Neha Pundlik Srikanth, **Jordan Boyd-Graber**, and Rachel Rudinger. **DiscoTrace: Representing and Comparing Answering Strategies of Humans and LLMs in Information-Seeking Question Answering**. *Empirical Methods in Natural Language Processing*, 2026.


```
    @inproceedings{Srikanth:Boyd-Graber:Rudinger-2026,
Title = {DiscoTrace: Representing and Comparing Answering Strategies of Humans and LLMs in Information-See
Author = {Neha Srikanth and Jordan Boyd-Graber and Rachel Rudinger},
Booktitle = {Empirical Methods in Natural Language Processing},
Year = {2026},
Url = {http://cs.umd.edu/~jbg//docs/2026_emnlp_discotrace.pdf},
}
```


Accessible Abstract: We use linguistic theory to understand how LLMs and humans answer questions. They're quite different: LLMs don't push back, humans share more personal experience. We hope to use this to guide LLMs to create answers to questions that are more human: safer, more discursive, more engaging, and maybe even funnier.

This was one of two papers selected as Best Paper at EMNLP 2026

Downloaded from `http://cs.umd.edu/~jbg/docs/2026_emnlp_discotrace.pdf`

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

DISCOTRACE: Representing and Comparing Answering Strategies of Humans and LLMs in Information-Seeking Question Answering

Neha Srikanth
Computer Science
University of Maryland
nehasrik@umd.edu

Jordan Boyd-Graber
Computing and Data Science
NTU Singapore
jordan.ying@ntu.edu.sg

Rachel Rudinger
Computer Science, UMIACS
University of Maryland
rudinger@umd.edu

Abstract

We introduce DISCOTRACE, a method to identify the rhetorical strategies answerers use when responding to information-seeking questions. DISCOTRACE represents answers as a sequence of question-related discourse acts paired with interpretations of the original question, annotated on top of rhetorical structure theory parses. Applying DISCOTRACE to answers from nine different communities reveals that communities have diverse preferences for answer construction. In contrast, LLMs do not exhibit rhetorical diversity in their answers, even when prompted to mimic specific human community answering guidelines. LLMs also systematically opt for breadth, addressing interpretations of questions that human answerers choose not to address. The rich, community-sensitive answering behavior structurally revealed by DISCOTRACE can guide the development of *pragmatic* LLM answerers that are more attuned to contextual information needs.

1 Introduction

Human–AI interactions have moved beyond simple fact retrieval ("*What is the capital of France?*") to more complex questions (Xu et al., 2022) requiring longer-form responses (Fan et al., 2019), such as asking for subjective interpretation of facts, empathy-seeking, or even advice (Shelby et al., 2025; Chatterji et al., 2025), more closely resembling the questions humans may often ask of each other. These questions may even be ambiguous (Stelmakh et al., 2022; Cole et al., 2023), vague, or contain false assumptions (Kim et al., 2021; Srikanth et al., 2024). Humans or AI can adopt different strategies when answering questions: ask a follow-up question, probe for more context, or make their best guess about the information needs of the user (Srikanth et al., 2025). As users turn to AI for questions that may be complex, flawed, domain-specific, or multifaceted, it is imperative that we understand *what* makes one answer better than another.[1]

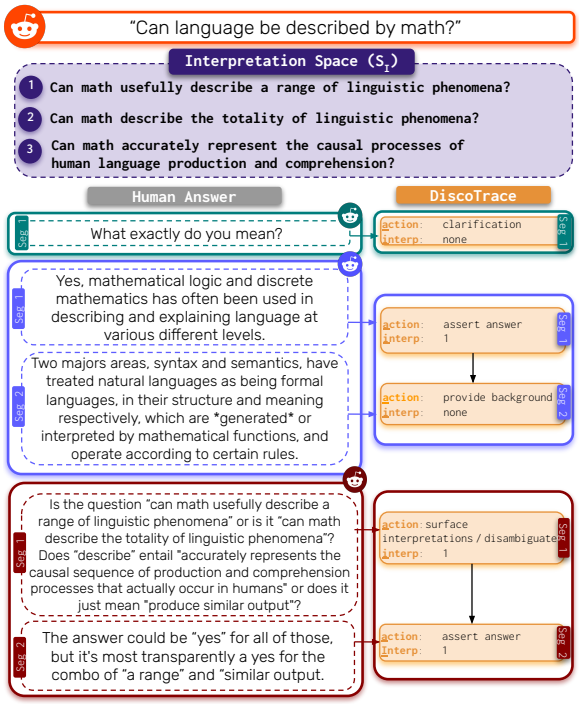


Figure 1: Given a question (top) and its interpretations (middle), DISCOTRACE provides a scaffold (bottom right) of answer rhetorical structure (bottom left) with a sequence of tuples of discourse acts and interpretations.

While reinforcement learning from human feedback (RLHF) and human preference judgments (Ethayarajh et al., 2022) aim to make answers more useful, they *implicitly* conflate many factors—factuality, completeness, style, length, relevance, or tone—with *rhetorical strategy* (Hu et al., 2023): the choices an answerer makes about *what* to address and *how* to structure a response. From a practical perspective, this diffuse signal makes it hard for models and researchers to learn *what* about an answer makes it good or how to generalize to

[1]Code and data released at https://github.com/nehasrikn/discotrace.

use this answer to help answer a different question. From a theoretical perspective, we as researchers lack the tools to model *rhetorical* differences, both in *content* (*what was addressed?*) and in *strategy* (*how was it addressed?*), in answers produced by humans in the same community, humans across communities, or LLMs.

We introduce DISCOTRACE (§3), a discourse-based representation of a long-form answer to an information-seeking question. An answer DISCOTRACE consists of a sequence of question-related *discourse acts* (Searle, 1969)—communicative moves (e.g., question an assumption or provide background)—each paired with specific *interpretations* of the original question (e.g., this is asking about Georgia the American state, not the country), annotated on top of rhetorical structure theory parses (Mann and Thompson, 1987, RST). The resulting sequence reveals the answerer's *strategy*, providing a high-level, interpretable mechanism to compare answers: which interpretations were covered, whether additional information was sought, etc. Figure 1 shows a question taken from a QA community on linguistics (`r/asklinguistics`). The DISCOTRACE of each answer surfaces three different answering strategies: the first answerer requests clarification, the second adopts an interpretation and answers it, and the third surfaces different readings of the question.

DISCOTRACE reveals just how differently humans answer questions in different online communities (§4) with different preferences and objectives (Kumar et al., 2025), which LLMs fail to do even when they are prompted to mimic answers from a particular community. Moreover, LLMs take a shallower approach to QA, hedging over many interpretations, unlike humans, who favor deeply answering only a few (§5). This extends prior observations that LLM text is lengthier and less lexically and grammatically diverse than human text (Poddar et al., 2025; Reinhart et al., 2025; Muñoz-Ortiz et al., 2024): **DISCOTRACE reveals that the homogeneity persists at a higher level of abstraction, in the strategic discourse choices LLMs make when constructing an answer to questions from vastly different domains.** Consequently, DISCOTRACE can help us develop **rhetorically flexible, pragmatic LLMs that adapt their strategies when answering questions** to diverse, sometimes competing, preferences or objectives—e.g., being more empathetic, more skeptical, or more evidence-based as context demands.

2 Motivation and Background

Unlike factoid questions that can be answered within a few tokens and typically have a single correct answer (Kwiatkowski et al., 2019), *long form* question-answering (LFQA) involves questions that may be open-ended, subjective, or complex (Dietz et al., 2017; Fan et al., 2019). Evaluating answers to these types of questions is challenging, since there are often many ways to respond, and **answerers may employ fundamentally different—but *valid*—strategies when responding to a question** (Rogers et al., 2023). The choice of strategy in an answer may be driven by the answerer's expertise (Malaviya et al., 2024), the context in which the question is asked, the answerer's reasoning or understanding of the user's information need (Taylor, 1962; Srikanth et al., 2025), and the question type (Cao and Wang, 2021), etc. **These factors admit a panoply of diverse answers with distinct rhetorical strategies.**

Much LFQA research has been centered on the ELI5 dataset (Fan et al., 2019; Krishna et al., 2021) from the subreddit `r/explainlikeimfive`, a community that rewards conceptual simplicity in explanations, but this is just *one example* of an objective that answerers may choose to prioritize (Adamic et al., 2008). Much work on learning from human preferences (Ouyang et al., 2022) also tries to align models with the preferences from a *prototypical* user, but this can similarly collapse the real-world diversity in objectives that may drive heterogeneous preferences (Bakker et al., 2022).

Given the complexity of the task, Xu et al. (2023) argue for *multifaceted* evaluation of answers, encouraging the use of finer-grained metrics, noting that in some cases, experts themselves were at odds in their preference judgments. In this spirit, we study answers and human preferences in diverse communities from a *rhetorical* perspective, characterizing their discourse *structure* (Xu et al., 2022) and choice of question interpretation.

Discourse Structure. Speakers use language to accomplish communicative goals (Searle, 2014). A *discourse act* is a functional label that helps characterize the communicative role that an utterance or a piece of text plays with respect to its context: the same surface form can perform different acts depending on the question it responds to. An explanation like *"because the liver already filters toxins continuously"* justifies an answer to *"how does the liver work?"* but rejects a presupposition

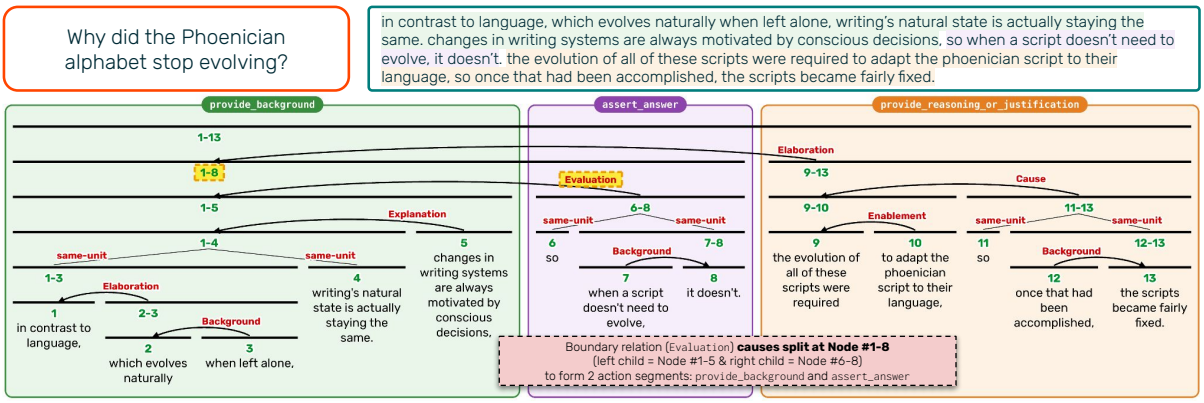


Figure 2: Rhetorical structure theory (RST) trees, designed to explain text structure by identifying relationship between text subsegments, provide useful discourse scaffolding, but not enough information alone to characterize high-level answering strategy. We first pre-process answers with RST, identifying *boundary* nodes, where children likely belong to different discourse acts and their subtrees define segment boundaries. Then, using the procedure in §3.1, we construct discourse segments which are then automatically labeled with answer-oriented discourse acts.

of *"what's the best liver detox cleanse?"*. Previous computational research has studied discourse acts in conversational speech (Stolcke et al., 2000), scientific articles (Teufel and Moens, 2002), and on Reddit (Zhang et al., 2017), but these acts are too general (e.g. *statement*, *question*, *announcement*) making them ill-fitting for studying QA. Others have characterized the discourse structure of *general* human and machine-generated text (Kim et al., 2024) using *rhetorical structure theory* (Mann and Thompson, 1987, RST), a theoretical linguistic framework to describe text coherence with relations between hierarchically organized text segments.

Perhaps most relevant, Xu et al. (2022) study the discourse structure of answers from the ELI5, WebGPT (Nakano et al., 2021), and Natural Questions (Kwiatkowski et al., 2019) datasets, presenting six roles of answer sentences (*summary*, *answer*, *example*, *auxiliary info*, *miscellaneous*, and *organizational*), but these roles focus more on how answers *organize* information, whereas our goal is to characterize what the answerer is *doing* in relation to the question asked—including moves like seeking clarification (Zhang et al., 2025) or challenging a question's premise (Yu et al., 2023)—especially across communities beyond `r/explainlikeimfive`. Additionally, they discount questions that are not "interpretable", have false presuppositions, or contain subquestions, but given that humans often ask these types of questions to LLMs (Kim et al., 2021; Min et al., 2020), we choose to model how answerers respond to these questions as a part of their answering strategy instead of filtering them out.

Question Interpretation. When responding to a question Q, answerers may address one or more readings, or *interpretations*, of Q. This may be because Q is ambiguous (Min et al., 2020) or vague (Stengel-Eskin et al., 2021) or because over-answering serves an underlying goal (e.g., to teach the asker something, Tsvilodub et al., 2023). As such, we also study how answerers address *interpretations* of questions when crafting a response. Modeling what answerers address is related to the Questions Under Discussion framework (QUD, Roberts, 2012), which proposes that discourse is organized around questions that sentences or utterances may implicitly or explicitly resolve. Computational work on QUDs (Ko et al., 2023; Wu et al., 2023; Namuduri et al., 2025) and "question-focused" discourse (Wang et al., 2025) recovers the implicit questions that drive *discourse coherence* rather than modeling how answerers select and respond to readings of an explicit question. DISCOTRACE pairs discourse acts in an answer with an interpretation of the original question to not only surface *how* the answerer responds, but also specifically *what* they responded to.

3 DISCOTRACE: Structured Discourse Representations of Answers

Here, we describe constructing the DISCOTRACE of an answer, revealing the high-level operations an answerer uses to address a question's information need. Our pipeline ingests a question Q and an answer A, and outputs a *strategy* S_{DT}—a *sequence* consisting of discourse act-interpretation tuples $\{(a_1, i_1), (a_2, i_2)...(a_n, i_n)\}$ (see Fig. 1), in

three stages: (1) rhetorical structure theory (RST)-based parsing (§3.1), (2) discourse act labeling (§3.2), (3) interpretation labeling on eligible discourse acts (§3.3).

3.1 Step 1: Preprocessing Answers with RST

Rhetorical structure theory (RST) explains the coherence and structure of text by identifying relationships between textual subsegments (Mann and Thompson, 1987). RST represents a text t as a binary constituent tree (Fig. 2) with clause-like leaves representing atomic spans of t (elementary discourse units, or EDUs) that recursively merge until the root represents the full text.

We preprocess an input answer A using a competitive end-to-end RST parser (Chistova, 2024)[2] based on `xlm-roberta-large` to construct its RST tree using 18 coarse-grained rhetorical relations (Appendix B). Preprocessing with RST provides a useful discourse scaffolding, but it alone is insufficient since: (1) RST relations hold *between* textual segments rather than between segments and some external question, and (2) EDUs are too short to serve as segments themselves to determine a higher-level discourse relation with the question (see leaf nodes, which are EDUs, in Fig. 2). Labeling segments with discourse acts (§3.2) addresses the first concern. For the second, we construct coarser **action segments** with one more preprocessing step.

Motivated by the observation that some relations are more likely to signal shifts than others (e.g. in subtopics (Cardoso et al., 2017)), we identify a subset of RST relations which we call *boundary relations* (*contrast*, *comparison*, *topic-change*, *evaluation*, *summary*, and *background*) where we expect a rhetorical shift between children; children with these boundary relations likely correspond to different discourse acts. Since a relation may signal a shift under some nuclearity configurations but not others, we define boundary relations as relation–nuclearity pairs (Table 3, Appendix).

We recursively traverse the RST tree, splitting a node's children into separate segments when the node's relation is a boundary relation, and preserving any splits that arise from deeper in the tree. Unsplit subtrees are collapsed into single spans. See appendix for the full procedure (Algorithm 1).

[2]This lightweight model (`https://huggingface.co/tchewik/isanlp_rst_v3`) pre-trained on the RST Discourse Treebank (Carlson et al., 2001) earns a span micro F1 score of 78.7 as compared to heavier-weight `llama-2-70b` earning only 1 point more at 79.8 (Maekawa et al., 2024).

This yields text spans (or action *segments*) with *suspected* distinct discourse acts which we label with our question-based act ontology next.

3.2 Step 2: Question-Related Discourse Acts

While prior work has studied specific phenomena in isolation (e.g., rejecting false presuppositions (Yu et al., 2023), follow-up questions (Meng et al., 2023), or clarifications (Majumder et al., 2021)), we develop a general ontology of 21 discourse action types designed to capture the full breadth of rhetorical moves answerers may employ across diverse question-answering settings.

Action Ontology. Our ontology of 21 discourse acts (Figure 3) is divided into five families. Answerers can *answer* the question (AQ), *comment* on the question (CQ), *seek* more information from the user (SI), *redirect* the question to a more appropriate source (RQ), or *no-op* (NO) by writing text that does not substantively advance the answer. We discuss our iterative, data-driven approach to ontology construction in Appendix C.

Parsing. Answers can, and often do, employ multiple discourse acts from our ontology (Fig. 2). We label each action segment from §3.1 with a discourse act (or, when no act is appropriate, `NONE`[3]) using `gpt-4.1-2025-04-14` (Prompt D.1). Since RST-based action segments are the result of a coarse preprocessing step, they *may* still contain multiple acts that must be broken down further. To handle this, we provide each EDU within each action segment as text subspans to the model, which can label the segment as a whole or assign different labels to individual EDUs if they serve different discourse functions (Appendix D). To reduce ambiguity at the boundaries between discourse acts, we also provide the model with the preceding segment and its predicted label as context: it can predict the *continuation* of an act or the *initiation* of a new act, as is standard in sequence tagging (Ramshaw and Marcus, 1995). This step produces a sequence of discourse act tags corresponding to subsegments.

Human Validation. Two external annotators label 200 segments (arising from 55 answers uniformly randomly sampled across the nine subreddits) using our action ontology, as well as have them match eligible actions to interpretations in

[3]The automatic labeler's usage of the `NONE` label remains under 4% of all segments in our dataset of Reddit questions, showing the broad coverage of our ontology (Table 5).

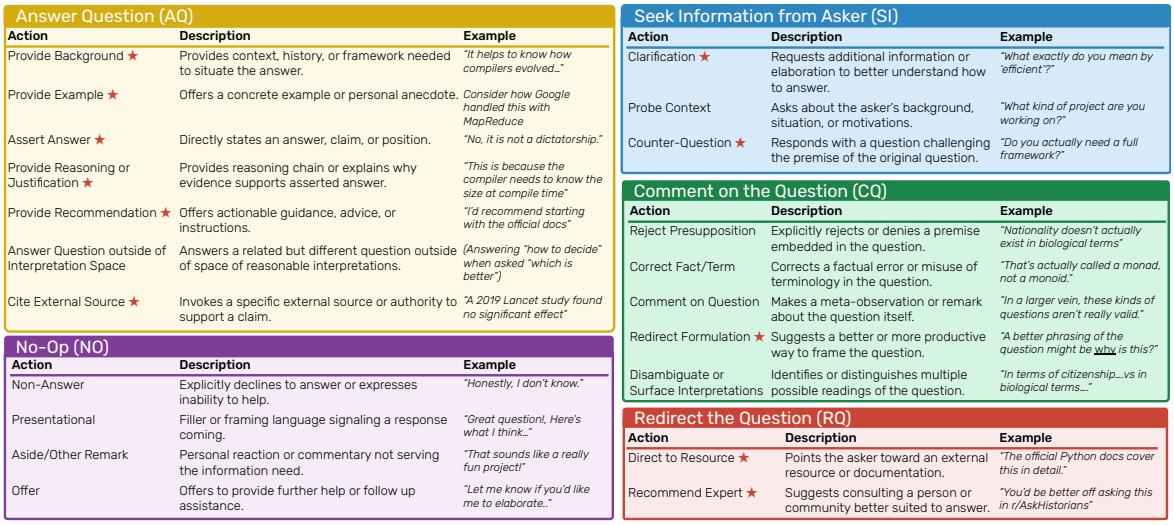


Figure 3: Each segment in a DISCOTRACE is labeled with one of these actions, grouped into five broad families. The full DISCOTRACE pairs each of these with a question-specific interpretation.

our generated space. Appendix E includes the guidelines we show to annotators. The discourse act agreement of the two external annotators is $\kappa = 0.75$ at the family level (six-way labeling: five families + NONE) as measured with Cohen's Kappa, and $\kappa = 0.62$ at the individual action level, indicating substantial agreement (Artstein and Poesio, 2008). Family-level agreement between each annotator A_1/A_2 and `gpt-4.1` is $\kappa = 0.74/0.73$ and act-level is $\kappa = 0.66/0.52$. Inter-annotator disagreement mostly concentrates within-family: of all segment disagreements, 71.9% are within-family, compared to 28.1% across family.[4]

3.3 Step 3: Pairing Acts with Interpretations

The rhetorical strategy S_{DT} produced by DISCOTRACE must not only describe *how* an answerer chooses to respond, but also to *which* interpretation(s) of the question. A question Q may admit multiple valid readings or interpretations (Min et al., 2020). This may be for a number of reasons: Q was not fully formed in an asker's mind (Taylor, 1962), the asker is not an expert in the topic or domain of Q, or Q is inherently open-ended. Understanding which *interpretation(s)* of Q an answerer addresses is therefore crucial to characterizing their strategy. For example, the question in Fig. 1 may reasonably admit at least three interpretations, and answerers may choose to adopt one or more of these interpretations, ignore others, or even surface the ambiguity explicitly.

Act Eligibility. Not all discourse acts assume or address a particular interpretation (presupposition rejection cannot adopt an interpretation, since it contests the question's premise rather than attempting to answer it), while others, like clarification, may adopt an interpretation, since they often seek to resolve ambiguity with respect to a specific reading the answerer has in mind. Our ontology marks which acts are eligible to be paired with an interpretation (marked ★ in Fig. 3), meaning that the act reasonably addresses that interpretation.

Interpretation Generation and Pairing with Discourse Acts. We first establish a set of reasonable interpretations of a question Q. Prior work identifies interpretations of Q via crowdsourcing (Min et al., 2020) or with retrieval-augmented generation (Kim et al., 2023) or direct prompting (Srikanth et al., 2025). Namuduri et al. (2025) use direct prompting with `gpt-4o` to generate questions under discussion. We generate a space of reasonable, distinct interpretations S_I from Q via direct prompting: first, we provide Q and some contextual information C[5] to two LLMs (`gpt-4.1-2025-04-14` and `claude-sonnet-4.6`) and prompt them to generate a variable number of

[4]The top three within-family confusion pairs were: ASSERT ANSWER⇔PROVIDE REASONING (AQ), ASSERT ANSWER⇔PROVIDE EXAMPLE (AQ), and PROVIDE BACKGROUND⇔PROVIDE REASONING (AQ). The top three cross-family confusion pairs were CITE EXTERNAL SOURCE⇔DIRECT TO RESOURCE (AQ vs. RQ), ASSERT ANSWER⇔REJECT PRESUPPOSITION (AQ vs. CQ), and ASSERT ANSWER⇔ASIDE/OTHER REMARK (AQ vs. NO).

[5]In our experiments, C is a short description of the subreddit in which the question was asked (Table 4), but can be adapted to any other setting simply by describing the context in which Q was asked.

generations using Prompt F.1. Models may abstain from generating any interpretations if they deem the question clear. We pool generated interpretations from both models and use the deduplication algorithm from Srikanth et al. (2025) to construct S_I (Table 6 shows $|S_I|$ for different subreddits).[6] Finally, using the act label predictions from §3.2, `gpt-4.1-2025-04-14` assigns an interpretation from S_I to each eligible act (Prompt F.2).

Human Validation. The same two annotators from §3.2 assign interpretations from S_I to their eligible acts or abstain if they believe the act does not adopt one for the 200 segments. On whether a segment addressed an interpretation at all (binary task): annotators achieve a Cohen's $k = 0.71$. For segments where *both* annotators assign an interpretation, they agree on the exact interpretation 74% of the time. Annotator-model agreement on the binary task for A_1/A_2 is $k = 0.78/0.51$ with exact interpretation agreement of $74\%/73\%$ when both the annotator and model assign one.

4 Discourse Across Humans and LLMs

DISCOTRACE can describe answers, compare answers for preference datasets, or help generate human discourse-aligned answers. Here, DISCOTRACE surfaces answering strategy differences (1) across QA communities (§4.2), signaling that the *objectives* that answerers optimize across communities are distinct, and (2) across humans and LLMs (§4.3), suggesting that by default, LLMs model a generic, narrow answering strategy. Even when prompted to mimic a particular human community, LLMs fail to capture its full strategy diversity.

4.1 Reddit Communities as Answer Data

We source answers from Reddit,[7] since it hosts communities (*subreddits*) with different answering objectives. Our nine selected communities (Table 4) vary along several axes: expert forums (`AskEconomics`, `AskHistorians`, `asklinguistics`), general QA (`NoStupidQuestions`, `OutOfTheLoop`, `explainlikeimfive`), experiential communities (`ScienceBasedParenting`, `beyondthebump`), and the topic-based general `history`.[8] For each community, we randomly sample 300 questions and their surviving top-level comments from a pool of posts filtered by quality heuristics (Appendix A). We then generate a DISCOTRACE for each of the 18,968 top-level comments (answers) to the 2,700 questions in our dataset.

4.2 Discourse Across *Human* Communities

Setup. We use the DISCOTRACE of each answer across all nine subreddits in our dataset (§4.1) to analyze distributions of answering strategies across communities. We train a bigram model over discourse act sequences for each community C, estimating $p(a_{\text{next}} \mid a_{\text{prev}})$ to capture both the acts and their ordering as an estimation of C's preferred answering "policies." We include start and end tokens to model how answers open and close. We compare community policies by computing a cross-perplexity matrix over all pairs of subreddits (C_1, C_2) by taking a trained bigram model on all answers in C_1 and evaluating it on answers in C_2. Perplexity here reflects how surprising one community's discourse act patterns are under another community's model—high perplexity indicates divergent answering policies. This cross-perplexity matrix reveals:

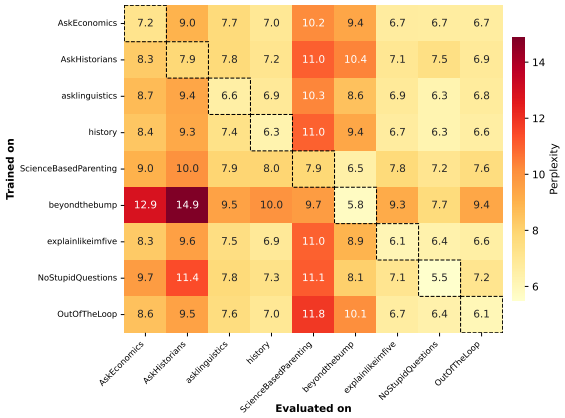
Expert and generic QA communities differ sharply in answering strategies. Domain-*agnostic* QA communities like `NoStupidQuestions` or `explainlikeimfive` are predicted well by virtually all other communities (Fig. 4a, cols. 7–9), suggesting a generic baseline answering style. In contrast, generic communities (rows/cols. 7-8) are poor predictors of expert communities (rows/cols. 1–3), while expert communities predict each other relatively well (`AskEconomics` → `asklinguistics`, `AskHistorians` → `AskEconomics`), suggesting a distinct expert answering style that may generalize across disciplines. Answerers in `AskHistorians` are more likely to ask clarification questions, cite external sources, and provide background context, while answerers in `explainlikeimfive` take a practical approach, directly asserting an answer as well as providing recommendations or advice to the asker (Appendix, Fig. 8). This expert-generic distinction is critical, as most LFQA research analyzes the ELI5 dataset, which reflects only a small slice of the diverse range of human answering policies that may satisfy different user preferences.

Shared topic does not imply shared rhetorical strategy. The communities `history` and

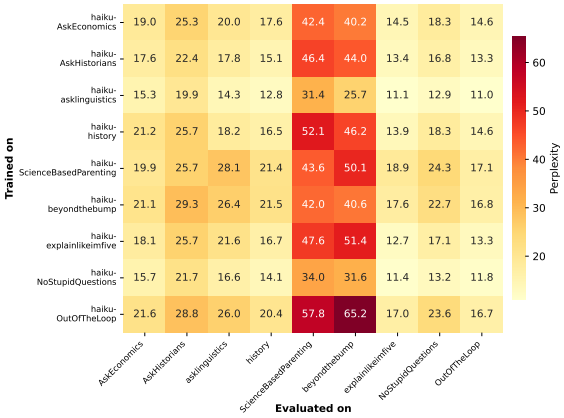
[6] We use `Qwen3-Embedding-8B` as the instruction-tuned embedder instead. All other hyperparameters are the same.

[7] Obtained from the `pushshift` dump (https://github.com/Watchful1/PushshiftDumps).

[8] We include `history` to compare with `AskHistorians` as it shares subject matter, but with less expertise asymmetry and no dedicated QA format.



(a) Cross-perplexity of the discourse act bigram model *across subreddits*: trained and evaluated on human answers.



(b) Cross-perplexity of the discourse act bigram model trained on `claude-haiku-4.5` answers, evaluated on human answers.

Figure 4: DISCOTRACE helps compare how different answering *strategies* are across different subreddits. We train bigram models on *act* sequences from one community C_{train} (rows) and evaluate on another C_{eval} (columns). **(a)** Bigram models are trained on human answers; the diagonal is each community's self-perplexity. **(b)** Bigram models are trained on `claude-haiku-4.5` answers to questions from C_{train}; here the diagonal compares the LLM's answers against human answers *in the same community*. **Note the differing scales: even the best LLM–human values (b) exceed most human–human values (a).**

`AskHistorians` both discuss history, but are poor predictors of each other's discourse structure. `AskHistorians`, with its strict moderation requiring sourced and comprehensive answers, develops a more constrained discourse style that can partially account for `history`'s looser patterns, but not vice versa. Similarly, `ScienceBasedParenting` and `beyondthebump` both revolve around parenting young children. Answering strategies in the latter are unsurprising to a model trained on the former, but the reverse is not true. While both of these communities emphasize experience sharing and community seeking when answering, `ScienceBasedParenting` additionally mandates the citation of peer-reviewed articles to support claims, resulting in a broader range of answering strategies.[9]

These patterns, recoverable from discourse act sequences alone, illustrate that QA **is not one size fits all**: similar questions posed in different communities may yield vastly different answering strategies. Where prior work conditions on subreddit name during preference optimization (Kumar et al., 2025), conflating topic, answering strategy, and community norms in a single token, DISCOTRACE offers an interpretable and generalizable alternative, surfacing explicit answering objectives (e.g. *contextualize* when answering) that future work may use when personalizing answer generation.

[9] https://www.reddit.com/r/ScienceBasedParenting/comments/1f9vwjt/welcome_and_introduction_september_2024_update/

4.3 Human vs. LLM Answering

LLMs are often post-trained on preference data that collapses distinctive preferences (Bakker et al., 2022). We ask two questions: whether LLMs' default answering strategies exhibit the cross-community rhetorical diversity we observe in humans (§4.2), and whether prompting with community context recovers the strategies of a particular community.

Default Behavior. We prompt both a closed (`claude-haiku-4.5`) and open source (`qwen3-32b`) LLM to answer questions from all communities without providing any context or instruction beyond the question itself, restricting responses to 1K tokens. Then, we train bigram models over LLM answer action sequences (§4.2). Cross-perplexity between LLM-generated answers and human answers is much higher than between any human communities (Fig. 5), showing that LLM discourse structures diverge significantly from human communities. Fig. 4b shows the full per-community cross-perplexity matrix for `haiku` (Fig. 9, Appendix, for `qwen`): both models best predict human strategies in generic, explanation-oriented communities (`explainlikeimfive`, `OutOfTheLoop`). Measuring perplexity between LLM answers across different subreddits (Figs. 10 and 11 in appendix) reveals that they deploy a largely uniform discourse structure *across question types and topics*.

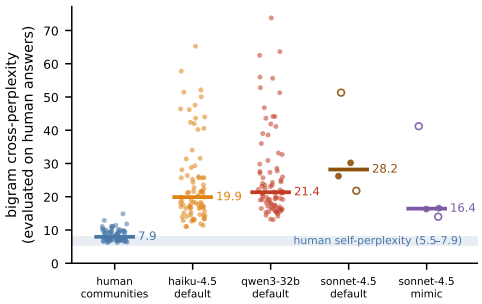


Figure 5: LLM answering strategies diverge from every human community (§4.3). Each point is the perplexity of a bigram model over discourse act sequences from one source, evaluated on human answers from one community (72 distinct human community pairs, 81 pairs per default LLM, and 4 per `sonnet-4.5` condition where 2 open points are cross-community and 2 filled are within-community). Lines are medians; the band spans human within-community self-perplexity. All LLM settings diverge from human communities more than communities do from one another; community guidelines ("mimic") narrow but do not close the gap.

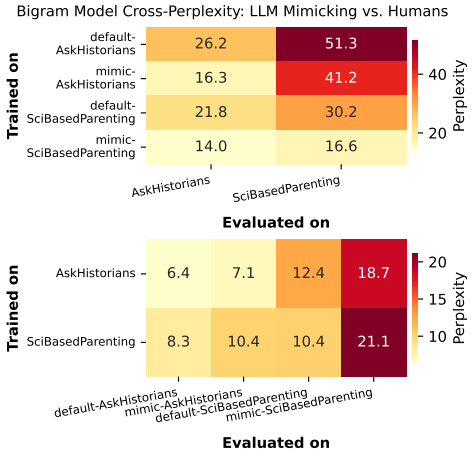


Figure 6: Prompting `claude-sonnet-4.5` with the community guidelines to answer questions as a member of two subreddits `AskHistorians` and `ScienceBasedParenting` (`mimic`) is insufficient to match the norms of the underlying community. Human answers are poorly predicted by bigram models trained on LLM answers (top), while LLM answers are far more predictable under human bigram models (bottom). This asymmetry indicates that LLMs deploy a narrower range of strategies, one that falls largely within (but does not cover) the human strategy distribution.

Adapting to a Community. In the default setting, LLMs are unaware of the community addressed by a question. While analyzing the default discourse structures of LLMs helps us model a straightforward user-LLM interaction without any intervention, it does not necessarily reveal whether LLMs have the capacity to mimic a certain community's answering "policy" when prompted to do so. To measure whether this absence of context inhibits rhetorical diversity in LLM answers, we provide a more powerful closed-source model (`claude-sonnet-4.5`) the community guidelines of two of the most distinctive communities in our dataset, `AskHistorians`[10] and `ScienceBasedParenting`, and ask it to answer as if it were a subreddit member (Prompt G.1). We also prompt it in the default setting for direct comparison. Mimicking `AskHistorians` this way drops perplexity from 26.2 to 16.3, but human self-perplexity in this community is 7.9 (similar in `SciBasedParenting`, Fig. 4a): a sizable gap remains even after the LLM is given explicit, detailed instructions on how to answer. Notably, these guidelines already prescribe behavior at the discourse level (avoid hedging and speculation, suppress fillers, write complete rather than partial answers), yet the **model does not translate these behavioral specifications into the rhetorical patterns of community answers.** This suggests that community rhetorical norms run deeper than what answering instructions can encode: the lack of rhetorical diversity is not merely an issue of context availability, but may reflect a limitation in LLMs' pragmatic reasoning about user information needs (Taylor, 1962) when constructing an answer.

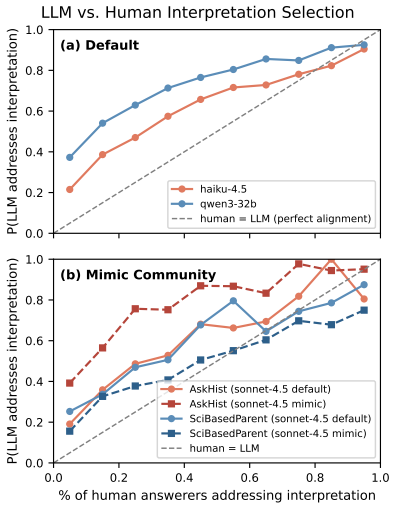


Figure 7: LLMs address low probability interpretations that humans do not address (top). This behavior can get *worse* when it mimics answerers in a specific community (bottom, `AskHist` default to mimic).

5 Interpretation Selection

Answering a question involves pragmatic reasoning about the asker's intent (Grice, 1975) and *informa-*

[10] https://www.reddit.com/r/AskHistorians/wiki/rules/#answers

prompt mode	answerer	mean	median
—	human	0.62	0.53
default	haiku-4.5	0.46	0.40
	qwen3-32b	0.40	0.31
default	claude-sonnet-4.5*	0.48	0.40
mimic	claude-sonnet-4.5*	0.41	0.33

Table 1: Mean/median proportion of eligible segments dedicated to each interpretation. Humans dedicate more segments per interpretation, producing more focused answers. * indicates the model was evaluated only on AskHistorians and ScienceBasedParenting (the two "mimic" communities).

tion need to provide maximally relevant information (Sperber and Wilson, 1986). DISCOTRACE not only enables the analysis of discourse *acts* as in §4, it also reveals the *content* choices of answerers by matching *interpretations* of the question with discourse acts. This lets us compute metrics such as interpretation overlap between two answers to the same question, the proportion of an answer dedicated to an interpretation, and the discourse act type used to address it. Our analyses reveal that LLMs deploy narrower, uniform answering policies than humans, but do they align in their choices of *interpretation*?

Humans. Humans in some communities address more of the interpretation space S_I of a question than others (Table 6): explainlikeimfive and AskHistorians answers cover 40-45% of S_I, versus only 34% in NoStupidQuestions. This illustrates that the proportion of interpretations addressed in an answer can be influenced by the objectives and preferences of a community.

LLMs. We study interpretation selection of the two LLMs from §4.3. For each interpretation $i \in S_I$ across questions from all communities, we (1) compute the proportion of human answers addressing i, (2) bin interpretations by this frequency, and (3) plot the average probability of an LLM addressing i per bin (Fig. 7a). The diagonal indicates perfect alignment with human interpretation preferences. Placement above it indicates overanswering: addressing interpretations that humans rarely address. With default prompting, both LLMs systematically overaddress: even when fewer than 10% of humans address i, haiku addresses it over 20% of the time and qwen 40%. Instead of reasoning about exact asker information need, LLMs opt to hedge and deluge the user with information.

Using the same DISCOTRACEs from sonnet-4.5 mimicking community members (§4.3) shows that mimicking does not help, and can even worsen the pattern (Fig. 7b, AskHist). Computing the average proportion of eligible discourse acts in an answer devoted to an interpretation i reveals that humans favor depth over breadth (Table 1), opting to address fewer interpretations more exhaustively. This pattern persists across communities with different rhetorical strategies, suggesting it is not simply a stylistic preference, but may be a core feature of human pragmatic reasoning when answering a question. LLM-generated answers are *systematically lengthier* (Table 7), corroborating prior work (Poddar et al., 2025), and additional length is the cost of addressing interpretations that the asker likely didn't intend. Future work must strengthen LLM pragmatic reasoning ability in QA to, like humans, hone in on likely interpretations that matter.

6 Conclusion

QA is not one size fits all. DISCOTRACE reveals that communities take different strategies to answer questions, driven by different preferences, needs, and norms, and that LLMs do not answer with the full rhetorical breadth of humans. More broadly, DISCOTRACE turns rhetorical strategy from an implicit property of answers into an explicit representation that systems can use to plan. It can characterize how humans respond to ambiguity (surfacing it versus committing to an interpretation). It can also support preference-sensitive answering: an LLM could generate a DISCOTRACE as a high-level plan conditioned on a user's needs, then populate content accordingly. For example, a consumer health assistant (Srikanth et al., 2024) that maximizes simplicity and minimizes harm might prioritize ASSERT ANSWER with PROVIDE REASONING and routinely RECOMMEND EXPERT; the same planner serving clinicians would instead favor CITE EXTERNAL SOURCE and SURFACE INTERPRETATIONS, with less PROVIDE BACKGROUND, since expertise can be assumed. DISCOTRACE can also disentangle preference judgments that may conflate strategy with execution (Hu et al., 2023): fixing a target DISCOTRACE across candidate answers isolates execution quality (clarity, accuracy, prose), while varying the plan isolates preferences over strategy itself. Finally, DISCOTRACE could also be used to study internal model representations of answering strategies.

Acknowledgments

We thank the anonymous reviewers as well as the members of the University of Maryland CLIP lab for their thoughtful feedback. We also thank our annotators: expert linguists from Upwork, as well as Maya Srikanth and Rupak Sarkar. This work was supported by NIH Award No. R01MD016037 (Srikanth and Boyd-Graber) and NSF CAREER Award No. 2339746 (Rudinger). Boyd-Graber is also supported by a gift from Adobe Corporation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Limitations

While this work studies answers across a variety of online communities, questions and answers are restricted to those from Reddit: we do not include QA pairs from other platforms such as Quora or StackOverflow. Though different subreddits have different answering strategies, these may be affected by the platform as a whole, and are not necessarily representative of humans across all contexts and platforms. All questions and answers in our study are in English, and our discourse act ontology was developed on English Reddit data. Whether the ontology and our findings transfer to other languages or answerers from diverse cultural and linguistic backgrounds and platforms is an open question.

Our experiments do not use the body of Reddit posts as extra context, similar to other datasets such as ELI5 (Fan et al., 2019) and Stanford Human Preferences 2 (SHP-2) (Ethayarajh et al., 2022). Beyond precedent, we opt to do this to ensure a fair comparison across questions, since some posts contain no body text and others have several paragraphs. We also study answers from top-level comments only, ignoring follow-up replies or multi-turn exchanges where the original asker may have refined their question or information need, possibly influencing the strategies of other answerers on the thread.

Because we filter answers by upvotes (Appendix A), the human answers we study are those a particular community *rewarded*, not the full range answerers *attempted*. Our characterization of human strategies is therefore best read as a characterization of *community-endorsed* strategies. This aligns with our aim of modeling community answering norms, but it means low-scoring strategies (which may include reasonable moves a community simply disfavors) are underrepresented, and that our human–LLM comparisons measure distance from *endorsed* norms rather than from *all* human answering behavior.

Lastly, our community-mimicking experiment covers two communities, and a full cross-community grid would more comprehensively test whether context can recover human strategy diversity.

References

Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems*, 35:38176–38189.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. 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 6424–6439, Online. Association for Computational Linguistics.

Paula CF Cardoso, Thiago AS Pardo, and Maite Taboada. 2017. Subtopic annotation and automatic segmentation for news texts in brazilian portuguese. *Corpora*, 12(1):23–54.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Elena Chistova. 2024. Bilingual rhetorical structure parsing with large parallel annotations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9689–9706, Bangkok, Thailand. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3558–3567.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. DecipherPref: Analyzing influential factors in human preference judgments via GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8344–8357, Singapore. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009, Singapore. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol-Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. 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 3932–3945.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2025. ComPO: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on*

Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. Qudsim: Quantifying discourse similarities in llm-generated text. In *Second Conference on Language Modeling*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2025. Brevity is the soul of sustainability: Characterizing llm response lengths. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21848–21864.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

John Searle. 2014. What is a speech act? In *Philosophy in America*, pages 221–239. Routledge.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Renee Shelby, Fernando Diaz, and Vinodkumar Prabhakaran. 2025. Taxonomy of user needs and actions. *arXiv preprint arXiv:2510.06124*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Neha Srikanth, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. No questions are stupid, but some are poorly posed: Understanding poorly-posed information-seeking questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3182–3199.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elias Stengel-Eskin, Jimena Guallar-Blasco, and Benjamin Van Durme. 2021. Human-model divergence in the handling of vagueness. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 43–57.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Robert S Taylor. 1962. The process of asking questions. *American documentation*, 13(4):391–396.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Polina Tsvilodub, Michael Franke, Robert D Hawkins, and Noah D Goodman. 2023. Overinformative question answering by humans and machines. *arXiv preprint arXiv:2305.07151*.

Yini Wang, Xian Zhou, Shengan Zheng, Linpeng Huang, Zhunchen Luo, Wei Luo, and Xiaoying Bai. 2025. Uncovering argumentative flow: A question-focus discourse structuring framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8241–8259, Suzhou, China. Association for Computational Linguistics.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. Qudeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3556–3572.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *proceedings of the international AAAI conference on web and social media*, volume 11, pages 357–366.

Michael Zhang, W. Bradley Knox, and Eunsol Choi. 2025. Modeling future conversation turns to teach llms to ask clarifying questions. In *International Conference on Learning Representations*, volume 2025, pages 60722–60742.

Relations
Elaboration, Attribution, Joint, Same-Unit, Explanation, Enablement, Background, Evaluation, Cause, Contrast, Temporal, Comparison, Topic-Change, Manner-Means, Textual-Organization, Condition, Summary, Topic-Comment

Table 2: RST-DT relations (https://huggingface.co/tchewik/isanlp_rst_v3/blob/rstdt/relation_table.txt.)

Relation	Nuclearity
Contrast	NN
Comparison	NN
Topic-Change	NN, NS, SN
Evaluation	NS, SN, NN
Summary	NN, NS, SN
Background	NS, SN

Table 3: Boundary relations used for RST-based discourse segmentation.

A Data Filtering

We use the PushShift[11] dump of our nine subreddits, which contains posts till the end of 2022. Our study is concerned with information-seeking questions and their answers, so we apply a series of filters accordingly to identify those posts.

Title-Based Filtering. We filter out posts with no title, less than five upvotes, and less than four tokens. To identify *information-seeking* questions instead of community-seeking or other commentary, we use surface form features of post titles, filtering out posts whose titles do not contain a *wh*-question word or that do not end with a question mark. We also remove posts with Reddit-specific terms (*subreddit*, *redditor*, *upvote* or *karma*) and posts that have multiple questions or statements in the title. We run a profanity and NSFW detection model (https://github.com/dimitrismistriotis/alt-profanity-check) on post titles, removing posts above probability 0.8. We also filter out posts that contain first-person pronouns, relationship terms (e.g. *my husband*), deictic references combined with first-person pronouns, or validation seeking phrases ("is this normal", "does anyone else", "is this okay"). These filters help narrow our pool to *information-seeking* questions.

Thread and Comment-Based Filtering. We want to ensure that our dataset has high quality comments, since we are studying answerer rhetorical strategy. We only consider top-level comments of a thread, and remove top-level comments with less than three upvotes. We filter out posts with more than `max` comments to avoid threads with unusually high engagement, which tend to be controversial, opinion-based, or discussion-oriented rather than information-seeking. We also filter out posts with less than `min` comments since we want enough answers for the same question to compare different strategies that answerers may use. We set `min_comments` to 5 for most subreddits and 4 for `ScienceBasedParenting` and `max_comments` to 12 for `AskHistorians`, `history`, `OutOfTheLoop`, and `beyondthebump`, 15 for `asklinguistics` and `explainlikeimfive`, 18 for `NoStupidQuestions`, 20 for `ScienceBasedParenting`, and 30 for `AskEconomics`.

Sampling. After filtering posts from each subreddit according to all the filters above, we sample 300 questions and their top-level comments from each, yielding 2,700 questions and 18,968 answers.

B Preprocessing Answers with Rhetorical Structure Theory

Rhetorical structure theory (Mann and Thompson, 1987), or RST, explains the coherence of a text by expressing how components of a text are related to each other. RST represents a text t as a binary tree. The root represents the whole text. Leaves are *elementary discourse units* (EDUs), which are clauses in t. Leaves are recursively connected through intermediate nodes which express the relation *between* children (Table 2) as well as the distinction between the primary ("nucleus") and supporting ("satellite") child (Fig. 2).

Coarsening EDUs to produce action segments. Elementary discourse units in RST are often sentence *clauses* (see Figure 2) and hence are too short to serve as segments that we can label with a higher-level discourse relation to the question. We opt to merge EDUs that we suspect are part of the same discourse act to form coarser segments that we label with an act. To do this, we define *boundary relations* (Table 3): a subset of the full set of RST relations (Table 2) that are likely to signal a rhetorical shift between a node's children. Boundary relations are defined as relation–nuclearity pairs, since the same relation (e.g., *Evaluation*) may signal a shift under some nuclearity configurations but not others.

Then, we execute the procedure detailed in Al-

[11] https://github.com/Watchful1/PushshiftDumps

Subreddit	Description
AskHistorians	Users ask questions about history and receive detailed answers from knowledgeable historians.
NoStupidQuestions	Users casually ask questions about general knowledge or any topic and receive answers.
AskEconomics	Users ask questions about economic theory, research, and policy and receive answers grounded in economic theory and empirical research.
asklinguistics	Users ask questions about linguistics and receive answers from knowledgeable linguists.
history	Users discuss historical topics.
OutOfTheLoop	Users ask questions about current events, pop culture, or internet trends they feel out of the loop on and receive answers.
ScienceBasedParenting	Users ask questions about parenting and discuss research and science-based guidance.
beyondthebump	Users discuss pregnancy, childbirth, and early parenting experiences.
explainlikeimfive	Users ask questions about any topic and receive answers that simplify complex concepts in a way that is accessible for laypeople.

Table 4: Descriptions of the subreddits included in our dataset.

gorithm 1 using these boundary relations. Starting at the root, we recursively traverse the RST tree to decide where to split. At an intermediate node v, we first recurse into both children to obtain a list of text spans L and R. If the relation–nuclearity pair of v is a boundary relation, we split v's children into separate segments, except for the *Background* relation, where we require both sides to be sufficiently large ($\geq k$ EDUs) to avoid creating trivially small segments from brief contextualizing clauses. If the size check fails, the entire subtree is collapsed into a single span. Even if a relation is not a boundary relation, we still check whether either child was split deeper in the tree. If so, we preserve those deeper splits, but otherwise, we collapse the subtree into one span. Our procedure produces a list of text spans, referred to as **action segments**, which we suspect correspond to distinct discourse acts that we label in §3.2.

Algorithm 1: RST-Based Discourse Segmentation

Input: RST tree node v, boundary relations $\mathcal{B}$, minimum span size $k = 3$
Output: List of EDU spans (each span is a list of EDUs)

Function GetSpans(v, $\mathcal{B}$, k):
  if v *is a leaf* **then return** $[[v]]$;
  $L \leftarrow$ GetSpans(v.*left*, $\mathcal{B}$, k);
  $R \leftarrow$ GetSpans(v.*right*, $\mathcal{B}$, k);
  if v.*relation*_v.*nuclearity* $\in \mathcal{B}$ **then**
    // $\|L\|, \|R\|$ = total number of EDUs in L, R
    if v.*relation* $\neq$ *Background* ***or*** $(\|L\| \geq k$ ***and*** $\|R\| \geq k)$ **then return** $L + R$;
    else return $[$GetLeaves(v)$]$;
  end
  // Not a boundary relation; preserve any splits from deeper in the tree
  if $|L| + |R| > 2$ **then return** $L + R$;
  return $[$GetLeaves(v)$]$;

// GetLeaves(v) returns all leaf EDUs of subtree v as a single flat list

C Discourse Act Ontology Construction

We developed our ontology iteratively, starting from the six sentence-level roles of Xu et al. (2022). These roles were designed to capture how answers *organize* information. However, our goal is to characterize the *function* each segment serves with respect to an answerer's rhetorical *strategy*. Labeling RST-derived segments from the Reddit community r/NoStupidQuestions with these roles, we found that, from a strategy perspective, roles like *summary* and *example* are realizations of a broader "answering" move, while their *organization* role spans both framing and substantive acts. Additionally, their role set did not include representation for the interrogative acts we observe in r/NoStupidQuestions. We therefore restructured these roles into functional families (answer question (AQ), seek information (SI), comment on the question (CQ), redirect the question (RQ), and no-op (NO)), and then repeatedly sampled new QA pairs, adding acts to each of these families when no existing act was appropriate or merging acts together until ten consecutive pairs were sampled without any modifications. We then stress-tested on other communities (e.g. AskHistorians) which yielded further acts (e.g. Cite External Source) and surfaced overloaded acts that required splits (e.g. Counter-Question from Redirect Formulation). Again, we randomly and uniformly sampled QA pairs from our nine different Reddit communities (§4) until ten consecutive samples required no ontology modifications. We treated all pairs used for the development of our ontology as "training samples" and discard them from any other experiments and analysis we do throughout the work.

D Discourse Act Labeling

We label the segments produced from Algorithm 1 using `gpt-4.1-2025-0414` using Prompt D.1.

Prompt D.1: Discourse Act Tagging

System: You are an expert discourse analyst trying to understand how people answer questions on Reddit. You will analyze answers by tagging each text segment from a Reddit answer with a discourse action.
Rules
1. **Select EXACTLY ONE action_id per segment or subsegment.**
2. If the current segment continues the previous action, reuse the previous action_id.
3. If a new rhetorical move begins, select the appropriate new action_id.
4. If no action fits, use "NONE".
Subsegment Labeling
Each segment you receive was produced by a discourse parser. You will also be shown the subsegments (sentences) that make up the segment. If all subsegments serve the same discourse function, return a single-element array with one action_id. If different subsegments serve **different discourse functions**, return an array with one entry per subsegment, each with its subsegment_index and action_id.
Common patterns worth splitting:
- Background/reasoning subsegments followed by an answer subsegment
- An answer subsegment followed by a redirect or recommendation
- A presupposition rejection followed by an alternative answer

Do NOT split when:
- A subsegment contains light framing for the next (e.g., "So basically," followed by an answer → single Assert Answer)
- The difference is just emphasis vs. substance within the same move

Caveats and Task Nuances
1. Consider the expected answer type of the question when labeling actions. Responding to "Where can I find X" with a website recommendation is an "Answer the Question" action, not a "Direct to Resource" action. If the resource is the answer itself, label it as "Assert Answer". If the resource is suggested as additional reading, use Direct to Resource.
2. When a segment contains both an answer and supporting reasoning: if subsegments are provided and the answer and reasoning fall in **different subsegments**, split them. If they are in the **same subsegment** (tightly integrated), label it as "Provide Reasoning or Justification" if the justification is non-trivial, otherwise "Assert Answer".
3. When a segment explains WHY something is the case, determine what it is explaining:
- If it explains why an answer is correct → "Provide Reasoning or Justification"
- If it explains why a premise of the question is wrong → "Reject Presupposition"
Example: For "Why is the sky blue?", the segment "Because of Rayleigh scattering" is justification. For "What's the best liver detox cleanse?", the segment "The concept of 'detoxing' your liver is misleading—your liver already filters toxins continuously" is rejecting the presupposition.
4. Sharing a personal anecdote or experience is "Provide Example", NOT "Provide Background." Background sets up context, frameworks, or history before answering. Examples use concrete cases (including personal ones) to support or illustrate an answer.
- "I have a doctorate, and sometimes introduce myself as Dr." → Provide Example
- "The use of honorifics has a long and contested history in academia." → Provide Background
5. When a segment follows a recommendation and provides supporting information, ask: does it explain why the recommendation is good in terms of the original question, or does it answer a different question?
- If it explains why the recommendation addresses the original question → "Provide Reasoning or Justification"
- If it introduces new information that answers a tangentially related but different question → "Answer a Question or Interpretation outside of Interpretation Space"
6. When a segment invokes an external source (study, statistic, law, quote, expert consensus) to support a claim, use "Cite External Source"—NOT "Provide Example" or "Provide Reasoning." The key test: does the credibility derive from an independently verifiable external source, or from the answerer's own experience/logic?
- "A 2019 Lancet study found no significant effect." → Cite External Source
- "I saw the same thing happen at my last job." → Provide Example
- "That's because the compiler needs type info at compile time." → Provide Reasoning

Action Ontology
{ontology}
Output Format
Always respond with ONLY a JSON array. No explanation, no reasoning, no commentary.
Single action for whole segment:
[{"action_id": "action_AQ_assert_answer"}]

Distinct actions per subsegment:
[{"subsegment_index": 0, "action_id": "action_CQ_reject_presupposition"},
{"subsegment_index": 1, "action_id": "action_AQ_assert_answer"}]

When no action fits:
[{"action_id": "NONE"}]
User:
Question
{question}
Full Answer
{answer}
Previous Segment action="{prev_label}"
{segment_prev}

Subreddit	Segments	NONE	NONE Rate
AskHistorians	10,619	332	3.13%
explainlikeimfive	9,190	224	2.44%
beyondthebump	6,950	75	1.08%
NoStupidQuestions	5,108	51	1.00%
asklinguistics	6,701	148	2.21%
ScienceBasedParenting	11,255	189	1.68%
history	7,442	163	2.19%
OutOfTheLoop	8,205	190	2.32%
AskEconomics	8,511	308	3.62%
Total	73,981	1,680	2.27%

Table 5: Discourse act NONE rates by community. Low rates across all nine communities indicate that our ontology covers the discourse acts present in answers. We inspect NONE segments, and attribute them to RST parsing errors rather than ontology gaps (§C).

```
Current Segment
{segment}
Subsegments
{subsegments}
Respond with ONLY a JSON array.
```

We observe fairly low usage rates of the label NONE by our automatic act labeler. These rates, stratified across our different communities, are shown in Table 5.

E Annotation Guidelines

The annotation guidelines shown to annotators are below. Annotators are not shown any PII: they are not shown the post ID or the user ID of the asker or the commentor. The only information they are shown is the name of the subreddit in which the question is asked. Both annotators have graduate degrees in Computer Science or linguistics and are native English speakers. Annotators were recruited from Upwork and through the department.

E.1 Overview

There are many ways to respond to a question when asked. An answerer may choose to ask a follow-up question, probe for some more context, or simply answer by making their best guess about the information needs of the user. Our goal is to identify the communicative moves an answerer makes and, where applicable, which interpretation of the question each move addresses. Each annotation instance consists of a question Q, an answer A that has been pre-segmented into text spans, and (optionally) a set of candidate interpretations of Q. Your task is to assign one discourse act label from the ontology (Fig. 3) to each segment, and for eligible acts, to indicate which interpretation the segment addresses.

E.2 Task Description

For each segment in the answer:

1. Read the segment in the context of the full answer and the original question Q.

2. Assign exactly one discourse act label from the ontology. If a segment appears to serve multiple functions, choose the primary or dominant one.

3. If the assigned act is marked as interpretation-eligible (starred), select the interpretation from the provided interpretation space that the segment most closely addresses. If no listed interpretation fits, mark it as "other."

E.3 Discourse Act Ontology

We provided annotators with the same content as in Figure 3 as well as a handful of examples.

E.4 Interpretation Labeling

Questions often admit multiple valid readings. For example, "Can language be described by math?" could be interpreted as asking about formal language theory, computational linguistics, or the philosophy of mathematical description. When a discourse act is interpretation-eligible (starred), you must select which interpretation from the provided set the segment addresses. If a segment addresses multiple interpretations, choose the primary one. If it does not clearly map to any listed interpretation, mark NONE. Acts that are NOT interpretation-eligible (e.g., Reject Presupposition, Presentational, Non-Answer) engage with the question's framing or serve a structural role rather than adopting a particular reading. Do not assign an interpretation to these.

E.5 Decision Rules and Caveats

- Reject Presupposition vs. Provide Reasoning: If the segment explains why a premise of the question is wrong, use Reject Presupposition. If it explains why an answer is correct, use Provide Reasoning or Justification.

- Provide Background vs. Assert Answer: Background situates the question within a broader context without directly answering. If the segment makes a direct claim that resolves the question, prefer Assert Answer.

- Cite External Source vs. Provide Example: If credibility derives from the authority of the

cited source, use Cite External Source. If a concrete case is offered to illustrate a point (including personal experience), use Provide Example.

- Cite External Source vs. Direct to Resource: If the answerer uses content from a source to support their claim in the answer, use Cite External Source. If they redirect the asker to go read or consult the resource themselves, use Direct to Resource.
- Comment on Question vs. Presentational: If the remark engages substantively with the question's nature or difficulty, use Comment on Question. If it is purely filler ("Great question!"), use Presentational.
- Redirect Formulation vs. Answer outside Interpretation Space: If the segment explicitly proposes a better way to ask the question, use Redirect Formulation. If it drifts to answering a related but different question without flagging the reframe, use Answer outside Interpretation Space.
- Counter-Question vs. Clarification: Counter-Questions challenge the premise or assumptions of the question. Clarifications simply request more information to better answer it.

F Interpretation Labeling

Prompt F.1: Interpretation Generation

System: Users in a question answering community typically try to express a need for information through a question. Sometimes, from the language of their question alone, it is not clear what their exact information need is. This leads to many distinct interpretations of their question, each representing different information needs. You will be given a **question** asked in a specific online community that may have many distinct interpretations. Your task is to output those interpretations as unambiguous distinct questions.
{subreddit_context}
Critical Rules
- Each interpretation must be a **different plausible reading** of the SAME question — a different thing the user could have MEANT by their words.
- Do NOT generate sub-questions, follow-up questions, related questions, or questions that explore different aspects of the topic.
- Ask yourself: "Could the user have typed this exact question while meaning THIS?" If the answer is no, it is not a valid interpretation.
- Interpretations should differ in WHAT the user is asking, not provide additional angles on the same clear question.

If the user's information need is already clear from their question, output 'NONE'. Otherwise, output the numbered list of interpretations as unambiguous questions and nothing else.
User:
Question
{question}

Prompt F.2: Interpretation Labeling

System: You are an expert discourse analyst. A Reddit answer segment has already been labeled with a discourse action. Your task is to determine which interpretation of the original question the segment best addresses.
Rules
1. You are given a question, its possible interpretations, and a segment from an answer that has been labeled with a discourse action.
2. Determine which question interpretation the segment most directly addresses, adopts, or targets. This may be explicit or implicit.
3. If the segment clearly and directly addresses one of the interpretations, return that interpretation's ID.
4. If the segment does not clearly target any specific interpretation, return "NONE".

Output Format
Respond with exactly ONE JSON object:
[{"interpretation_id": "id_1"}]

When no specific interpretation is targeted:
[{"interpretation_id": "NONE"}]

User:
Question
{question}
Question Interpretations
{interpretations}
Full Answer
{answer}
Segment (labeled as "{action_label}")
{segment}
Respond with EXACTLY ONE JSON dictionary (NOT an array).

G Community Mimicking

Prompt G.1: Mimicking Community Guidelines

System: r/{subreddit} is a subreddit for {subreddit_explanation}. The community guidelines for r/{subreddit} are as follows: {community_guidelines}.
User: Answer the question as if you were a redditor in that subreddit: {question}

G.1 r/AskHistorians Community Guidelines

Answers in r/AskHistorians are held to a higher standard than is generally found on Reddit. Answers should be in-depth, comprehensive, accurate, and based off of good quality sources. Answers should be in line with the existing historiography on the topic, and written in a manner respecting the Historical Method.

Write an in-depth answer. An in-depth answer provides the necessary context and complexity that the given topic calls for, going beyond a simple cursory overview. "Good answers aren't good just because they are right — they are good because they explain." Your answer should give context to the events being discussed, not simply list related facts. Ask yourself: Do I have the expertise needed to answer this question? Have I done research on this topic? Can I cite academic quality primary and secondary sources? Can I answer follow-up questions?

Sources. Answers should reflect current academic understanding or debates on the subject. Both primary and secondary sources are accepted. Secondary literature should be from academic or respected general publishers. Tertiary sources such as Wikipedia may be cited for basic undisputed facts, but sole reliance on tertiary sources is not allowed. Blog posts and random web articles are not acceptable. You are not a source: personal experience and anecdotes are not acceptable.

No personal anecdotes. Personal anecdotes are unreliable, unverifiable, and of very little real interest.

No speculation. Suppositions and personal opinions are not a suitable basis for an answer. Warning phrases include "I guess," "I believe," "I think," "to my understanding," and "it makes sense to me that."

No partial answers or placeholders. An answer should be full and complete in and of itself. Do not post partial answers with the intention of prompting further discussion or as a placeholder to expand on later.

No political agendas or moralising. Answers should represent a sincere effort to make an argument from the historical record, constructed in keeping with the principles of the historical method. Evidence should not be chosen selectively to support a predetermined argument.

Do not just post links or quotations. A good answer is a balanced mix of context, explanation, and sources. Always provide context for any source you cite. Answers should not consist only or primarily of copy-pasted text.

No plagiarism. Directly copying and pasting another person's words and passing them off as your own will result in an instant ban.

G.2 r/ScienceBasedParenting Community Guidelines

Be respectful. Discussions and debates are welcome but must remain civilized. Inflammatory content is prohibited.

Read linked material before commenting. Make sure you know what you are commenting on to avoid misunderstandings.

Respect post flair. All top-level comments must adhere to the flair type guidelines set by the OP. If you reply to a top-level comment with additional or conflicting information, a link to flair-appropriate material is also required.

Sources must match flair type. For "Question — Link To Research Required" flair, top-level answers must link directly to peer-reviewed research published in scientific journals. For "Question — Link to Expert Consensus Required" flair, links to sources containing expert consensus are permitted (e.g. CDC, WHO, American Academy of Pediatrics). Parenting books, podcasts, and blogs are not peer-reviewed and should not be referenced as scientific sources.

No individualized medical advice. General questions are allowed; specific questions about one's own child's symptoms or treatment are not. Nothing posted constitutes medical advice.

Keep comments relevant. All threads and comments must be directly relevant to the discussion. Off-topic threads and comments will be removed.

Community	$\lvert S_I \rvert$ $\mu \pm \sigma$	Unmatched rate	# Interps $\mu \pm \sigma$	Coverage $\mu \pm \sigma$
r/explainlikeimfive	3.1±1.4	0.19	1.3±0.8	0.45±0.25
r/AskHistorians	3.3±1.7	0.25	1.3±0.9	0.41±0.27
r/beyondthebump	3.2±1.5	0.24	1.2±0.7	0.41±0.24
r/ScienceBasedParenting	3.1±1.5	0.32	1.2±0.8	0.39±0.25
r/asklinguistics	3.1±1.6	0.19	1.2±0.7	0.39±0.25
r/OutOfTheLoop	3.4±1.6	0.20	1.1±0.7	0.36±0.24
r/history	3.8±2.0	0.23	1.2±0.8	0.35±0.24
r/NoStupidQuestions	3.6±1.6	0.17	1.1±0.7	0.34±0.22
r/AskEconomics	3.9±1.9	0.25	1.3±0.9	0.34±0.24

Table 6: Interpretation space statistics across communities. Coverage is computed over questions with $\lvert S_I \rvert \geq 2$.

prompt mode	answerer	segments per answer	matched per answer	dedication mean	dedication median
—	human	3.3	2.5	0.62	0.53
default	`haiku-4.5`	5.7	4.6	0.46	0.40
	`qwen3-32b`	12.1	10.4	0.40	0.31
default	`sonnet-4.5*`	5.8	4.8	0.48	0.40
mimic	`sonnet-4.5*`	9.8	8.0	0.41	0.33

Table 7: Average number of eligible segments per answer, segments matched to an interpretation, and proportion of eligible segments dedicated to each addressed interpretation. Higher dedication values indicate more focused answers.

No self-promotion or product promotion. Do not use this subreddit to advertise or sell a product, service, podcast, or book.

General discussion. Questions that cannot be answered by direct research or expert consensus, or requests for anecdotes and parent-to-parent advice, belong in the weekly General Discussion Megathread.

H Computational Resources

All experiments were run on two NVIDIA RTX A6000s. Models greater than 11B parameters were run using 4 bit quantization. All models including closed-source models were run with a temperature of 0.01.

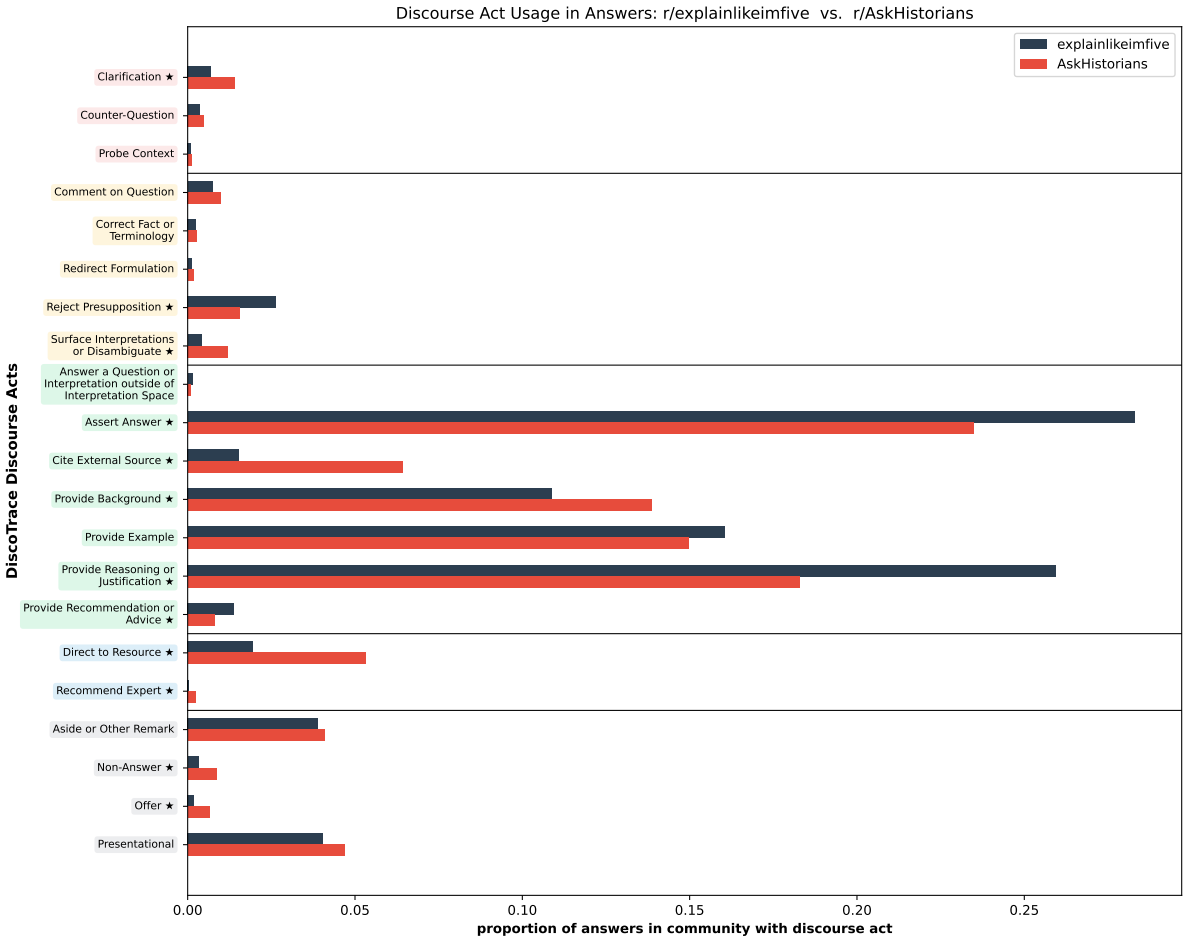


Figure 8: We compare the proportion of answers that address each of the 21 acts in our ontology in `explainlikeimfive` and `AskHistorians`, revealing interesting differences in their act usage. Proportions differences that are statistically significant are starred, and measured using a chi-squared test with Bonferroni correction.

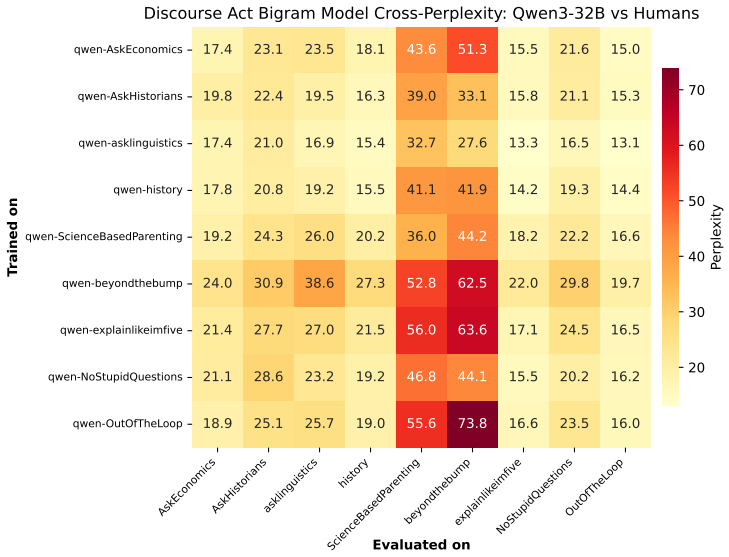


Figure 9: We train bigram models on *act* sequences from one community C_{train} (rows) and evaluate on another C_{eval} (columns). The diagonal compares the LLM's answers against human answers in the same community. Here we train a bigram model on `qwen3-32b` answers to questions in C_{train} and evaluate on human answers.

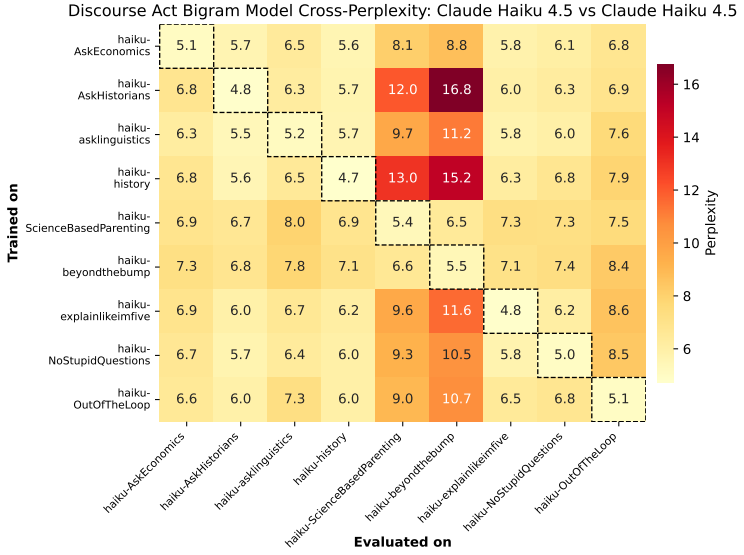


Figure 10: We train bigram models on *act* sequences from one community C_{train} (rows) and evaluate on another C_{eval} (columns). The diagonal is the self perplexity. Here we train a bigram model on `claude-4.5-haiku` answers to questions in C_{train} and evaluate on `claude-4.5-haiku` answers.

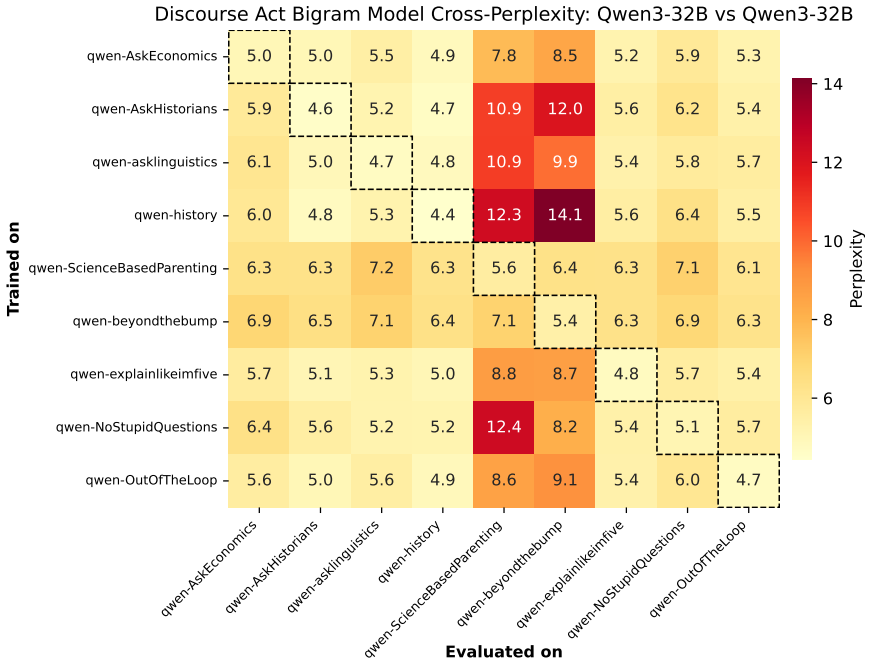


Figure 11: We train bigram models on *act* sequences from one community C_{train} (rows) and evaluate on another C_{eval} (columns). The diagonal is the self perplexity. Here we train a bigram model on `qwen3-32b` answers to questions in C_{train} and evaluate on `qwen3-32b` answers.