FineWeb-Conv: A Method for Finding Good Conversation Data

Robert J. Moore, Sungeun An, Jay Pankaj Gala, Divyesh Jadav

IBM Research, Almaden
650 Harry Road
San Jose, CA 95120 USA
rjmoore@us.ibm.com

Abstract

In principle, large language models could talk more like humans naturally do if they are trained on data containing the interaction patterns of human conversation. However, one challenge to training a "conversation" model is that natural conversation data are relatively difficult to find. In this paper we demonstrate a method for annotating documents at scale with a 0-5 conversation score. We use a large language model to score a sample of documents for how conversational they are. Using the annotated samples, we trained Snowflake-arctic-embed with a classification head that outputs a single regression score from 0 to 5 for conversation rating. When converted to a binary classifier using a score threshold of 4, the model achieved a precision of 94%. Our conversation score approach offers significant implications for data preparation in generative AI, particularly enhancing data annotation, filtering, and quality control.

Introduction

While large language models are generally good at interacting with users, there is still much room for them to improve at natural conversation. Even with the latest voice interfaces to popular automated agents, the models still tend to be too verbose for voice (Moskovitz et al. 2023). For example, chat models tend to generate long responses that include unsolicited information. Users must continually interrupt them to stop their output. Although large language models are routinely fine tuned for "chat," such models are not optimized for natural conversation, that is, conversation that can be done through the words alone, without a visual display.

In order to fine tune a large language model for natural conversation style output, it is essential to have data containing the interaction patterns of natural conversation. However, high-quality conversation data are not easy to find online. While online forum discussions and social media retweets are publicly available, instant messaging logs and customer service call transcripts are typically not. Moreover, the most authentic form of conversational content, transcripts of everyday face-to-face or phone conversations, are rarely shared online and, when they are, they may lack accurate transcripts. Currently, there isn't a comprehensive source of unedited, ordinary conversation transcripts available online, unlike resources for general knowledge (e.g., Wikipedia), social media (e.g., Twitter), video (e.g., YouTube), or government proceedings (e.g., C-SPAN). This gap underscores the need for more extensive public conversation datasets and the need to actively seek out such content wherever it may exist on the Internet.

In order to find conversation data where it exists on the Internet, we used the FineWeb-Edu approach by Penedo et al. (Penedo et al. 2024) for scoring any textual English content for its conversation naturalness. Our method focuses on the style of the language, the number of participants and whether it contains key phrases or actions, rather than on the quality of the information communicated. In other words, "good" conversation data, for us, means it takes the form of natural conversation and not some other form of communication, such as an encyclopedia entry, blog post, email, etc. The method detailed below results in an overall conversation score, between 0 and 5, with 5 indicated natural conversation content and 0 indicating no conversation.

Existing methods for conversation scoring primarily focus on evaluating the correctness or engagingness of conversation, not the naturalness of it (Takehi, Watanabe, and Sakai 2023; Jiang, Vakulenko, and de Rijke 2023; Yi et al. 2019; Ghazarian et al. 2020; Xu et al. 2022; Demszky et al. 2021; Mehri and Eskenazi 2020; Sakai 2023). Most existing approaches to measuring naturalness focus on speech data, analyzing respiratory patterns or acoustic cues (Zhang et al. 2010; Bari et al. 2018; Wyatt, Choudhury, and Bilmes 2007; Liang, Zhang, and Thomaz 2023). In contrast, our goal is to assess the conversational nature of language in written documents and differentiate between conversational and non-conversational content.

Our conversation scoring technique has multiple uses. Initially, it can be employed to identify high-quality conversation data within a collection of diverse documents, like Fineweb or Common Crawl. Here, quality refers to the presence of natural interaction patterns, not the information or knowledge quality. This method, to a degree, mimics a subject matter expert in the field of conversation science. Secondly, our scoring method can assess the naturalness of synthetic conversation data. Given the need for substantial amounts of data to pretrain large language models, there are numerous attempts to generate high-quality conversation data. Our method offers a way to measure how closely the

synthesized data mirrors natural data.

In this paper, we describe the FineWeb-Edu approach, the conversation prompt we developed, and the evaluation results.

FineWeb-Edu Replication

To enhance the quality of the FineWeb dataset, Penedo et al. (Penedo et al. 2024) developed an educational quality classifier designed to assess the educational value of web pages on a scale from 0 to 5. This classifier, trained on 450K annotations generated by the LLama3-70B-instruct model, was created to filter and curate educational content from web datasets. The resulting subset, FineWeb-Edu, comprises 1.3T tokens of educational web pages extracted from the FineWeb dataset (Lozhkov et al. 2024). LLMs pretrained on FineWeb-Edu demonstrate significantly improved performance on knowledge- and reasoning-intensive benchmarks, including MMLU and ARC.

Following their approach, we implemented the following steps to train our classifier and curate high-quality conversational content:

- Randomly sample 600K documents from the 15T-token FineWeb dataset.
- Develop a prompt to annotate the 600K sampled FineWeb documents.
- Annotate these 600K samples using the Mixtral 8X22B model.
- Train the classifier on 450K annotated samples (split into 450K for training, and 75K each for validation and test sets).
- Evaluate classifier performance to ensure accurate classification of educational quality.
- Use the classifier to annotate the complete FineWeb dataset for the FineWeb-Conv subset.

Our prompt for conversation annotation, adapted from FineWeb-Edu's original prompt (Penedo et al. 2024), retained the template structure while modifying the content to suit our needs. The original FineWeb-Edu prompt, designed to assign a 5-point educational score, leverages an additive scale as described by Yuan et al. (Yu and Dhillon 2022). This approach enables the LLM to evaluate each criterion independently and incrementally build the score, contrasting with a fixed-category scoring approach (Li et al. 2023).

In their annotation prompt, five criteria are presented to assess a page's educational value, particularly for primary to secondary education contexts. Points are accumulated based on the degree to which each criterion is satisfied. After assessing the content, the model is instructed to justify the total score in up to 100 words, concluding with the format: "Educational score: <total points>". In the next section, we describe the design and theoretical framework of our conversational prompt.

Prompt for Conversation

Our approach is adapted from the IBM Natural Conversation Framework (Moore and Arar 2019; Moore, An, and Ren 2023; Moore, An, and Marrese 2024), which identifies generic conversational actions that occur across domains and use cases. For example, whether talking about dinner or bank accounts or medical symptoms, speakers in a natural conversation may say, "What did you say?" or "Never mind" These generic actions function to help participants manage the conversation as a form of synchronous interaction (Moore and Arar 2019). Unlike entities and domain-specific knowledge, which may occur in any type of communication, such as letters or text books or encyclopedia entries or conversation transcripts, the generic interaction features are distinctive of natural conversation. Therefore, where there is a concentration of these features, the language content is natural conversation.

In other work (Moore et al. forthcoming), we used a natural language classifier to label each line of documents to identify generic conversation features. We scored the content in terms of the range and density of the features. In this paper, we attempt to achieve the same result but through a different method. Instead of classifying each line of text and analyzing the distribution of features, we instruct a large language model to classify language content using a subset of these features, as well as, other criteria.

We adapt these methods for finding educational data and for finding conversational data (Penedo et al. 2024; Moore et al. forthcoming) into a single method. We retain the additive approach, but we replace the five educational criteria with five criteria indicative of natural conversation (see Appendix for full text of prompt). The five criteria used in our new prompt include whether it contains: 1) an oral or speech style of language, characterized by simple sentences, common vocabulary, idioms and more; 2) language from multiple speakers; 3) language from multiple speaker who are interacting with each other, that is, responding to each other; 4) common phrases used to achieve mutual understanding or affiliation, such as "what do you mean" and "yeah"; and 5) common phrases characteristic of real-time interactions, such as "hello" and "okay."

While the second criterion, multiple speakers, is obvious for conversation content, the remaining four are not. The language in natural conversation tends to have an informal style (criterion 1) in contrast to the formal style of language in writing. We discovered that indicating multiple speakers is not enough. Some content, such as reviews, contains multiple speakers but not conversation. Therefore the speakers must be interacting with each other (criterion 3). In addition, in natural conversation ever-present goals are to understand the other participants and often to affiliate with them, and certain generic keywords and phrases are used to achieve those goals (criterion 4). Finally, natural conversations are real-time interactions and certain generic phrases are used to coordinate that interaction (criterion 5). The presence of all five criteria tends to indicate that the language content is from a natural conversation, for example, a transcript, whereas the absence of all of them tends to indicate a written style document.

Discrepancy Score	Occurrence
0	465582
1	115669
2	16859
3	1730
4	150
5	10

Table 1: Discrepancy scores (0–5) indicating the difference between classifier predictions (student model) and ground truth scores from the Mixtral 8x22 model, across a sample of 600,000 cases.

Results

Model Training

The classifier was trained on 450,000 pairs of web samples and their scores from 0 to 5, generated by Mixtral 8x22. The samples were annotated based on their conversationality with 0 being not conversational and 5 being highly conversational. The previous section presents the prompt used for Mixtral annotation and the theoretical framework underlying its development.

We added a classification head with a single regression output to Snowflake-arctic-embed and trained the model for 20 epochs with a learning rate of 3e-4. During training, the embedding and encoder layers were frozen to focus on the classification head.

Evaluation: Discrepancy Metrics

We compared the scores between the trained classifier and the ground truth annotated by the Mixtral 8x22 model across 600,000 samples. Table 1 presents each discrepancy score (ranging from 0 to 5) and its frequency of occurrence within the dataset. Each score quantifies the degree of difference between the classifier's predictions and the Mixtral model's ground truth. A score of 0 indicates exact agreement between the classifier output and Mixtral 8x22, while larger scores represent greater discrepancies, with higher values indicating less alignment between the classifier's predictions and the ground truth.

The distribution in Table 1 shows that classifier predictions generally align well with the ground truth, as smaller discrepancies are more frequent than larger ones. The most common score, 0, occurs 465,582 times. As the discrepancy score increases, the occurrence decreases substantially, with only 10 instances for the maximum score of 5.

Evaluation: Model Performance Metrics

We created a manually annotated dataset consisting of 180 samples and evaluated the trained classifier in a binary classification setting. We selected 60 samples from group 1 (scores of 0, 1, and 2), 60 samples from group 2 (score of 3), and 60 samples from group 3 (scores of 4 and 5) from a 75,000-sample test split for manual ground truth annotation. By manually annotating the binary ground truth, we obtained the following true labels:

- Negative class (n = 123): samples that do not contain conversation or are borderline.
- Positive class (n = 57): samples that contain conversation.

Next, we determined the threshold for classification. Using a threshold of 3, any sample rated below 3 by the conversation classifier is classified as the negative class, while samples rated 3 or above are classified as the positive class. Table 2 compares the model's performance metrics when using thresholds of 3 and 4.

As shown in Table 2, the choice of threshold significantly impacts the performance metrics of the classifier. For a threshold of 3, the model achieved a precision of 0.75 and a recall of 0.70, resulting in an F1 score of 0.73 and an accuracy of 0.83. In contrast, with a threshold of 4, precision increased to 0.94, indicating a higher proportion of true positive predictions among those classified as positive. However, this threshold also resulted in a recall of only 0.26, suggesting that many positive instances were missed.

Given these trade-offs, we selected a threshold of 4 for our final evaluation. This choice prioritizes precision, reducing the number of false positives and thereby ensuring that when the classifier predicts a sample contains conversation, it is highly likely to be accurate. Although the recall is lower, the increased precision aligns better with our goal of minimizing false positives in contexts where incorrect classifications could lead to significant misunderstandings or errors.

Metric	**Threshold 3**	**Threshold 4**
Precision	0.75	0.94
Recall	0.70	0.26
F1 Score	0.73	0.41
Accuracy	0.83	0.76

Table 2: Comparison of precision, recall, F1 score, and accuracy at Thresholds 3 and 4. Support for class 0: 123. Support for class 1: 57

Examples

We provide examples of high score conversation example and low score conversation example.

High score conversation example (TP)

```
TS: Uh-huh. About that, actually.
Prior to Thor's passing, you and
he were together. And, whereas you
initially appeared to be but an
in-shape woman with an interest in
being a hero, you were suddenly a
super-powered warrior astride a
winged horse. What happened?
V: I don't understand....
TS: It's okay to be scared. Someone
you loved was killed and it not
only means you are without him, but
that you may not be able to do what
you once did.
```

```
V: That's not true!
TS: But you are still worried that
it might be, aren't you?
TS: And you, Captain, sir? You
haven't had an opportunity to speak
yet.
CAPTAIN AMERICA: I'm...fine.
TS: Just fine? I thought someone
said earlier you were in a
relationship with Janet Pym, the
Wasp. Did her death affect you in
any way?
CA: What do you think?
TS: I think it must have.
TS: I'm sorry.
```

In this high scoring example (score=5), we can see evidence of all 5 criteria: 1) it is written in an oral style, with first and second person perspective and simple sentences, 2) there is evidence of more than one speaker, 3) there is evidence that the multiple speakers are talking to each other, responding to what the other says, 4) there are phrases used to achieve mutual understanding and affiliation, such as "I don't understand", "what do you think?" and "I'm sorry", and 5) there are phrases characteristic of real-time, synchronous interaction, such as "uh-huh" and "aren't you?"

Low score conversation example (TN)

```
Carbon Black is the leading
provider of a next-generation
endpoint-security platform designed
to enable organizations to stop
the most attacks, see every
threat, close security gaps, and
evolve their defenses. The Cb
Endpoint Security Platform helps
organizations of all sizes replace
legacy antivirus technology, lock
down systems, and arm incident
response teams with advanced
tools to proactively hunt down
threats. Today, Carbon Black has
approximately 2,000 worldwide
customers, including 25 of the
Fortune 100 and more than 650
employees. Carbon Black was
voted Best Endpoint Protection by
security professionals in the SANS
Institute's Best of 2015 Awards.
```

In contrast to the previous case, this low scoring example (score=0) lacks the five conversation criteria: 1) the language is in a written style, for example using third-person perspective, technical vocabulary and complex sentences, 2) there is no evidence of multiple speakers, 3) there is no evidence of multiple speakers talking to each other, 4) there are no phrases indicative of mutual conversation and affiliation, and 5) there are no phrases indicative of real-time interaction. This bit of language is from a written document, not from a spoken conversation.

Test with Additional Datasets

We further evaluated our classifier using additional datasets that serve as gold standards for negative and positive cases. For a negative case, we used the PubMed Central (PMC) dataset, a free, full-text archive of biomedical and life sciences journal literature (Sayers et al. 2022). For a positive case, we used the Newport Beach dataset, which includes a collection of phone call transcripts curated by Conversation Analysts (Jefferson 2015). We randomly sampled 20 entries from each dataset. As shown in Table 3, the score range for PubMed Central was 0-1 (non-conversational, written documents), with a median score of 0, while for Newport Beach it was 4-5 (conversational, transcripts), with a median score of 5. These results align with our expectations, further validating the classifier's performance.

Table 3: Descriptive Statistics of Conversation Scores by PubMed Central and Newport Beach.

Dataset	N	Mean	Std. Dev	Min	25%	Median	75%	Max
PubMed Central	20	0.05	0.22	0.0	0.0	0.0	0.0	1.0
Newport Beach	20	4.90	0.31	4.0	5.0	5.0	5.0	5.0

Discussion and Conclusions

We have demonstrated one method for finding "good" conversation data, where good means that the data contain the generic interaction patterns of natural conversation, not that the topics discussed are true, valuable or interesting. We do this by identifying several generic interaction features using a large language model. The resulting student snowflake model is lightweight enough to analyze documents at scale. As stated above, the intended use of this method is to find language data containing the interaction patterns of natural conversation in order to train a large language model to engage in natural conversation style interaction.

An advantage of using a large language model to classify documents is that it introduces broader intent classification and general reasoning. While our alternative method (under review) relies on classification using a set of over 2,000 training example phrases for 80 intents, it is still limited to that set. Instead of inserting all of those examples and classes into a prompt for an LLM, we described them at a higher level. Based on our results, the LLM appears to be quite good at reasoning from our limited instructions regarding short phrases indicative of "mutual understanding," "affiliation" and "real-time interaction," along with only a few examples. And the LLM method can also understand higher level instructions like "multiple speakers" and "multiple speakers talking to each other." Specifying those two criteria through rules would be quite challenging. We still must conduct a systematic comparison to discover more precisely where the two approaches do well and poorly.

Our conversation scoring approach supports generative AI by enabling more effective data annotation, filtering, and

quality assessment. In terms of data annotation, the conversation score provides meta-information about the style of the language, indicating whether a text is a formal document, casual conversation, or something in between. By capturing these distinctions, we can efficiently extract and filter conversational data from mixed-content datasets like FineWeb, PR, Reddit, and YouTube, ensuring that training data aligns with specific conversational standards. In pretraining, models greatly benefit from high-quality conversation data, which helps them learn natural phrasing, natural length, turn-taking structures, and coherent response generation. This filtering capability enables models to better capture real-world dialogue patterns, enhancing their performance in natural language tasks. For quality control, the conversation score can serve as a benchmark for evaluating the naturalness and human-like qualities of synthetic data. If synthetic conversations consistently receive low scores, it may indicate deficiencies in natural conversational structure, highlighting areas that need refinement. Conversely, high scores suggest that synthetic data closely resembles real conversational patterns, making it more suitable for training interactive models.

Acknowledgments

We thank our colleagues in IBM Research, especially Alexei Karve for training the model and Masayasu Muraoka for assisting with experiments on the Mixtral 8×22B model.

References

Bari, R.; Adams, R. J.; Rahman, M. M.; Parsons, M. B.; Buder, E. H.; and Kumar, S. 2018. rconverse: Moment by moment conversation detection using a mobile respiration sensor. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(1): 1–27.

Demszky, D.; Liu, J.; Mancenido, Z.; Cohen, J.; Hill, H.; Jurafsky, D.; and Hashimoto, T. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

Ghazarian, S.; Weischedel, R.; Galstyan, A.; and Peng, N. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7789–7796.

Jefferson, G. 2015. *Talking about troubles in conversation*. Oxford University Press.

Jiang, S.; Vakulenko, S.; and de Rijke, M. 2023. Weakly supervised turn-level engagingness evaluator for dialogues. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 258–268.

Li, X.; Yu, P.; Zhou, C.; Schick, T.; Levy, O.; Zettlemoyer, L.; Weston, J.; and Lewis, M. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Liang, D.; Zhang, A.; and Thomaz, E. 2023. Automated face-to-face conversation detection on a commodity smartwatch with acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3): 1–29.

Lozhkov, A.; Ben Allal, L.; von Werra, L.; and Wolf, T. 2024. FineWeb-Edu.

Mehri, S.; and Eskenazi, M. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. *arXiv preprint arXiv:2006.12719*.

Moore, R. J.; An, S.; Gala, J. P.; and Jadav, D. forthcoming. Finding the Conversation: A Method for Scoring Documents for Natural Conversation Content. In *CHI '25: CHI Conference on Human Factors in Computing Systems*.

Moore, R. J.; An, S.; and Marrese, O. H. 2024. Understanding is a Two-Way Street: User-Initiated Repair on Agent Responses and Hearing in Conversational Interfaces. *Proceedings of the ACM on Human-Computer Interaction*, 8(187): 1–26.

Moore, R. J.; An, S.; and Ren, G.-J. 2023. The IBM natural conversation framework: a new paradigm for conversational UX design. *Human–Computer Interaction*, 38(3-4): 168–193.

Moore, R. J.; and Arar, R. 2019. *Conversational UX design: A practitioner's guide to the natural conversation framework*. Morgan & Claypool.

Moskovitz, T.; Singh, A. K.; Strouse, D.; Sandholm, T.; Salakhutdinov, R.; Dragan, A. D.; and McAleer, S. 2023. Confronting reward model overoptimization with constrained rlhf. *arXiv preprint arXiv:2310.04373*.

Penedo, G.; Kydlíček, H.; Lozhkov, A.; Mitchell, M.; Raffel, C.; Von Werra, L.; Wolf, T.; et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Sakai, T. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. *arXiv preprint arXiv:2305.08290*.

Sayers, E. W.; Bolton, E. E.; Brister, J. R.; Canese, K.; Chan, J.; Comeau, D. C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. 2022. Database resources of the national center for biotechnology information. *Nucleic acids research*, 50(D1): D20–D26.

Takehi, R.; Watanabe, A.; and Sakai, T. 2023. Open-Domain Dialogue Quality Evaluation: Deriving Nugget-level Scores from Turn-level Scores. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 40–45.

Wyatt, D.; Choudhury, T.; and Bilmes, J. A. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Interspeech*, 586–589.

Xu, G.; Liu, R.; Harel-Canada, F.; Chandra, N. R.; and Peng, N. 2022. EnDex: Evaluation of Dialogue Engagingness at Scale. *arXiv preprint arXiv:2210.12362*.

Yi, S.; Goel, R.; Khatri, C.; Cervone, A.; Chung, T.; Hedayatnia, B.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tur, D. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*.

Yu, Y.; and Dhillon, P. 2022. Deconstructing the structure of online conversations on Reddit. *arXiv preprint arXiv:2209.14836*.

Zhang, Y.; Bern, M.; Liu, J.; Partridge, K.; Begole, B.; Moore, B.; Reich, J.; and Kishimoto, K. 2010. Facilitating meetings with playful feedback. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 4033–4038.

Appendix

Prompt

Below is an extract from a web page. Evaluate whether the page contains social interaction between people using the additive 5-point scoring system described below.

Starting with 0, add 1 point to the score for each of the following criteria it satisfies...

Criteria 1: Add 1 point if the extract consists mostly natural language in the style of human speech. This style consists of simple sentences or short phrases, common vocabulary, first and second person perspective, contractions and idioms and colloquialisms, repetitions.

Criteria 2: Add 1 point if the extract contains natural language from multiple speakers. Two or more people, individuals, authors or agents contribute to the content in the extract. Their names may or may not be given.

Criteria 3: Add 1 point if the extract contains natural language from multiple people speaking to each other or interacting with each other. That is, each turn responds to the previous turn, forming a chain, thread or sequential order.

Criteria 4: Add 1 point if the extract contains multiple generic keywords or short phrases that show attempts by the speakers to achieve mutual understanding, agreement or affiliation, for example, 'yeah', 'yep', 'sure', 'nope', 'youre correct', 'exactly!', 'thats wrong', 'thanks', 'thank you', 'no problem', 'sorry', 'I mean', 'you know', 'what do you mean', 'right?', 'can you give an example', 'i don't understand', 'no I mean', 'oh you mean', 'what's your point?', 'never mind', 'I get it', 'great!', 'love it!', 'haha', 'lol', 'that sucks' or equivalent examples.

Criteria 5: Add 1 point if the extract contains multiple generic keywords or short phrases that are characteristic of transcripts of synchronous, real time interactions. It may include common phrases that involve the live nature of the interaction, such as, 'now?', 'wait', 'hold on', 'ready', 'next on the agenda', 'almost done', 'look here', 'over there', 'this one', 'one moment please', 'be right back', 'listen', 'go ahead', 'well', 'ok', 'okay', 'oh', 'ahh', 'aww', 'uh huh', 'mhmm', 'say again', 'what did you say?', 'what?', 'hello', 'hi there', 'how are you', 'what's your name', 'how can I help you', 'anything else', 'gotta go', 'have a nice day', 'see ya later', 'goodbye', or equivalent examples.

BEGIN EXTRACT

$< extract >$

END EXTRACT

After examining the extract: - Sum up the added points into a total score for the extract, between 0 and 5 points. - Briefly justify your total score, up to 100 words. - You must prepend the score exactly using the following format: "Conversation score: $< totalpoints >$."