Relations of Linguistic Features to Guideline-Based Rankings of Medical Answers are Nontrivial

Davis Bartels, Brandon C. Colelough, and Dina Demner-Fushman
National Library of Medicine, NIH
firstname.lastname@nih.gov

Abstract

We present an exploratory analysis of how simple linguistic features relate to annotator rankings of machine-generated medical answers under task-specific evaluation guidelines. We examine eight interpretable features of the answer text: length in words, average words per sentence, percentage of polysyllabic words, medical named entity density, perplexity, coherence, lexical diversity, and dependency distance. Our preliminary results show how the role of certain features depends on the ranking guidelines. When the guidelines explicitly instruct annotators to prefer shorter answers if informativeness is equal, annotators tend to be broadly consistent with the stated criteria, though with substantial annotator-level variation. In rankings where brevity is not part of the stated criteria, length still shows a strong, but less consistent association, suggesting that answer length may influence rankings even outside explicit guideline instructions. Other features show weaker and more variable associations. These findings suggest that linguistic features may provide insight into how annotators apply ranking guidelines, but that these relationships are heterogeneous and not fully explained by simple linear trends. The results motivate further analysis of guideline adherence, annotator variation, and possible confounding between linguistic form and answer content in medical answer evaluation.

1 Introduction

Human judgments of answer quality not only reflects whether an answer is correct, but also how that answer is written. Prior work in community question answering has shown that linguistic properties such as answer length, relevance to the question, and readability can influence which answers users prefer or select as best (Beygelzimer et al., 2015; Banjar et al., 2024). The influence of writing style on task-specific quality judgments is especially important in the medical domain, where readers must evaluate information that is often technical and difficult to process. Prior studies have shown that standard readability formulas do not fully capture the difficulty of medical text, and that features beyond simple word and sentence length may be needed to better characterize how understandable a document is (Wu et al., 2013; Zheng and Yu, 2018). Relatedly, Roberts and Demner-Fushman (2016) demonstrate the value of analyzing medical questions through multiple linguistic levels, comparing consumer and professional health questions using lexical, readability, language-model, semantic-type, and question-decomposition features. The findings from Roberts *et al.* demonstrate that medical text differs not only in surface readability, but also in terminology, background information, question structure, and semantic focus. Motivated by evidence that medical text varies across linguistic, semantic, and readability dimensions, we examine how a small set of interpretable linguistic features is associated with annotator rankings under task-specific guidelines. Our exploratory analysis is intended to establish whether associations are consistent with the guidelines, vary across annotators, or point to additional factors requiring future study.

2 Data

We use the ClinIQLink dataset, which comprises textbook-grounded medical questions each paired with eight machine-generated answers from a diverse set of open-source and internal language model systems (Colelough et al., 2025). Specifically, we draw from Task 2 of the dataset, a medical student ranking study that restricts evaluation to four open-ended question types, namely short answer, short inverse, multi-hop, and multi-hop inverse. Students from accredited medical schools ranked each set of eight answers according to a provided set of task-specific guidelines. We focus on the short answer and short inverse question types

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1080–1088

Table 1: Example question and selected ranked model-generated answers.

Question: *What type of cancers do gene-based tests serve as the standard of care?*		
Rank	**Judgment**	**Answer**
1	Good	Gene-based tests are standard of care in certain breast (BRCA1, BRCA2), ovarian (BRCA1, BRCA2), colon (MLH1, MSH2, MSH6, PMS2, EPCAM), and lung (EGFR, ALK, ROS1, BRAF, KRAS) cancers, as well as melanoma (BRAF) and thyroid (BRAF, RET) cancers. They help determine risk, diagnosis, prognosis, and targeted therapy options.
2	Good	Gene-based tests serve as the standard of care for certain types of cancers, including breast cancer (BRCA1/BRCA2 mutations), colorectal cancer (MLH1, MSH2, MSH6, PMS2 mutations), ovarian cancer (BRCA1/BRCA2 mutations), prostate cancer (BRCA1/BRCA2, HOXB13 mutations), and lung cancer (EGFR, ALK, ROS1 mutations). These tests help in identifying hereditary cancer syndromes and guiding treatment decisions.

in our preliminary results, as these two types offer the clearest contrast in annotation guidelines. For short answer questions, annotators were instructed to rank the shorter of two equally informative answers higher, whereas for short inverse questions, annotators were instructed to rank answers solely by accuracy, strength of argument, and grounding. In addition to ranking, annotators were asked to assign a judgment to each answer, choosing from "Good", "Ok", or "Bad". "Good" indicates that the answer is complete and correct, "Ok" indicates the answer is acceptable with slight discrepancy, and "Bad" indicates that the answer is not factually accurate or does not address the question. To reduce the influence of overt factual incorrectness and focus on variation among acceptable answers, only answers judged "Good" were considered for this study. Examples of ranking may be found in Table 1.

3 Methods

We define the below set of metrics for linguistic features to analyze in the answer texts. We aim to see how they associate with the annotators' rankings under the provided guidelines. The metrics were selected to sample a range of commonly studied linguistic features and reveal preliminary patterns, before expanding to a broader feature set in future work. The linguistic features we use to analyze the CLinIQLink dataset include:

Lexical Count We measure the length of the text in terms of the number of words.

Average Words Per Sentence We measure sentence length as the average number of words in a sentence.

Percentage Polysyllabic Words The percentage of polysyllabic words is frequently included as a component of readability metrics, defined as the percentage of words with three or more syllables. If a word exists in CMUdict (Bartlett et al., 2009), we count the number of vowel phonemes in the word. If it does not, we approximate syllable count using Hunspell hyphenation dictionaries.

Medical Entity Density Using MetaMapLite (Demner-Fushman et al., 2017), we record the number of unique spans recognized as a medical named entity. We compute density by normalizing by the number of words in the text.

Dependency Distance Dependency distance, described by Oya (2011), is used as a measure of syntactic complexity, where a greater distance represents greater complexity. We average over the dependency distance for each sentence.

Coherence We measure the coherence between two sentences as the cosine similarity between their embedding vectors. We use `all-MiniLM-L6-v2` [1] to produce embedding vectors. We average over the coherence for each pair of consecutive sentences.

MATTR To capture lexical diversity, we measure the Moving-Average Type-Token Ratio (MATTR) of the text (Covington and McFall, 2010). MATTR is the average of the ratio of unique words to total words in a specified window size, for each window in the text. We use a window size of 50.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Perplexity Perplexity is the exponentiated average negative log-likelihood of a sequence. We use GPT-2 to compute the log-likelihoods for each token.

3.1 Additive Utility Paired Comparison Model

To measure the relationships between each metric and the orderings, the ranked lists are converted into pairwise data. We model each answer i's latent utility, η_i, as an additive function of its M metrics of linguistic features. Specifically, x_{im} denotes the value of feature m for answer i, and $f_m(\cdot)$ is a feature-specific spline function that captures the potentially nonlinear contribution of feature m to the answer's overall utility, as shown in Equation 1:

$$\eta_i = \sum_{m=1}^{M} f_m(x_{im}) \tag{1}$$

We then model the probability that answer i is ranked over answer j, denoted $P(i \succ j)$, as a logistic function of the difference between their utilities, $\eta_i - \eta_j$, as shown in Equation 2:

$$P(i \succ j) = \text{logit}^{-1}(\eta_i - \eta_j) \tag{2}$$

Thus, answers with higher estimated utility are more likely to be preferred in pairwise comparisons.

3.2 Normalization and Cross-Validation

Metrics are first normalized and missing data, such as MATTR for texts shorter than the window size or coherence for texts with only one sentence, are imputed with the average value for that feature after normalization. A missingness indicator is introduced as a covariate for each feature with imputed values. We use 5-fold cross validation to reduce variance in our performance metrics and split folds by question to prevent leakage. Annotators with fewer than 20 ranked lists for a particular question type were excluded for that question type.

3.3 Annotator Summary Metrics

We summarize the models across annotators through a variety of metrics. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators, giving an overall direction of the curve for that feature. We count the percentage of annotators that exhibit a positive and negative direction for a given feature with a threshold of 0.05, such that a value x is not counted if $-0.05 \leq x \leq 0.05$. Utility range is $\max f(x) - \min f(x)$ and nonlinearity is measured as the Root Mean Square Error between the estimated smooth curve and its best-fitting straight line, normalized by the range of the smooth curve. To analyze annotator agreement, we use curve correlation and curve variability. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean of the standard deviations of the dispersion of annotator-specific smooth curves across the feature range.

3.4 Feature Ablation

To assess the importance of each feature to the model, we conduct a leave-one-out ablation study. We fit the model again, with one feature removed, and measure the change in held-out negative log-likelihood. Positive values indicate that removing the feature worsened predictive performance.

4 Results

4.1 Short Answer Feature Relationships

Table 2 summarizes the feature relationship metrics across annotators for the short answer questions. Lexical count shows the strongest directional indication with a median direction of -0.238, and also the largest utility range at 1.168, though its IQR is large at $[-0.747, -0.081]$, reflecting substantial variation across annotators. A large percentage of annotators demonstrate a negative direction for both lexical count (76.5%) and dependency distance (70.6%), while perplexity has the largest percentage of annotators below the threshold at 52.9%. None of the remaining features show strong directional indication. Nonlinearity is broadly similar across all features.

4.2 Short Answer Curve Agreement and Ablation Results

Table 3 summarizes curve agreement and ablation results for the short answer questions. Lexical count and dependency distance mean have the greatest curve correlations at 0.824 and 0.822 respectively, though only lexical count exhibits high curve variability at 0.439, suggesting that even among annotators for whom length is a strong predictor, the shape of the relationship or magnitude of the effect size differs considerably. All features show very low magnitude changes in negative log-likelihood across the ablation study.

Table 2: Annotator-level feature relationship summaries for short answer questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.238 [-0.747, -0.081]	17.6	76.5	1.168 [0.834, 1.795]	0.084 [0.017, 0.131]
Pct. Polysyllabic Words	-0.005 [-0.112, 0.108]	41.2	35.3	0.641 [0.396, 0.875]	0.129 [0.098, 0.216]
Avg. Words Per Sentence	-0.082 [-0.175, 0.080]	41.2	52.9	0.637 [0.393, 0.971]	0.106 [0.067, 0.190]
Medical Entity Density	-0.059 [-0.127, 0.033]	23.5	52.9	0.570 [0.321, 0.678]	0.122 [0.068, 0.182]
MATTR	-0.058 [-0.249, 0.059]	35.3	52.9	0.496 [0.174, 0.618]	0.140 [0.109, 0.208]
Dependency Distance Mean	-0.090 [-0.158, 0.005]	17.6	70.6	0.486 [0.326, 0.642]	0.109 [0.072, 0.181]
Perplexity	0.000 [-0.072, 0.019]	11.8	35.3	0.413 [0.272, 0.665]	0.043 [0.020, 0.135]
Coherence	0.056 [-0.113, 0.099]	52.9	35.3	0.359 [0.237, 0.534]	0.162 [0.130, 0.180]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 3: Annotator-level curve agreement and ablation summaries for short answer questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.824 [-0.539, 0.990]	0.439	-0.002 [-0.017, 0.007]
Pct. Polysyllabic Words	-0.125 [-0.932, 0.900]	0.131	-0.001 [-0.004, 0.004]
Avg. Words Per Sentence	0.091 [-0.906, 0.973]	0.108	0.000 [-0.004, 0.003]
Medical Entity Density	0.421 [-0.940, 0.959]	0.085	-0.001 [-0.004, 0.003]
MATTR	-0.116 [-0.925, 0.930]	0.159	0.006 [-0.016, 0.023]
Dependency Distance Mean	0.822 [-0.955, 0.992]	0.067	-0.004 [-0.007, -0.001]
Perplexity	0.650 [-0.996, 0.999]	0.091	-0.000 [-0.004, 0.007]
Coherence	-0.196 [-0.939, 0.945]	0.127	-0.006 [-0.008, 0.021]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

4.3 Short Inverse Feature Relationships

Table 4 summarizes the feature relationship metrics across annotators for the short inverse questions. Lexical count again shows the strongest directional indication with a median direction of -0.301 and the largest utility range at 1.932, though the IQR is considerably wider at $[-1.126, 0.491]$ than for the short answer questions, reflecting a more even split between annotators favoring shorter and longer answers. Average words per sentence shows the next strongest negative direction at -0.118, while medical entity density is the only feature with a notably positive direction at 0.101, with 60.7% of annotators exhibiting a positive association. Nonlinearity remains broadly similar across all features.

4.4 Short Inverse Curve Agreement and Ablation Results

Table 5 summarizes curve agreement and ablation results for the short inverse questions. Lexical count and average words per sentence have the greatest curve correlations at 0.571 and 0.489 respectively, and lexical count again exhibits the highest curve variability at 0.496. Lexical count is the only feature to show a meaningful positive change in negative log-likelihood upon removal, with a median ablation ΔNLL of 0.019 $[0.003, 0.050]$.

5 Discussion

5.1 Length and Annotator Guidelines

For the short answer questions, lexical count shows a very strong negative direction, meaning annotators, on average, ranked shorter text over longer text. This trend is closely aligned with the guidelines asking annotators to rank shorter answers higher for the short answer question type. However, the large IQR reveals variation in degree of direction, meaning that some annotators may not have strictly adhered to the brevity guideline, at least for some questions. We also see a strong negative direction for short inverse questions, despite annotators not being asked to consider length in their rankings. The IQR is even larger, showing many annotators tended to rank longer text over shorter text, despite the overall trend. For the short inverse questions, the IQR points to annotators having a strong, yet varied, association between length and perceived accuracy, strength of argument, and grounding. It is worth noting that annotators could have been biased by the instruction to prefer shorter answers on the short answer questions.

5.2 Feature Associations

While lexical count is the only feature with significant direction for the short answer question, the short inverse answer rankings show modest asso-

Table 4: Annotator-level feature relationship summaries for short inverse questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.301 [-1.126, 0.491]	39.3	60.7	1.932 [1.109, 3.613]	0.075 [0.044, 0.116]
Perplexity	-0.017 [-0.169, 0.042]	25.0	42.9	0.684 [0.356, 0.832]	0.089 [0.055, 0.153]
Avg. Words Per Sentence	-0.118 [-0.189, 0.078]	28.6	60.7	0.661 [0.465, 0.875]	0.099 [0.057, 0.170]
MATTR	-0.032 [-0.230, 0.152]	35.7	50.0	0.631 [0.395, 0.870]	0.126 [0.101, 0.181]
Pct. Polysyllabic Words	-0.038 [-0.356, 0.086]	32.1	50.0	0.613 [0.347, 0.868]	0.134 [0.097, 0.195]
Medical Entity Density	0.101 [-0.083, 0.213]	60.7	32.1	0.501 [0.352, 0.739]	0.139 [0.084, 0.197]
Coherence	-0.032 [-0.127, 0.156]	35.7	46.4	0.371 [0.339, 0.488]	0.166 [0.139, 0.198]
Dependency Distance Mean	-0.051 [-0.144, 0.044]	21.4	50.0	0.304 [0.238, 0.494]	0.163 [0.092, 0.212]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 5: Annotator-level curve agreement and ablation summaries for short inverse questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.571 [-0.975, 0.985]	0.496	0.019 [0.003, 0.050]
Perplexity	0.339 [-0.916, 0.940]	0.092	-0.001 [-0.004, 0.001]
Avg. Words Per Sentence	0.489 [-0.921, 0.974]	0.122	-0.001 [-0.005, 0.003]
MATTR	-0.081 [-0.938, 0.940]	0.175	-0.002 [-0.007, 0.002]
Pct. Polysyllabic Words	-0.054 [-0.918, 0.895]	0.145	0.000 [-0.005, 0.003]
Medical Entity Density	0.160 [-0.934, 0.945]	0.140	-0.003 [-0.005, 0.001]
Coherence	-0.156 [-0.904, 0.897]	0.107	-0.002 [-0.008, 0.002]
Dependency Distance Mean	-0.055 [-0.840, 0.894]	0.090	-0.005 [-0.012, -0.002]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

ciation with shorter sentences and higher medical entity density. It is to be expected that we may see stronger relationships with more features when one specific feature is not singled out in the guidelines. Both the positive and negative directions claim a significant portion of annotators for the short inverse questions, which signal variation in the associations across annotators. The significant percentage of annotators with negative directions for dependency distance, despite a low direction value for the short answer questions, could mean agreement, but weak association, or a nonlinear relationship not captured by the direction metric.

5.3 Nonlinearity and Annotator Variation

All features, for both question types, show some degree of nonlinearity, meaning some may have complex relationships. There may be, for example, a parabolic relationship between a specific feature and an annotator's rankings that would not be observable in the direction metric. For the short answer questions, lexical count has a high curve correlation, but also high variability. This again suggests even across annotators for whom length is a strong predictor of ranking, the relationships are very different. This variance in relationship implies that the brevity guideline was not applied uniformly across annotators.

5.4 Ablation and Feature Concurvity

The leave-one-out ablation study describes only lexical count as having a significant positive difference in negative log-likelihood and only for the short inverse question type. Though some features likely truly have little impact on the model, the negligible deltas of others could be explained by concurvity among features. Despite normalizing for length or selecting metrics relatively robust to length, some other features likely carry some signal for length through specific behavior at extremely short lengths. The percentage of polysyllabic words, for example, must be either 0% or 100%, when lexical count is 1.

6 Conclusion

This analysis of linguistic features in medical text and annotator rankings shows strong relationships for some annotators, as well as broader trends across the cohort. These findings highlight the nontrivial relationships between linguistic features and ranking judgments under task-specific guidelines. Future work includes assessing the nature of specific complex relationships, the inclusion of a broader set of features, and more detailed analysis.

Limitations

There are far more linguistic features that are frequently analyzed in both medical and open-domain texts than included in this study. The answer choices for each question were generated by the same eight systems and system identity was not ruled out as a potential confounder. Furthermore, because answers not judged "Good" were excluded from the study, it may be that some trends are better explained by the factuality of the answer, rather than the features analyzed. Questions were not analyzed by medical domain, which will be necessary to understand if the results depend on the topic of the text. Finally, because this study infers guideline application from associations between linguistic features and rankings, it cannot determine why annotators deviated from guideline-consistent patterns. Such deviations may reflect misunderstanding, disagreement with the guideline, attention to unmeasured answer qualities, or residual differences in correctness or informativeness among answers judged "Good."

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Ameen Banjar, Awais Shaheen, Tehmina Amjad, Riad Alharbey, and Ali Daud. 2024. Users' satisfaction based ranking for yahoo answers. *Multimedia Tools and Applications*, 83(28):71265–71284.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316, Boulder, Colorado. Association for Computational Linguistics.

Alina Beygelzimer, Ruggiero Cavallo, and Joel Tetreault. 2015. On yahoo answers, long answers are best. In *Proceedings of the ICML 2015 Workshop on Crowdsourcing and Machine Learning (CrowdML)*.

Brandon Colelough, Davis Bartels, and Dina Demner-Fushman. 2025. Overview of the ClinIQLink 2025 shared task on medical question-answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 378–387, Vienna, Austria. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17:100 – 94.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *J. Am. Med. Inform. Assoc.*, 23(4):802–811.

Danny T. Y. Wu, David A. Hanauer, Qiaozhu Mei, Patricia M. Clark, Lawrence C. An, Jianbo Lei, Joshua Proulx, Qing Zeng-Treitler, and Kai Zheng. 2013. Applying multiple methods to assess the readability of a large corpus of medical documents. *Studies in Health Technology and Informatics*, 192:647–651.

Jiaping Zheng and Hong Yu. 2018. Assessing the readability of medical documents: A ranking approach. *JMIR Medical Informatics*, 6(1):e17.

A Appendix

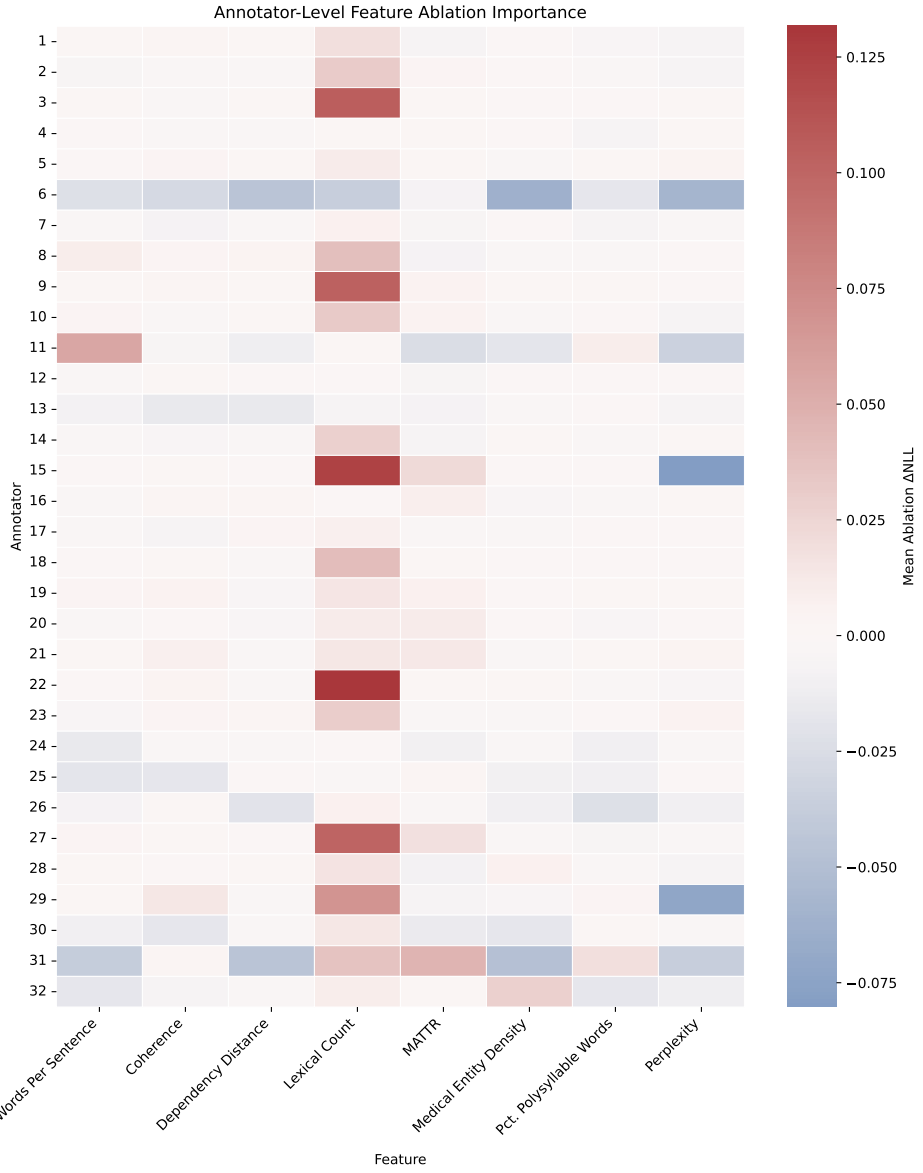

Figure 1: Heatmap of annotators and features. Color indicates the change in negative log-likelihood when each feature is removed.

Table 6: Annotator-level cross-validated model performance.

Annotator	Mean NLL	NLL SD	Pairwise acc.	Pairwise acc. SD	Kendall's τ	Kendall's τ SD
1	0.667	0.037	0.613	0.044	0.142	0.094
2	0.573	0.069	0.728	0.029	0.417	0.062
3	0.470	0.041	0.816	0.022	0.591	0.087
4	0.689	0.013	0.542	0.029	0.096	0.050
5	0.656	0.020	0.611	0.022	0.197	0.051
6	0.783	0.063	0.480	0.076	-0.012	0.113
7	0.698	0.011	0.520	0.033	0.037	0.031
8	0.524	0.075	0.760	0.078	0.483	0.145
9	0.416	0.030	0.818	0.020	0.589	0.053
10	0.621	0.058	0.662	0.036	0.321	0.071
11	0.692	0.072	0.577	0.046	0.129	0.087
12	0.698	0.005	0.508	0.016	0.016	0.035
13	0.719	0.013	0.479	0.036	-0.013	0.116
14	0.668	0.019	0.576	0.021	0.173	0.057
15	0.471	0.177	0.838	0.025	0.609	0.054
16	0.621	0.034	0.656	0.028	0.325	0.089
17	0.615	0.032	0.662	0.039	0.309	0.063
18	0.607	0.017	0.674	0.017	0.368	0.035
19	0.572	0.011	0.715	0.020	0.396	0.066
20	0.666	0.018	0.586	0.020	0.123	0.110
21	0.585	0.051	0.696	0.036	0.383	0.077
22	0.420	0.017	0.812	0.026	0.620	0.076
23	0.649	0.019	0.600	0.032	0.153	0.043
24	0.689	0.036	0.551	0.037	0.098	0.077
25	0.667	0.035	0.604	0.022	0.195	0.053
26	0.719	0.050	0.548	0.051	0.094	0.143
27	0.494	0.041	0.784	0.025	0.571	0.066
28	0.647	0.034	0.636	0.036	0.214	0.103
29	0.573	0.023	0.727	0.027	0.473	0.074
30	0.691	0.018	0.560	0.053	0.138	0.138
31	0.706	0.042	0.551	0.065	0.202	0.162
32	0.676	0.010	0.596	0.026	0.165	0.119

Note. Mean NLL, pairwise accuracy, and Kendall's τ are weighted cross-validated estimates. SD columns are fold-level standard deviations.

Table 7: Annotator-level feature relationship summaries for multi-hop questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.566 [-0.898, 0.039]	24.000	68.000	1.333 [0.684, 2.596]	0.079 [0.058, 0.145]
Coherence	-0.009 [-0.260, 0.097]	36.000	48.000	0.695 [0.444, 0.918]	0.159 [0.118, 0.204]
Medical Entity Density	-0.093 [-0.175, 0.146]	44.000	52.000	0.535 [0.392, 0.792]	0.136 [0.077, 0.201]
Dependency Distance Mean	-0.061 [-0.202, 0.118]	32.000	56.000	0.524 [0.256, 0.750]	0.155 [0.106, 0.185]
MATTR	-0.006 [-0.088, 0.152]	44.000	36.000	0.507 [0.293, 0.792]	0.168 [0.149, 0.212]
Pct. Polysyllabic Words	-0.043 [-0.168, 0.175]	36.000	48.000	0.462 [0.287, 0.609]	0.141 [0.099, 0.190]
Perplexity	0.003 [-0.079, 0.151]	36.000	36.000	0.441 [0.213, 0.760]	0.105 [0.067, 0.194]
Avg. Words Per Sentence	-0.008 [-0.087, 0.101]	28.000	36.000	0.430 [0.297, 0.700]	0.133 [0.095, 0.195]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 8: Annotator-level curve agreement and ablation summaries for multi-hop questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.715 [-0.899, 0.965]	0.337	0.012 [-0.001, 0.040]
Coherence	-0.039 [-0.828, 0.788]	0.193	-0.001 [-0.004, 0.002]
Medical Entity Density	-0.026 [-0.922, 0.946]	0.117	-0.002 [-0.009, -0.000]
Dependency Distance Mean	0.070 [-0.910, 0.902]	0.129	-0.003 [-0.005, 0.002]
MATTR	-0.068 [-0.912, 0.891]	0.132	-0.001 [-0.003, 0.003]
Pct. Polysyllabic Words	-0.438 [-0.953, 0.941]	0.120	-0.003 [-0.007, 0.000]
Perplexity	-0.662 [-0.962, 0.945]	0.112	-0.003 [-0.008, 0.000]
Avg. Words Per Sentence	0.011 [-0.902, 0.903]	0.071	-0.002 [-0.006, -0.001]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

Table 9: Annotator-level feature relationship summaries for multi-hop inverse questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.136 [-0.874, 0.092]	33.333	62.500	1.800 [0.908, 3.066]	0.062 [0.040, 0.110]
Dependency Distance Mean	-0.010 [-0.146, 0.107]	45.833	50.000	0.514 [0.304, 0.637]	0.136 [0.074, 0.206]
Medical Entity Density	0.079 [0.012, 0.202]	62.500	16.667	0.507 [0.379, 0.767]	0.170 [0.091, 0.226]
Avg. Words Per Sentence	-0.034 [-0.183, 0.047]	25.000	45.833	0.494 [0.318, 0.716]	0.129 [0.096, 0.212]
Coherence	-0.075 [-0.184, 0.056]	29.167	54.167	0.451 [0.301, 0.591]	0.166 [0.112, 0.197]
MATTR	-0.053 [-0.172, 0.045]	25.000	50.000	0.393 [0.260, 0.648]	0.150 [0.089, 0.212]
Pct. Polysyllabic Words	-0.018 [-0.189, 0.137]	41.667	50.000	0.340 [0.262, 0.582]	0.142 [0.120, 0.200]
Perplexity	-0.001 [-0.127, 0.067]	37.500	41.667	0.287 [0.211, 0.891]	0.121 [0.063, 0.196]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 10: Annotator-level curve agreement and ablation summaries for multi-hop inverse questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.244 [-0.966, 0.975]	0.408	0.029 [0.000, 0.069]
Dependency Distance Mean	-0.274 [-0.934, 0.870]	0.112	-0.003 [-0.009, 0.002]
Medical Entity Density	0.428 [-0.610, 0.878]	0.127	-0.002 [-0.006, 0.001]
Avg. Words Per Sentence	0.097 [-0.850, 0.900]	0.105	-0.002 [-0.008, 0.002]
Coherence	0.167 [-0.893, 0.902]	0.107	-0.005 [-0.009, -0.000]
MATTR	-0.108 [-0.839, 0.891]	0.116	-0.004 [-0.007, 0.005]
Pct. Polysyllabic Words	-0.737 [-0.963, 0.956]	0.109	-0.003 [-0.007, 0.000]
Perplexity	0.020 [-0.966, 0.956]	0.076	-0.001 [-0.009, 0.005]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.