QUILT: Query–User Interaction Linguistic Taxonomy

Yuyang Dai
Florida State University
y9657422@gmail.com

Abstract

Large language models serve people with different backgrounds and goals, but analyses of their use commonly emphasize task categories over the expertise expressed in individual queries. We introduce QUILT, a two-dimensional framework that jointly characterizes task type and query-level expertise from observable language. Applying QUILT to 3,000 WildChat dialogues, we construct a cross-domain annotated corpus. The labels reveal systematic associations: debugging queries are disproportionately labeled Expert, Novice labels are uncommon in reasoning and planning tasks, and Expert-labeled queries are longer within each task category. These associations describe linguistic evidence rather than verified user competence or causal effects. Automatic classifiers recover both dimensions from first-turn queries. A downstream response-adaptation evaluation further compares unconditioned, expertise-only, and joint task–expertise conditioning; the mean scores favor joint conditioning on terminology, knowledge calibration, and response depth. We provide annotation decision rules, annotation prompts, within-task robustness analyses, and a discussion of how the resource can support interpretable, expertise-aware evaluation and personalization.

1 Introduction

Large language models (LLMs) are used for information seeking, problem solving, and content generation across diverse user populations. Public conversation datasets such as LMSYS-Chat-1M (Zheng et al., 2024) and WildChat (Zhao et al., 2024) make these interactions available for systematic study. Yet superficially similar requests can call for different responses. A person encountering a concept for the first time may need definitions and concrete examples, whereas a query expressing substantial domain knowledge may call for assumptions, trade-offs, or a concise technical argument.

Task type describes the user's objective; query-level expertise describes presupposed knowledge. Consider two queries that both display advanced programming vocabulary: repairing an existing implementation and designing a new implementation require different response structures. Conversely, two requests for the same explanation may differ in the terminology and intermediate steps that would make an answer useful. Expertise-only conditioning cannot represent the first distinction, while task-only conditioning cannot represent the second. Their combination provides an interpretable description of the response requirements.

Existing work has identified recurring interaction patterns, including information seeking, content creation, and instruction following (Ouyang et al., 2023; Trippas et al., 2024). Work on personalization and user modeling studies how systems estimate or adapt to individual differences (Chen et al., 2024; Thu and Kocaballi, 2025). Closely related research examines expertise alignment between users and assistants and its relationship to user experience (Palta et al., 2025). These studies motivate a reusable annotation resource that connects task intent with expressed expertise across domains, without treating language as a verified profile of the person producing it.

We introduce QUILT (Query–User Interaction Linguistic Taxonomy), a two-dimensional annotation framework applied to the first user turn of a conversation. It assigns a primary task type and one of three expertise levels using only observable linguistic evidence. Expertise is domain-relative and query-specific: the same person can express advanced knowledge in one interaction and seek introductory help in another. The label is an operational description of the query, not a persistent judgment about its author.

We apply the framework to 3,000 WildChat dialogues through LLM annotation, targeted escalation, and human review. The resulting joint distri-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3018–3035
November 6-10, 2026 ©2026 Association for Computational Linguistics

bution contains structure that marginal summaries conceal. Expert labels form the plurality in debugging, whereas Novice labels are uncommon in reasoning and planning. Query-length differences persist within task types. Because the expertise labels and linguistic measurements derive from the same text, these findings are associational; they do not establish that underlying competence causes a particular query style.

Our contributions are threefold. We provide a cross-domain annotation framework with explicit boundary rules for task intent and expressed expertise. We construct and analyze a 3,000-dialogue corpus, including within-task language statistics and an analysis of the scope imposed by English-only filtering. We evaluate automatic classification and downstream response calibration, supplemented by an expertise-only ablation, annotation prompts, and a discussion of the resource's reproducibility and limitations.

2 Related Work

Analysis of real-world LLM interactions. LMSYS-Chat-1M and WildChat support empirical study of conversational behavior at scale (Zheng et al., 2024; Zhao et al., 2024). Task-oriented analyses identify recurring uses of assistants (Ouyang et al., 2023), while interaction studies examine experience and information-seeking behavior (Wang et al., 2024; Trippas et al., 2024). These resources provide the setting for QUILT, but task labels alone do not describe the background knowledge expressed within a request.

Intent taxonomies. Classical search taxonomies distinguish informational, navigational, and transactional goals (Rose and Levinson, 2004; Jansen et al., 2008). Conversational requests extend this picture because an assistant may explain, derive, generate, repair, or plan. OpenAI's usage analysis distinguishes practical guidance, information seeking, and writing, as well as asking, doing, and expressing (OpenAI, 2025). The Anthropic Economic Index relates interactions to occupational tasks and distinguishes patterns such as learning, validation, and task iteration (Anthropic, 2025). QUILT adapts recurring distinctions from this literature into annotation decisions at the query level. It does not propose that six categories exhaust all forms of interaction, nor that its labels map one-to-one onto these different taxonomies.

Expertise and response alignment. Palta et al. (2025) analyze 25,000 Bing Copilot conversations across topical domains, relating user–agent expertise alignment to user experience and engagement. Their work demonstrates the relevance of expertise matching. QUILT complements this analysis by operationalizing the interaction between task type and expressed expertise as a joint annotation resource, and by comparing joint conditioning with expertise-only conditioning. Cross-domain coverage by itself is not a distinction from that prior work; the contribution here is the explicit task–expertise framework and its use for controlled comparisons.

Personalization and user modeling. Personalization research considers how LLM systems adapt to user information and preferences (Chen et al., 2024; Thu and Kocaballi, 2025). QUILT addresses a narrower setting in which only the current query is available. It supplies an interpretable, revisable signal for a cold-start interaction. An expertise label should complement an explicit request such as "explain briefly" or "show every step," not override it. The distinction between expressed knowledge and a stable user attribute is central to both the annotation procedure and the intended applications.

3 The QUILT Annotation Framework

3.1 Design Principles

We use three design principles. *Observability* requires that annotations be supported by the query text, without demographic metadata or external profiles. *Mutual exclusivity* requires one primary task label and one expertise label; an Other category handles task requests outside the six substantive categories. *Cross-domain applicability* requires the same decisions to be usable in programming, medicine, education, finance, and everyday application settings.

The two dimensions describe different response requirements. For debugging, a useful answer should inspect the supplied artifact, identify a failure, and propose a repair. For creation, it should deliver a new artifact. Within either task, expressed expertise informs the level of terminology, assumed prerequisites, and explanatory detail. Combining the labels therefore represents both an objective and a level of exposition. Assigning the dimensions separately does not imply that they will be statistically independent: their association is one of the properties the corpus is designed to measure.

Category	Operational target	Decisive evidence or boundary
Task type: the primary objective of the request		
Information-seeking	Factual knowledge, definitions, or direct explanation.	Expected answer primarily retrieves or explains knowledge.
Reasoning/Analysis	Inference, comparison, or multi-step derivation.	Requires connecting evidence or assumptions to a conclusion.
Creation/Generation	A new text, code, or other usable artifact.	Produces an artifact from scratch rather than repairing input.
Debugging/Problem-solving	Diagnosis, correction, or revision of supplied material.	A prior artifact, error, or failed attempt is present in the input.
Planning/Design	A roadmap, architecture, workflow, or strategy.	Main output organizes actions or decisions rather than delivering the finished artifact.
Opinion/Debate	A preference, recommendation, or evaluative stance.	Multiple reasonable answers depend on context or preferences.
Expertise: background knowledge expressed in the query		
Novice	Requests introductory support or expresses little domain familiarity.	Basic definitions, explicit beginner framing, or misuse of central terms.
Intermediate	Uses domain concepts correctly with a targeted request.	Concrete context, but seeks explanation, validation, or standard practice.
Expert	Presupposes advanced shared knowledge.	Multiple independent signals: unexplained terminology, trade-offs or constraints, and professional or research framing.

Table 1: The two annotation dimensions. Other is a fallback task label. Expertise labels describe observable query language; they do not certify the user's actual competence. The same decision rules apply to requests from different topical domains.

3.2 Task Type: Six Categories

Table 1 summarizes the operational definitions. The categories were refined through a pilot sample and informed by prior interaction and search taxonomies. Information-seeking requests factual content or a direct explanation. Reasoning/Analysis requests a derivation, comparison, or inference. Creation/Generation requests a new artifact. Debugging/Problem-solving requests diagnosis or repair of an existing artifact or concrete failed attempt. Planning/Design requests a strategy, architecture, or sequence of decisions. Opinion/Debate requests an evaluative position for which the answer depends on preferences or context.

The output sought by the user takes precedence over isolated keywords. "Explain" can introduce either a factual answer or a multi-step derivation. "Write" can refer to a finished document or to a strategic plan. The most consequential boundary is the *prior-artifact criterion*: repair or revision requires existing material, an error, or a failed attempt in the input. A request to produce code from scratch is creation even when its wording mentions solving a problem. Appendix A gives the extended boundary rules and examples.

3.3 Expertise: Three Query-Level Labels

Novice queries typically request basic explanations, express uncertainty about elementary concepts, or assume little shared domain knowledge. Intermediate queries use relevant concepts correctly and specify a concrete problem, while seeking explanation, validation, or established practice. Expert queries presuppose substantial shared knowledge and introduce advanced constraints, trade-offs, edge cases, or professional-level problem specifications.

Expert assignment requires multiple independent signals. Dense terminology alone is insufficient, and asking for the definition of a term does not by itself establish low competence. A professional can ask a basic question outside a familiar subfield; a non-specialist can paste sophisticated terminology. The extended guideline uses a conservative Intermediate default when evidence is mixed. At least two signals among unexplained domain terms, explicit constraints or trade-offs, and professional or research framing are required for the conservative Expert decision.

Query length is not an annotation criterion. Nevertheless, a longer query provides more opportunities for expertise-related evidence to appear. Consequently, length and inferred expertise can be associated partly through the measurement procedure. We explicitly preserve this caveat when interpreting the corpus analysis. External validation against independently measured domain knowledge would test a stronger claim about construct validity.

3.4 Why the First User Turn?

The first-turn design describes the information available before the assistant has influenced the dialogue. Later turns can incorporate vocabulary, explanations, and suggestions supplied by the model. Including them would change the target from an initial query-level judgment to an estimate conditioned on an evolving interaction. The design also supports cold-start classification without user history or a persistent identity.

Section 5.4 examines whether group-level linguistic differences persist across turns. Such stability does not establish individual-level reliability. A deployed system should let users correct the inferred level and revise its response strategy as new evidence appears.

3.5 Pilot Study and Decision Workflow

A pilot study on 200 randomly sampled dialogues used three annotators to identify ambiguous boundaries. Inter-annotator agreement increased from $\kappa = 0.58$ to 0.74 after guideline revision. Disagreements particularly concerned Information-seeking versus Reasoning/Analysis, and Intermediate versus Expert. Refinements emphasized the expected reasoning process and the need for multiple expertise signals.

Figure 1 summarizes the decisions and the escalation procedure. For multi-intent queries, annotators select the primary objective rather than assigning several task labels. This convention produces tractable categories at the cost of losing secondary intents. Expertise is then assigned from its own evidence rules rather than mechanically inferred from the task label. The full interface guidance and examples are provided in Appendix A.

4 The QUILT Corpus

4.1 Source, Filtering, and Sampling

We draw from WildChat-1M, a public collection of real user–assistant conversations (Zhao et al., 2024). The source includes conversation text, language metadata, and toxicity indicators. Our sampling pipeline uses the source's English label, requires at least two user turns, excludes records flagged as toxic or redacted, and retains first queries between 10 and 500 whitespace-delimited words. The scan yielded 39,544 eligible conversations from approximately 300,000 inspected records.

We use reservoir sampling with random seed 42 to select 3,000 eligible conversations. We distinguish this corpus-construction step from the stratification used for train, development, and test partitioning. The resulting distribution is imbalanced, with creation accounting for about half the corpus. Filtering out single-turn, very short, long, and non-English interactions also limits which usage patterns the sample represents.

4.2 Three-Stage Annotation

First-pass annotation. The pipeline queries `google/gemini-2.5-flash-lite` with the first user query and the complete annotation guideline. The annotation client uses temperature zero and requests a JSON task label, expertise label, a single confidence score, and a rationale. Confidence is a self-reported annotation score, not a calibrated probability of correctness. Its interpretation may depend on the prompting and scale design (Dai, 2026).

Escalated review. The annotation procedure partitions 3,000 conversations into 2,920 retained first-pass cases and 80 review cases. The selector checks low confidence, the Other label, and ambiguity markers in rationales. Review cases are re-annotated with `google/gemini-2.5-flash`, using the same message construction and guidelines. The two models disagree on 54 conversations, which are forwarded for human review.

Human adjudication. Two trained annotators independently reviewed the 54 flagged cases and resolved remaining disagreements through discussion. The review focuses on the observable evidence supporting each label and on the boundaries identified during the pilot study. Appendix E provides the annotation prompts, output schema, and implementation details.

4.3 Agreement and Coverage

On the 80 escalated cases, Lite–Flash agreement is $\kappa = 0.0504$ for task type and 0.4740 for expertise. Accuracies are 35.0% and 71.25%, respectively. These values refer to a deliberately selected difficult subset and should not be interpreted as agreement on a random sample of the whole corpus. Human inter-annotator agreement on the 54 adjudication cases is $\kappa = 0.84$ for task and 0.83 for expertise. High agreement within adjudication does not validate the much larger automatically retained subset.

The corpus covers all 18 substantive task–expertise combinations, plus 26 Other cases. The

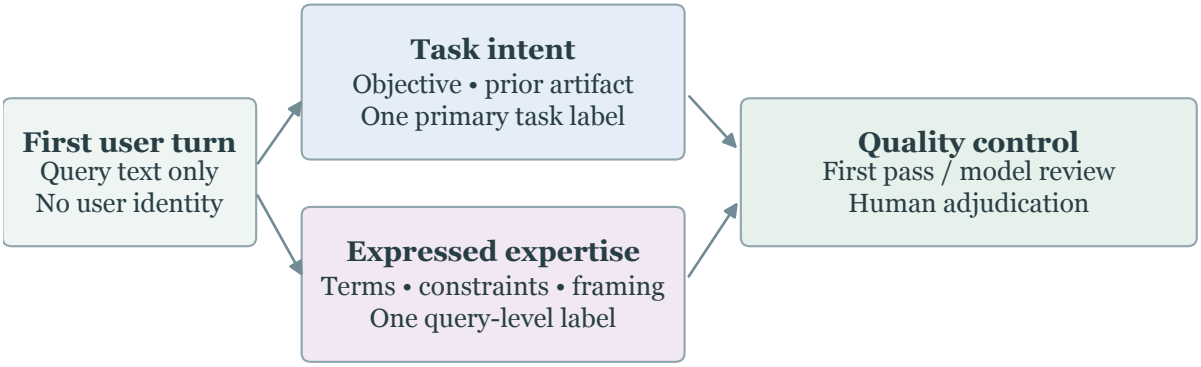


Figure 1: QUILT annotation workflow. Task intent and expressed expertise are separate judgments over the same first user turn. Difficult cases receive model escalation and human review. The two output labels can be associated without being interchangeable.

Comparison	n	Task κ	Expertise κ
Lite–Flash review	80	0.0504	0.4740
Human adjudication	54	0.84	0.83

Table 2: Agreement on the selected review and adjudication subsets. The model and human comparisons involve different populations; neither is an estimate of whole-corpus annotation accuracy.

largest expertise group is Intermediate (1,453), followed by Novice (865) and Expert (682). We use a 70/10/20 partition with 2,102 training, 298 development, and 600 test conversations, stratified by the joint labels. Data splits, class frequencies, and the label set should be held fixed when comparing classifiers; otherwise macro-averaged scores may describe different evaluation problems.

4.4 Cross-Lingual Validation

The English sample represents approximately 13.2% of inspected records under all eligibility filters jointly. To assess transfer beyond English, we apply the same pipeline to Chinese ($n = 120$), Japanese ($n = 95$), and Spanish ($n = 110$). A balanced subset of 50 conversations across these languages is translated into English and independently annotated by a human. Agreement with the pipeline is $\kappa = 0.78$ for task type and $\kappa = 0.76$ for expertise annotation.

Creation remains the largest task category and Intermediate the largest expertise category across the four language samples. Appendix C reports the full distributions and translation-based validation. These initial findings support further multilingual investigation, but do not establish cross-language equivalence. English technical vocabulary and explanatory conventions can differ from those in other languages. The first-query length bounds and multiple-user-turn requirement also restrict generalization: concise expert queries, successful single-turn interactions, and long technical documents may be excluded.

5 Corpus Analysis

We examine associations between task intent, inferred expertise, and linguistic behavior. The unit is the conversation, represented for labeling by its first user query. All references to Novice, Intermediate, and Expert below denote annotation categories. They do not describe verified demographic or professional groups.

5.1 Structure in the Joint Distribution

Figure 2 shows the joint task–expertise distribution. Counts are also tabulated in Appendix B. A chi-square test over the six substantive task categories finds a strong association between task and expertise ($\chi^2 = 364.72$, $df = 10$, $p < 10^{-70}$). The test excludes Other and describes the observed sample rather than the prevalence of expertise in all LLM users.

Debugging is the only substantive task where Expert labels form the plurality: 168 of 345 cases (48.7%), compared with 24 Novice cases (7.0%). By contrast, Novice labels account for 259 of 605 information-seeking queries (42.8%) and only 34 of 349 reasoning queries (9.7%). Intermediate, rather than Novice, is narrowly the largest group in information seeking. Opinion/Debate is the category in which Novice labels form a majority (27 of 52, or 51.9%). In planning, Intermediate and Expert together account for 119 of 130 queries (91.5%).

	Novice	Interm.	Expert
Information	259 8.6%	265 8.8%	81 2.7%
Reasoning	34 1.1%	203 6.8%	112 3.7%
Creation	490 16.3%	737 24.6%	266 8.9%
Debugging	24 0.8%	153 5.1%	168 5.6%
Planning	11 0.4%	71 2.4%	48 1.6%
Opinion	27 0.9%	21 0.7%	4 0.1%

Figure 2: Joint task–expertise distribution. Cells show counts and percentages of all 3,000 conversations; the 26 Other cases are omitted from this six-task view.

These patterns motivate joint analysis: the marginal frequency of Expert labels does not reveal their concentration in debugging. They do not show that a person's competence determines which task they choose. Task conventions can influence the visibility of expertise cues, and the same language supports both annotations. Prior artifacts and technical constraints, for example, may be especially common in repair requests. This measurement dependence remains even when task and expertise follow separate decision rules.

5.2 Query Length Within Task Categories

Across the corpus, first queries average 46.0 words for Novice, 77.5 for Intermediate, and 140.0 for Expert labels. A Kruskal–Wallis test yields a statistic of $H = 416.99$ with $p < 10^{-90}$ and pairwise Mann–Whitney tests with $p < 10^{-32}$. These measurements are descriptive indicators of linguistic differences, not evidence that greater competence causes verbosity.

Task composition alone does not explain the pattern. Table 3 reports word counts within every substantive task, computed across the 3,000 annotated conversations. Expert-labeled queries are longer than Novice-labeled queries in all six categories. The absolute gap is largest in debugging (111.5 words), followed by planning (107.6); reasoning has a large relative gap because its Novice mean is particularly low. Figure 3 visualizes the same within-task comparisons without changing the scale between groups.

Task	Nov.	Int.	Exp.	Δ
Information	26.3	39.3	82.6	56.3
Reasoning	25.1	60.9	112.5	87.4
Creation	58.3	92.4	149.0	90.7
Debugging	61.2	95.5	172.7	111.5
Planning	33.9	72.1	141.5	107.6
Opinion	22.8	58.8	95.0	72.2

Table 3: Mean first-query word counts within task type. Δ is Expert minus Novice, computed from the displayed one-decimal means. All comparisons use the same whitespace-based word-count definition.



Figure 3: Within-task query length by expertise label. The common horizontal axis shows means from the annotated corpus. No length threshold is used to assign expertise.

Stratification controls for differences in the mix of task categories, but it does not remove all confounding. Artifact length, domain, writing style, language proficiency, and the presence of detailed constraints can vary within a category. Longer text can also reveal more of the signals used in expertise annotation. The persistence of the association within tasks should therefore be read as a descriptive robustness check, not as identification of an expertise effect.

5.3 Distributional and Task-Balanced Comparisons

Means can be influenced by long artifact-containing queries. Figure 4 therefore complements the mean comparisons with medians and quartiles. Median first-query lengths are 18 words for Novice, 35 for Intermediate, and 88 for Expert. The corresponding interquartile ranges are 13–40, 17–94, and 31–212 words. The ordering of group medians matches the ordering of group

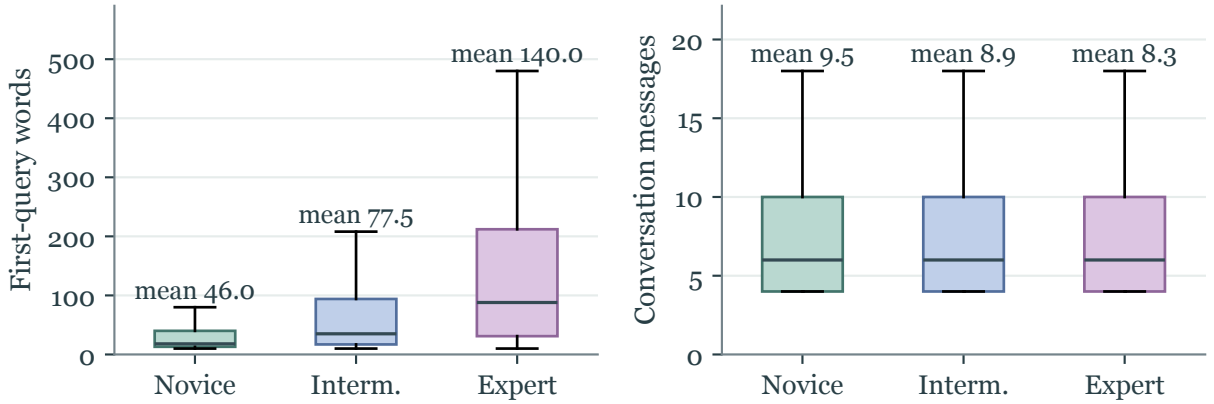


Figure 4: Query-length and conversation-length distributions by expertise label. Boxes mark quartiles and medians; whiskers cover observed values within 1.5 interquartile ranges. Outlying points are omitted only from display. Means use all conversations.

means, while the overlapping ranges show that length does not deterministically identify expertise.

To compare groups under a common task mixture, we average the six within-task means with equal weights. The resulting means are 37.9, 69.8, and 125.6 words for Novice, Intermediate, and Expert. The task-balanced Expert–Novice gap is 87.6 words, compared with 94.0 words in the overall corpus. Thus, the ordering persists when task proportions are held constant. Equal weighting defines a standardized comparison, not an estimate of natural task prevalence; it does not remove differences within task categories.

Appendix D provides rank-based comparisons and bootstrap intervals. Expert-labeled queries tend to be longer within every task, but sparse cells yield imprecise estimates. These distributional checks support the descriptive association without making query length an annotation rule.

5.4 Conversation Length and Change

Conversation lengths average 9.52 turns for Novice, 8.89 for Intermediate, and 8.30 for Expert labels. Here a turn count is the stored total number of conversation messages, not solely the number of user messages. The Novice–Expert comparison was not significant ($p = 0.386$, Mann–Whitney test). Opinion/Debate and Information-seeking have longer mean conversations (10.7 and 10.3 messages) than Planning/Design and Creation/Generation (8.0 and 8.2).

Expert query lengths average 141.2, 138.5, and 142.8 words across early, middle, and late conversation stages; the corresponding Novice means are 44.6, 51.3, and 46.9. These group averages suggest broadly persistent differences in query length. A middle-stage increase among Novice-labeled dialogues could reflect several processes, including changed tasks or uptake of model-provided vocabulary. Without a separate design that identifies those mechanisms, the pattern cannot establish learning, adaptation, or a change in underlying expertise.

6 Classification and Downstream Utility

We separate two questions: whether a classifier can recover the annotation labels, and whether conditioning on them helps calibrate responses. Classification agreement measures reproducibility of the operational taxonomy; downstream evaluation measures one proposed use of the labels. Neither alone demonstrates that the labels capture a stable psychological trait.

6.1 Automatic Classification

Two classification subtasks predict task type (seven classes, including Other) and expertise (three classes) from the first query. We use the 600-query held-out set and macro-F1 as the primary metric because class frequencies differ substantially. Baselines are majority-class prediction, TF–IDF with logistic regression, fine-tuned RoBERTa-base (Liu et al., 2019), and GPT-4o zero-shot prompting with the QUILT guidelines.

Table 4 summarizes classification performance. Lexical features improve over the majority baseline, particularly for expertise. RoBERTa and GPT-4o achieve higher macro-F1 for both dimensions, indicating that contextual and instruction-following models can recover more of the annotation distinctions. These results concern agreement with QUILT labels rather than independently measured user competence. Appendix B.2 discusses error categories; Table 11 includes the complete accuracy results.

Method	Task F1	Expertise F1
Majority	.095	.217
TF–IDF + LR	.388	.578
RoBERTa-base	.743	.718
GPT-4o zero-shot	.761	.732

Table 4: Automatic classification results on 600 test queries. F1 denotes macro-F1. Task prediction includes the fallback Other class.

Classification performance is lower for rare categories such as Planning, Opinion/Debate, and Other than for Creation and Debugging. RoBERTa and GPT-4o reduce the performance gap between high- and low-frequency categories, especially Opinion/Debate and Other. These patterns motivate macro-averaged evaluation and analysis of category-specific errors rather than reliance on aggregate accuracy alone.

6.2 Response-Adaptation Setup

The downstream study uses the same 600 test queries. GPT-4o produces a response under each of three conditions: a baseline without explicit QUILT labels, expertise-only conditioning, and joint task–expertise conditioning. The joint condition uses gold annotations, so it estimates utility under an oracle-label setting. It does not incorporate classification errors from the preceding experiment.

The expertise-only ablation holds the presence of the expertise label constant while removing the task label. This comparison is necessary because an improvement over the unconditioned baseline could otherwise arise entirely from knowing the user's apparent background. The joint comparison tests whether the immediate task contributes additional information about an appropriate response. A full factorial study would also include a task-only arm; that comparison remains outside this reported experiment.

The evaluation uses a blinded GPT-5.1 judge and scores responses from one to five on terminology appropriateness, knowledge calibration, depth appropriateness, and factual correctness. Terminology asks whether technical vocabulary is explained or presupposed appropriately. Knowledge calibration concerns the prerequisites assumed by the response. Depth concerns the amount of explanation or analysis relative to the query's objective. Factual correctness serves as a separate check rather than a proxy for calibration.

Blinding reduces evaluator bias, but LLM judges may favor particular styles. An independent Claude Sonnet 4.6 evaluation yields positive calibration deltas of 0.61, 0.58, and 0.29, with inter-judge preference agreement of $\kappa = 0.81$ (Appendix A.5). Human evaluation and task-success measures remain necessary complements.

6.3 Expertise-Only Ablation

Table 5 compares the three conditioning strategies. Expertise conditioning alone improves terminology from 3.64 to 4.10 and knowledge calibration from 3.52 to 3.95. Joint conditioning increases these means further to 4.31 and 4.18, corresponding to incremental gains of 0.21 and 0.23. Depth rises from 3.78 to 4.02 when task information is added, an incremental gain of 0.24. Figure 5 displays the three conditions on the same score range.

The depth improvement is consistent with task type informing answer structure: an expert requesting a repair may need a different response from an expert requesting a derivation. This is an interpretation of the mean-score pattern, not a separately identified mechanism.

Factual correctness changes only slightly, from 4.39 to 4.40 to 4.41. These close means do not establish equivalence or prove that correctness is preserved. Such a claim would require a prespecified equivalence margin and uncertainty estimates. We therefore describe the observed correctness means as similar, while distinguishing that descriptive result from a statistical safety or non-inferiority testing conclusion.

Paired significance testing across the 600 evaluation queries provides additional statistical support for the observed response-adaptation gains. The calibration gains are largest for Novice users, for whom explicit expertise signals help specify the appropriate explanatory level.

6.4 Implications for Response Adaptation

The incremental benefit of joint conditioning concerns the relationship between the requested objective and the level of exposition. Debugging calls for diagnosis and repair of supplied material, while creation calls for a new artifact. An expertise label alone does not identify which of these outputs the user needs. Conversely, a task label does not specify which prerequisites should be explained. This separation gives the two labels a clear role in a response policy without requiring a persistent user profile.

The gold-label setting separates the potential value of the annotation framework from the diffi-

Response dimension	Baseline	Expertise only	Joint	Joint − baseline	Joint − expertise
Terminology appropriateness	3.64	4.10	4.31	+0.67	+0.21
Knowledge calibration	3.52	3.95	4.18	+0.66	+0.23
Depth appropriateness	3.71	3.78	4.02	+0.31	+0.24
Factual correctness	4.39	4.40	4.41	+0.02	+0.01

Table 5: Mean judge scores on the 600-query response-adaptation evaluation. Joint conditioning supplies both task and expertise. The added expertise-only arm isolates the incremental contribution of the task label. Scores use a five-point rubric across four distinct response dimensions.

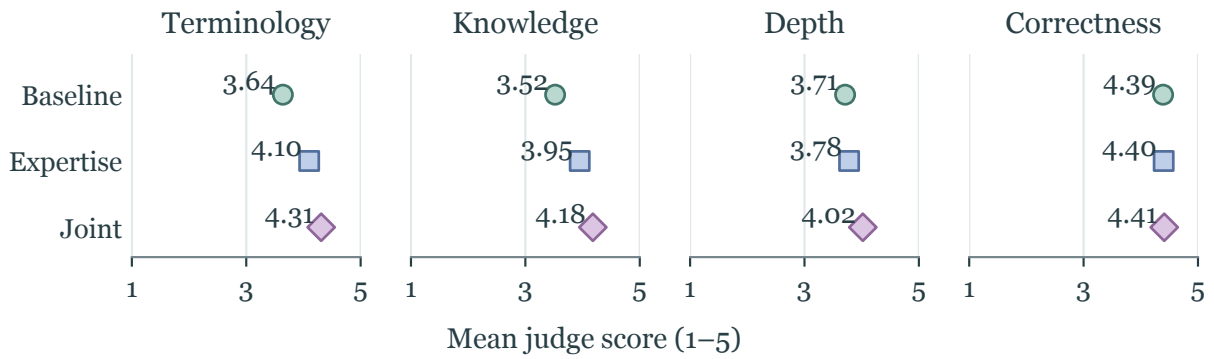


Figure 5: Response calibration across the three conditioning strategies. All panels use the same 1–5 score scale; direct numerical labels preserve readability in grayscale. Points show mean scores for terminology, knowledge calibration, depth, and correctness.

culty of predicting its labels. In practice, task misclassification may change the output objective, and expertise misclassification may change its exposition. A complete deployment evaluation should compare predicted-label conditioning with gold-label conditioning and allow users to correct the inferred response level. The present results motivate this evaluation, rather than establishing the performance of an end-to-end deployed system.

7 Research Uses and Conclusion

7.1 Broader Research Utility

QUILT supports several research directions beyond a single prompting experiment. For response adaptation, it supplies a compact representation of the current objective and expressed background knowledge. A system can expose or revise this representation with the user, compare explicit preferences with inferred labels, and evaluate the effects of misclassification.

For evaluation, the joint labels enable matched comparisons within task categories and expertise levels. Researchers can ask whether an assistant's errors, explanation quality, or calibration differ across cells rather than reporting one average over heterogeneous requests. Rare cells require careful uncertainty reporting; a balanced benchmark constructed from QUILT should not be confused with an estimate of natural usage frequencies.

For interaction analysis, first-turn labels support studies of subsequent dialogue. Such work should separate task changes, model-introduced vocabulary, and independent evidence of learning. Cross-platform and multilingual replications can test whether the rules retain their meaning under different linguistic conventions.

For annotation research, escalation records identify boundaries for targeted audits and active learning. Released classifiers could extend annotation to new samples. Their evaluation should include random high-confidence cases as well as escalated cases to assess whole-corpus accuracy.

7.2 Conclusion

We introduced QUILT, a framework for jointly annotating task intent and query-level expertise, and applied it to 3,000 WildChat dialogues. The corpus reveals systematic task–expertise associations and within-task differences in query length, while supporting automatic label prediction. The response-adaptation means favor joint conditioning over expertise-only conditioning on three calibration dimensions. Together, the taxonomy, corpus, and annotation materials provide a basis for expertise-aware interaction research. Their use should preserve the distinction between observable linguistic evidence, verified user knowledge, and causal claims about behavior.

Limitations

Expertise is inferred from language. QUILT labels describe expressed knowledge, not verified qualifications or stable ability. Knowledgeable users can ask elementary questions, and novices can reproduce specialized language. The conservative Expert rule may underrecognize expertise expressed concisely or unconventionally. Validation against independently collected domain knowledge and user-reported needs remains necessary.

Associations are not causal effects. Task type, expertise, and length derive from overlapping textual evidence. Within-task comparisons do not remove annotation dependence or confounding by domain, artifact length, or writing style. Longer queries may reveal more expertise cues even though length is not a criterion. The corpus and temporal statistics do not identify causal effects. Bootstrap intervals condition on the observed labels and do not account for labeling uncertainty or dependence between conversations from the same user.

Sampling and language coverage. The sample excludes non-English, flagged, single-turn, very short, and very long interactions. Its frequencies do not represent all LLM use. Language conventions and translation can change the signals used for expertise annotation. Initial Chinese, Japanese, and Spanish validation is encouraging; broader multilingual and cross-cultural testing remains necessary.

Annotation quality and taxonomy. Most labels are automatic; human agreement on selected difficult cases does not estimate whole-corpus accuracy. Self-reported confidence may be overconfident. Single-label annotation suppresses secondary intents, and boundaries can depend on the intended deliverable. Independent audits should include random high-confidence records as well as difficult cases. Rare cells, particularly Opinion/Debate with Expert labels, limit the precision of fine-grained comparisons.

Downstream evaluation. Gold-label conditioning and LLM judging do not measure the effects of classifier errors, user satisfaction, or real-world task success. Mean-score improvements describe the evaluated response dimensions, while similar correctness means do not establish equivalence or non-inferiority. The evaluation uses one response model and does not include a task-only conditioning arm. Human studies and end-to-end predicted-label evaluation are needed to assess broader practical benefit.

Ethics Statement

Data provenance and privacy. QUILT derives from the publicly released WildChat dataset (Zhao et al., 2024). Public availability and upstream anonymization do not guarantee that free-text conversations contain no identifying or sensitive information. Any redistribution should follow the source license and access conditions, review quoted examples, minimize retained metadata, and support removal of sensitive records. We do not attempt to identify the people behind conversations or use demographic metadata to assign expertise labels.

Responsible interpretation. Expertise labels must not be used as certified assessments of individuals or as grounds for discrimination, ranking, or restricting access to services. Linguistic style is affected by language proficiency, disability, context, and communication preference. A helpful adaptation mechanism should remain contestable: users should be able to request a different explanation level, and inferred expertise should never override safety requirements or explicit user instructions.

Annotator welfare. Human adjudicators received compensation, and toxic-flagged conversations were excluded before annotation. Filtering reduces exposure but cannot guarantee that all harmful content is removed. Future annotation efforts should provide clear content notices and allow annotators to skip distressing material. Quality review should account for annotator fatigue and disagreement.

Research use and transparency. The framework is intended for research on interaction analysis, evaluation, and user-adaptive responses. Appendix E documents the annotation messages, guideline, structured output, and review procedure. We plan to release the QUILT Corpus, guidelines, fine-tuned RoBERTa-base, and prompting-based classifiers with supporting code. Applications should document their adaptation policy, monitor performance across language and expertise groups, and provide a way for users to correct inappropriate assumptions about their knowledge.

References

Anthropic. 2025. Introducing the Anthropic Economic Index. February 10.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Yuyang Dai. 2026. Rescaling confidence: What scale design reveals about LLM metacognition. *Preprint*, arXiv:2603.09309.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

OpenAI. 2025. How people are using ChatGPT. September 15.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393. Association for Computational Linguistics.

Shramay Palta, Nirupama Chandrasekaran, Rachel Rudinger, and Scott Counts. 2025. Speaking the right language: The impact of expertise (mis)alignment in user-AI interactions. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 58–69. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, pages 13–19. ACM.

Si Thu and A. Baki Kocaballi. 2025. From prompting to partnering: Personalization features for human–LLM interactions. *Preprint*, arXiv:2503.00681.

Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do users really ask large language models? an initial log analysis of Google Bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2703–2707. ACM.

Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding user experience in large language model interactions. *Preprint*, arXiv:2401.08329.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1m ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

A Annotation Guidelines and Edge Cases

This appendix presents extended decision rules and reference examples for consistent annotation. The operational aim is to label the first query, not infer the writer's identity or qualifications. Annotators should document the observable evidence supporting each decision and retain ambiguity for subsequent adjudication.

A.1 Task-Type Boundaries

Information versus reasoning. The decisive question is whether the expected answer requires a derivation. "What causes malaria?" primarily requests established factual knowledge. "Why does X cause Y under constraint Z?" may require combining evidence or applying a causal model. The word "explain" alone is insufficient: explaining why entanglement cannot transmit information faster than light calls for a reasoning process. Annotators should use the intended answer, rather than the question word, to resolve this boundary.

Creation versus debugging. Repair requires a prior artifact, error message, or failed attempt in the input. Requests to fix, improve, refactor, or continue supplied material therefore belong to Debugging/Problem-solving when revision is the primary objective. "Fix my essay style" with a pasted essay supplies such an artifact; it is not creation from scratch. When no existing material is provided and a finished artifact is requested, Creation/Generation is appropriate.

Creation versus planning. The distinction concerns the primary output: a finished artifact versus an organization of actions or design decisions. A request for a ready-to-use document is creation, while a request to organize stages, dependencies, or architecture is planning. "Write a marketing plan" can be ambiguous. The intended use and requested components should guide adjudication, with the rationale preserving why one objective was treated as primary.

Reasoning versus opinion. Reasoning seeks a conclusion supported by evidence and explicit assumptions. Opinion seeks a recommendation or evaluative stance whose answer depends substantially on preference. A technical comparison can be reasoning even when phrased as "which is better?" Conversely, a recommendation involving personal priorities may remain opinion despite including technical terms. The framework does not assume that every recommendation lacks evidence; it classifies the primary goal of the request.

Multiple intents and Other. For a query that asks for an explanation and then a generated artifact, select the dominant objective and record the secondary intent in the rationale when useful. Do not assign several task labels under the current scheme. Use Other only when the substantive categories fail to describe the request, not merely because a boundary is difficult. Ambiguous cases should be escalated for review.

A.2 Expertise-Level Boundaries

Conservative Expert assignment. Require at least two independent signals among: multiple unexplained domain terms; explicit trade-offs, edge cases, constraints, or performance considerations; and professional or research framing such as system specifications or benchmarks. A single specialized word does not establish expertise. An Intermediate default is appropriate when evidence is insufficient or mixed, while the rationale should identify the unresolved evidence.

Short or terminology-free queries. Short queries are not automatically Novice. A specific decision about index funds and individual stocks can express targeted familiarity without dense terminology. Basic definitional requests or explicit beginner framing support Novice; a well-specified standard decision can support Intermediate. Query length, demographic attributes, country, and inferred identity must not substitute for textual evidence.

Role statements and domain shifts. A self-description such as "I am a nurse" supports an expertise judgment only when accompanied by relevant problem specification. Role statements are not independently verified. A query that explicitly expresses uncertainty should be interpreted from its content even if the author names a profession. Expertise evidence in programming must not be transferred to a medical or legal question without domain-specific textual support.

A.3 Reference Examples

Table 6 provides representative query excerpts and their label rationales. Quotations are shortened for readability. Expertise reference cases in Table 7 similarly explain the annotation target. Their role is instructional rather than an additional measurement of annotation accuracy.

Task	Query excerpt	Annotation rationale
Information	"Which pressure sore stages can be managed non-surgically with local wound care and pressure relief?"	Requests clinical facts rather than a derivation. The task label does not establish the user's expertise.
Reasoning	"Explain to a beginner the difference between expected value and variance and why a game with smaller EV but larger variance may be worse."	Requires relating probabilistic concepts through an argument; explicit beginner framing is considered separately.
Creation	"Please write a console Java application that uses a startup parameter to launch worker threads and wait for them to finish."	Requests a new code artifact without supplying an existing implementation for repair.
Debugging	"Whitelabel Error Page: There was an unexpected error (type=Internal Server Error, status=500). File UserController…"	The error trace is an existing problem instance; diagnosis and repair are the intended output.
Planning	"I work in apparel and we are making a leaflet showing our sustainability roadmap. What should I include?"	Asks for components of a roadmap rather than a completed leaflet; context determines the primary objective.
Opinion	"Act like a business analytics expert. Should I learn Tableau, Power BI, or KNIME to be most valuable in the current job market?"	Requests a recommendation whose answer depends on goals and context. The requested assistant role does not identify the user's expertise.

Table 6: Representative task-type query excerpts and their annotation rationales.

A.4 Human Annotation Workflow

The annotation interface contains the decision tree, category definitions, extended rules, and reference examples. The protocol uses a 30-item qualification test from the pilot adjudication set, with a target of $\kappa \geq 0.70$, and a limit of 150 items per session. Annotators reviewed flagged disagreements independently before discussion. Independent annotation precedes discussion so that agreement measures reflect decisions made without access to another annotator's judgment.

A.5 Judge Model Ablation

We repeat blinded response evaluation with Claude Sonnet 4.6 as an independent judge. Both judges agree on the direction of the score differences, and inter-judge agreement on pairwise preference labels is $\kappa = 0.81$. Table 8 compares the calibration and correctness deltas. The three calibration gains remain positive under Claude, while the correctness difference is small under both judges. This supports consistency across the two evaluators without establishing correctness equivalence.

A.6 End-to-End Evaluation Considerations

An operational evaluation should connect the classifier to the response generator and compare predicted-label conditioning with the gold-label upper bound. Errors in task and expertise need separate analysis: a wrong task label can change the requested output, whereas a wrong expertise label can change its exposition. Allowing user correction offers a further condition for testing practical utility.

Response depth should also be distinguished from response length. A technically detailed answer can be concise when it identifies a specific fault or states the relevant assumption. A beginner-oriented answer can require more words to explain unfamiliar concepts. Optimizing for brevity or verbosity alone would therefore miss the dimensions evaluated here. A useful policy should preserve the requested deliverable while varying terminology, prerequisites, and explanatory structure.

B Corpus Statistics and Classifier Details

Table 9 gives the complete joint counts, including Other. Table 10 summarizes the marginals and language statistics. The table includes all annotated conversations. Expertise percentages are 28.8%, 48.4%, and 22.7% after one-decimal rounding. The smallest substantive cell, Opinion/Debate × Expert, contains only four conversations; fine-grained comparisons require caution.

B.1 Classification and Human Agreement

Table 11 gives both macro-F1 and accuracy for task and expertise classification. Human agreement on the 54 adjudicated cases is $\kappa = 0.84$ for task type and $\kappa = 0.83$ for expertise, with accuracy values of 85.2% and 89.3%, respectively.

B.2 Classifier Error Analysis

Approximately 31% of Reasoning/Analysis instances are misclassified as Information-seeking

Level	Query excerpt	Interpretation of observable evidence
Novice	"I want to create a YouTube channel but don't know which one to create. I have knowledge in sports, animals, movies, celebrities."	Broad introductory guidance and explicit uncertainty; no detailed problem specification.
Novice	"Is this a good mashup title? … Mashup title: The Fusion Girls."	Requests general evaluative help, with little evidence of specialized criteria in the excerpt.
Intermediate	"I am having trouble understanding the word order in the French phrase 'nous ne songeons pas combien la femme réelle y tient peu de place.' Is it unusual?"	A targeted language question with concrete context, while still seeking explanation.
Intermediate	"Please write an IELTS essay: 'Governments are encouraging industries to move from cities to regional areas. Do the advantages outweigh the disadvantages?' "	Specifies a recognizable writing task and assessment setting, but does not supply multiple advanced expertise signals.
Expert	"Why is gamma-ray emission from a black hole jet not attracted back by gravity, while light cannot escape?"	Dense astrophysics framing presupposes relativistic jets, photon escape velocity, and distinctions between massive and massless particles near a black hole.
Expert	"Complete the financial model. What is the Revenue in 2025 under Base Case Drivers? 63.971M / 61.833M / 64.331M / 58.013M."	The annotation uses professional financial-model context. The short excerpt alone may not expose all evidence, so the full query is required to audit the label.

Table 7: Expertise reference excerpts. The examples illustrate language-based labels rather than independently verified user competence. Ambiguous excerpts require the full query and conservative application of the evidence rules.

Dimension	GPT-5.1 Δ	Claude Δ
Terminology	+0.67	+0.61
Knowledge calibration	+0.66	+0.58
Depth	+0.31	+0.29
Factual correctness	+0.02	+0.03

Table 8: Judge-model ablation. Differences are QUILT-conditioned minus baseline scores on the 1–5 rubric. Claude Sonnet 4.6 provides the independent evaluation.

by TF–IDF + logistic regression. The most important task boundary is Information-seeking versus Reasoning/Analysis. Both categories often use question forms such as *why* or *explain*, but differ in the inferential work expected in the answer. Lexical cues alone can therefore be insufficient. Planning/Design and Creation/Generation share artifact-oriented language and require attention to whether the primary deliverable is a roadmap or a finished product.

Approximately 28% of Novice instances are predicted as Intermediate. Expertise errors often concern the Novice–Intermediate boundary. A targeted request may contain few technical terms, while a query with specialized words may still seek elementary clarification. The annotation framework resolves these cases through the combination of context, assumed knowledge, and explicit evidence rules. These shared boundaries motivate evaluating label predictions with their rationales, rather than interpreting all errors as equivalent.

Macro-F1 gives each class equal weight and is therefore sensitive to rare categories. The Other task category has only 26 corpus examples, and the smallest task–expertise combinations contain fewer examples still. Small changes in predictions can consequently alter macro-averaged performance without materially changing behavior on the dominant creation category. Comparisons should preserve the same label inventory and data partition, and broader transfer should be assessed on independent conversational samples.

C Language Filtering and Generalization

C.1 Cross-Lingual Validation

We apply the same annotation pipeline to Chinese ($n = 120$), Japanese ($n = 95$), and Spanish ($n = 110$) conversations. A balanced subset of 50 conversations across these languages is translated into English and independently labeled by a human annotator using the QUILT guidelines. Agreement with the pipeline labels is $\kappa = 0.78$ for task type and $\kappa = 0.76$ for expertise. Table 12 summarizes the cross-language distributions. Creation is the largest task category and Intermediate the largest expertise category in all four language samples. These initial observations motivate broader multilingual validation rather than establishing equivalence across languages.

Task type	Novice	Intermediate	Expert	Total
Information-seeking	259 (8.6%)	265 (8.8%)	81 (2.7%)	605
Reasoning/Analysis	34 (1.1%)	203 (6.8%)	112 (3.7%)	349
Creation/Generation	490 (16.3%)	737 (24.6%)	266 (8.9%)	1,493
Debugging/Problem-solving	24 (0.8%)	153 (5.1%)	168 (5.6%)	345
Planning/Design	11 (0.4%)	71 (2.4%)	48 (1.6%)	130
Opinion/Debate	27 (0.9%)	21 (0.7%)	4 (0.1%)	52
Other	20 (0.7%)	3 (0.1%)	3 (0.1%)	26
Total	865 (28.8%)	1,453 (48.4%)	682 (22.7%)	3,000

Table 9: Complete joint distribution. Parenthesized percentages use all 3,000 conversations as the denominator, not the row total.

Category	Count	Corpus %	Mean words	Mean messages
Information-seeking	605	20.2	39.5	10.3
Reasoning/Analysis	349	11.6	74.0	9.9
Creation/Generation	1,493	49.8	91.3	8.2
Debugging/Problem-solving	345	11.5	130.7	8.5
Planning/Design	130	4.3	94.5	8.0
Opinion/Debate	52	1.7	42.9	10.7
Other	26	0.9	–	–
Novice	865	28.8	46.0	9.5
Intermediate	1,453	48.4	77.5	8.9
Expert	682	22.7	140.0	8.3

Table 10: Marginal counts and language statistics for the annotated corpus. Words refer to the first user query; messages refer to the full conversation. Other is excluded from the six-task linguistic comparisons.

C.2 Scope and Transfer

The corpus is restricted to conversations marked English in WildChat. This choice narrows the annotation setting but also conditions the sample on a particular language label. The resulting frequencies should not be extrapolated to multilingual assistant usage. A source-language field can also miss code-switching, translated material, or mixed-language technical terminology within an otherwise English conversation.

Expertise is inferred from linguistic evidence that can vary across languages. Technical terms may be borrowed, abbreviated, or expressed through different conventions. Explicitly defining a term may be a courtesy to the assistant rather than an indication of unfamiliarity. The absence of an English technical expression cannot be treated as absence of the corresponding domain knowledge. Multilingual annotation should therefore validate equivalent decision criteria rather than translate a list of keywords mechanically.

Translation can change the evidence being measured. It may expand abbreviations, standardize terminology, or make implicit relations explicit. Agreement between an English translation and its source query can consequently reflect both translation quality and annotation consistency. A suitable multilingual evaluation would examine native-language judgments alongside translated versions and analyze where the inferred task or expertise changes.

Language and task composition also need to be separated. A language sample containing more debugging requests could contain more visible code and constraints even if there were no systematic difference in domain knowledge. Matching or stratifying by task and topical domain helps distinguish compositional differences from changes in how the annotation rules apply. Within-task statistics offer an interpretable starting point, but sufficient sample sizes in each expertise cell remain necessary.

The other eligibility rules have related effects. Requiring multiple user turns excludes self-contained interactions that end after a satisfactory first response. The 10-word lower bound excludes very short requests, including concise technical questions. The 500-word upper bound excludes long artifacts and document pastes. Toxicity and redaction filters further restrict the represented interaction population. These decisions should be revisited when adapting QUILT to a platform with different conversation conventions.

Method	Task macro-F1	Task accuracy	Expertise macro-F1	Expertise accuracy
Majority class	.095	.497	.217	.483
TF–IDF + logistic regression	.388	.687	.578	.578
RoBERTa-base	.743	.821	.718	.741
GPT-4o zero-shot	.761	.834	.732	.756

Table 11: Complete automatic-classification results on the 600-query QUILT test set. Macro-F1 is the primary metric.

Language (n)	Info.	Reason.	Creation	Debug.	Planning	Opinion
English (3,000)	20.2	11.6	49.8	11.5	4.3	1.7
Chinese (120)	18.9	13.2	47.5	12.8	5.1	2.1
Japanese (95)	22.6	10.8	46.9	10.5	6.2	2.6
Spanish (110)	19.4	9.9	51.2	12.3	4.9	1.9

Language (n)	Novice	Intermediate	Expert
English (3,000)	28.8	48.4	22.7
Chinese (120)	26.5	47.9	25.6
Japanese (95)	31.2	45.8	22.9
Spanish (110)	29.4	46.3	23.9

Table 12: Cross-language task-type and expertise distributions (percentages). The task panel lists the six substantive categories. The translation-based human validation uses 50 conversations in total across the three non-English languages.

Task	n_N/n_E	A_t	Gap 95% CI
Information	259/81	.726	[31.0, 85.1]
Reasoning	34/112	.824	[64.8, 112.0]
Creation	490/266	.774	[73.4, 107.7]
Debugging	24/168	.758	[70.0, 147.0]
Planning	11/48	.845	[64.1, 151.8]
Opinion	27/4	.889	[2.8, 143.4]

Table 13: Within-task length comparisons. A_t is empirical probability of superiority; the last column gives a percentile bootstrap interval for the Expert-minus-Novice mean word-count gap. Intervals condition on the observed labels.

D Within-Task Length Robustness

We further compare Expert and Novice queries within each task using the empirical probability of superiority,

$$A_t = \Pr(L_{E,t} > L_{N,t}) + \tfrac{1}{2}\Pr(L_{E,t} = L_{N,t}),$$

where $L_{E,t}$ and $L_{N,t}$ denote word counts drawn from the corresponding annotated groups. We estimate A_t from the Mann–Whitney statistic divided by the number of cross-group pairs. This rank-based measure is less sensitive than a mean difference to the magnitude of a few extremely long queries. It describes the ordering of observed lengths, not a probability that one person is more competent than another.

Table 13 shows $A_t > 0.5$ in all six categories. For example, the value of .726 in information seeking means that a randomly selected Expert-labeled query exceeds a randomly selected Novice-labeled query in length in approximately 72.6% of cross-group comparisons, with ties receiving half weight. This complements the mean gap of 56.3 words without requiring a normal length distribution. A large rank separation still does not imply that any particular short query must receive a Novice label.

For each task, we also estimate uncertainty in the mean difference by resampling conversations independently within the Expert and Novice cells, preserving their observed sizes. We use 10,000 bootstrap resamples, random seed 1268, and the 2.5th and 97.5th percentiles of the resampled differences. The resulting intervals describe sampling variability conditional on the labels and the selected corpus. They do not incorporate label uncertainty, user-level dependence, or uncertainty introduced by the source and filtering process.

Opinion/Debate illustrates the importance of cell size: only four Expert-labeled queries contribute to that comparison, producing a wide interval despite a high probability-of-superiority estimate. Planning also has few Novice cases. These comparisons should not be ranked by point estimate alone. The larger information-seeking and creation groups provide more observations, but their estimates remain specific to the sampled interaction population.

E Annotation Prompts and Implementation

E.1 Models and Request Configuration

The annotation pipeline uses `google/gemini-2.5-flash-lite` for first-pass annotation and `google/gemini-2.5-flash` for escalated review through the OpenRouter chat-completion interface. Requests use temperature zero and a strict JSON-schema response format. Both stages receive the same guideline and the first user query; review changes the model, not the annotation target.

The annotation client batches input items, each containing an identifier and query text. Its command-line interface exposes batch size, model identifier, temperature, and retry count. The default batch size is 20 and the retry limit is four. Requests use a 120-second timeout with exponential retry backoff for transient failures. Unicode query text is preserved during JSON serialization.

E.2 Annotation Message Construction

The system message consists of the following prefix, two newline characters, and the complete guideline in the next subsection. The user message consists of its prefix, one newline, and the JSON batch. The batch identifiers link returned annotations to source queries; no country, identity, or demographic fields are supplied to the annotator.

```
SYSTEM PREFIX:
You are a careful research annotation
↪  assistant. Classify each query using
↪  only observable text evidence.
↪  Return valid JSON only.

USER PREFIX:
Annotate these QUILT queries:

BATCH FORMAT:
[{"id": "<id>", "first_user_query":
↪  "<query>"}, ...]
```

The response schema contains an `annotations` array. Every object requires `id`, `task_type`, `expertise_level`, `confidence`, and `rationale`. Additional properties are disallowed. Task values are restricted to the six substantive categories and Other; expertise values are Novice, Intermediate, and Expert. Confidence is a single number between zero and one for the annotation, not a separately calibrated probability for each dimension.

E.3 LLM Annotation Guideline

The following guideline is supplied to both annotation models. Appendix A gives the extended decision rules used to explain and resolve difficult cases during human review.

```
# QUILT Annotation Guideline

## Task Type

Assign exactly one label.

- `information_seeking`: The user asks
↪  for factual knowledge, definitions,
↪  or explanations, and mainly expects
↪  informative content.
- `reasoning_analysis`: The user asks
↪  for inference, comparison,
↪  evaluation, causal explanation,
↪  diagnosis, or multi-step analysis.
- `creation_generation`: The user asks
↪  the model to create a new artifact
↪  such as text, code, plans, images,
↪  names, or examples from scratch.
- `debugging_problem_solving`: The user
↪  provides existing content, an error,
↪  a failed attempt, or a concrete
↪  problem and asks for a fix or
↪  solution.
- `planning_design`: The user asks for a
↪  plan, architecture, workflow,
↪  research design, multi-step
↪  strategy, or system design.
- `opinion_debate`: The user asks for a
↪  recommendation, preference,
↪  judgment, or stance where multiple
↪  reasonable answers may exist.
- `other`: Use only when none of the
↪  above labels fits.

Boundary rules:

- Existing content plus a request to
↪  fix, revise, debug, or solve points
↪  to `debugging_problem_solving`.
- A concrete artifact requested from
↪  scratch points to
↪  `creation_generation`.
- A roadmap, architecture, or multi-step
↪  procedure points to
↪  `planning_design`.
- A preference or recommendation points
↪  to `opinion_debate`.
- A "why/how/compare" question with
↪  inference or trade-offs points to
↪  `reasoning_analysis`.
- A definition or direct explanation
↪  with no substantial inference points
↪  to `information_seeking`.

## Expertise Level

Assign exactly one label based only on
↪  observable language in the query.

- `novice`: The query asks for basic
↪  explanation, uses few or no domain
↪  terms, explicitly says the user is
↪  new, or misuses core terminology.
```

```
- `intermediate`: The query uses domain
↪  terms correctly and has concrete
↪  context, but still asks for
↪  explanation, validation, or standard
↪  practice.
- `expert`: The query has dense domain
↪  terminology, omits basic
↪  definitions, and involves
↪  trade-offs, edge cases, constraints,
↪  methods, standards, or recent
↪  developments.

Important:

- Expertise is query-level, not
↪  user-level.
- Expertise is domain-relative. The same
↪  person may be expert in programming
↪  and novice in medicine.
- Do not infer expertise from
↪  demographics, metadata, country, or
↪  identity.

## Output Requirement

Return only valid JSON matching the
↪  requested schema.
```

E.4 Review Selection

The review selector checks annotation confidence, the Other label, and ambiguity markers in the rationale. Its configurable confidence threshold defaults to 0.75. The ambiguity markers are `ambiguous`, `unclear`, `could be`, `borderline`, `possibly`, and `hard to tell`. These criteria target uncertain or difficult decisions rather than selecting a random validation sample.

After escalated annotation, the pipeline compares task and expertise predictions separately. A disagreement in either dimension flags the conversation for human review. Retaining the model rationales supports adjudication of recurring boundaries, including reasoning versus information seeking and intermediate versus expert framing. The distinction between targeted review and random quality assessment is essential when interpreting agreement statistics.

E.5 Statistical Definitions

Word count is the number of whitespace-delimited tokens in the first user query. Conversation length counts all messages, including assistant messages. The joint chi-square analysis excludes Other and uses the six-by-three contingency table. Cramer's V is computed as $\sqrt{\chi^2/(n \min(r-1, c-1))}$ with $r = 6$, $c = 3$, and $n = 2{,}974$.

For the length comparisons, the unit of analysis is a conversation. Bootstrap intervals resample within task–expertise cells, preserve observed cell sizes, and use 10,000 resamples with random seed 1268. Percentile intervals describe uncertainty in mean length differences conditional on the observed labels. They are separate from the downstream mean judge scores. Task-balanced means weight each of the six substantive task categories by one sixth and exclude Other.