# complex arithmetic - ge ometric mean ↵
↵ It is also possible to define the arithmetic - ge ometric mean for complex numbers . To do this , we first must make the geometric mean un ambigu ous by
choosing a branch of the square root . We may do this as follows : Let $ a $ and $ b $ br two non - zero complex numbers such that $ a \ neq sb $ for any real
number $ s < 0 $. Then we will say that $ c $ is the geometric mean of $ a $ and $ b $ if $ c ^{ 2 }= ab $ and $ c $ is a convex combination of $ a $ and $ b $ ( i . e
. $ c = sa + tb $ for positive real numbers $ s $ and $ t $). ↵
↵ Ge omet r ically , this may be understood as follows : The condition $ a \ neq sb $ means that the angle between $ 0 a $ and $ 0 b $ diff ers from $\ pi $. The
square root of $ ab $ will lie on a line bis ect ing this angle , at a distance $\ sqrt {| ab | }$ from $ 0 $. Our condition states that we should choose $ c $ such that $ 0
c $ bis ect s the angle smaller than $\ pi $, as in the figure below : ↵
↵ $\ xy ,( 2 ,- 1 )* { 0 }, ( 0 , 0 ); ( 5 0 , 5 0 ) ** @ {- }; ( 5 2 , 5 2 )* { b }, ( 0 , 0 ); (- 1 6 , 1 6 ) ** @ {- }, (- 1 8 , 1 8 )* {% a }, ( 0 , 0 ); ( 0 , 4 0 ) ** @ {- }, ( 0 , 4 2 )* { c
}, ( 0 , 0 ); ( 0 ,- 4 0 ) ** @ {- }, ( 0 ,- 4 2 )* {- c }$ ↵
↵ Anal yt ically , if we pick a polar representation $ a = | a | e ^{ i \ alpha }$, $ b = | b | e ^{ i \ beta }$ with $|\ alpha -\ beta | <\ pi $, then $ c =\ sqrt {| ab | } e ^{ i {\
alpha +\ beta \ over 2 }}$. Having clar ified this pre lim inary item , we now proceed to the main definition . ↵
↵ As in the real case , we will define sequences of geometric and arithmetic means recurs ively and show that they conver ge to the same limit . With our
convention , these are defined as follows : ↵
↵ $\ displaystyle g _{ 0 }$ $\ displaystyle = a $ $\ displaystyle a _{
```

Figure 14: Math example highlighted for mean attention entropy at each token for layer 34, head 7. Attention diffusion is significantly lower as compared to the human-human conversation as shown in Figure 11 and equivalent to the web example in Figure 12.