Language Models Don't Know What You Want: Evaluating Personalization in Deep Research Needs Real Users

Nishant Balepur[1,2,3*] **Malachi Hamada**[2] **Varsha Kishore**[2] **Sergey Feldman**[2]
Amanpreet Singh[2] **Pao Siangliulue**[2] **Joseph Chee Chang**[2]
Eunsol Choi[3] **Jordan Boyd-Graber**[1] **Aakanksha Naik**[2]

[1]University of Maryland [2]Allen Institute for Artificial Intelligence [3]New York University
nbalepur@umd.edu aakankshan@allenai.org

MyScholarQA: https://personalized-scholarqa.apps.allenai.org/

Abstract

Deep Research (DR) systems help researchers cope with ballooning publishing counts. Such tools synthesize scientific papers to answer research queries, but lack understanding of their users. We address this with MySCHOLARQA (MySQA), a personalized DR agent that: 1) infers a profile with a user's research interests; 2) proposes personalized actions for a user's input query; and 3) writes a multi-section report for the query that follows user-approved actions. We first test MySQA with NLP's standard protocol: we build a benchmark with synthetic users and LLM judges, where MySQA beats baselines in citation metrics and personalized action-following. However, we suspect this process does not cover all aspects of personalized DR users value, so we interview users in an online version of MySQA to unmask them. We reveal nine nuanced errors of personalized DR undetectable by our LLM judges, and we study qualitative feedback to form lessons for future DR design. In all, we argue for a pillar of personalization that easy-to-use LLM judges can lead NLP to overlook: real progress in personalization is only possible with real users.[1]

1 When Deep Research Gets to Know You

Scholars increasingly turn to LLMs to support their scientific research (Liao et al., 2024), such as to learn new concepts (August et al., 2023) or brainstorm ideas (Pu et al., 2025). With publishing rates skyrocketing and literature becoming daunting to track (Parolo et al., 2015), a new use case of LLMs emerges: **Deep Research (DR)** tools that answer researchers' queries by retrieving, organizing, and synthesizing papers into multi-section, attributed reports (Asai et al., 2024; Huang et al., 2025).

DR has advanced in giving well-cited reports for queries (Bragg et al., 2026), but few capture the individual needs of *who* asks them, lacking **personalization**. By knowing a researcher's background, DR could create more helpful reports: focusing on papers in certain domains, framing explanations in familiar terms, or showing how to use new ideas in users' ongoing work. While decades of search engine research motivates personalization's benefits in AI search tools (Teevan et al., 2005; Dou et al., 2007), work on personalized DR remains sparse.

We introduce MySCHOLARQA (MySQA): the first open-source personalized DR tool that: 1) infers *user profiles* via papers users pick to capture their interests (Fig 1, left); 2) suggests tool *actions* tailored to the user's profile and query (Fig 1, center); and 3) writes a *report* to answer the query and execute actions via a multi-LLM system (Figure 1, right). For transparency and control, our users can edit profiles in (1) so MySQA best captures them, adjust the actions that it executes in (2), and review highlights where it personalizes content in (3).

A formative study showed MySQA's promise but that the system was imperfect (§2.4), leading us to ask: how should we evaluate MySQA to reveal common failures and track improvements? In surveying recent NLP work on personalization, offline benchmarks dominate (Voorhees et al., 1999): for 31 ACL'25 works on personalization, all evaluate offline (18 with synthetic user datasets and 17 with LLM judges), but **only two run user studies with real end users** (Appendix B). However, excelling in offline evaluation does not ensure the system is helpful (Mozannar et al., 2025), as offline metrics can neglect what users value (Venkit et al., 2025b).

To test what offline personalization evaluations miss, we first build a synthetic dataset pairing DR queries with paper sets simulating users (§3.1), and 16 offline metrics (§3.2). Here, MySQA excels:

[1]We release code and data at: https://github.com/allenai/personalized-scholarqa-eval

[*]Work primarily completed during internship at Ai2.

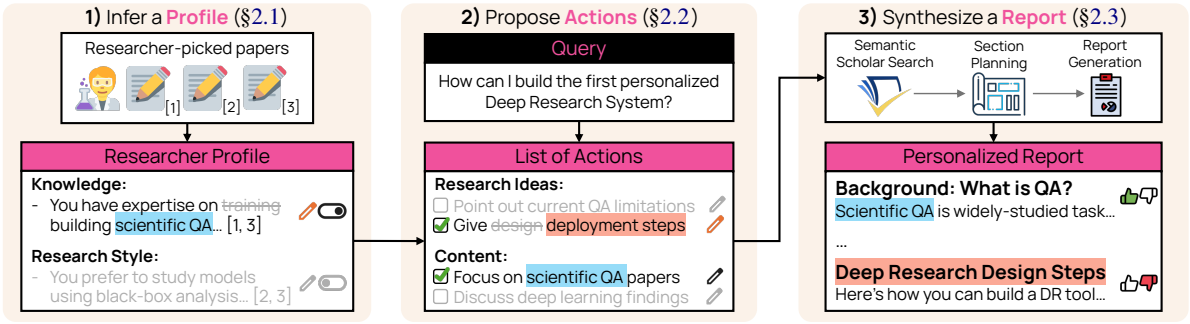

Figure 1: Overview of MySQA: our three-step personalized Deep Research system. (1) Researchers upload papers from Semantic Scholar, from which an LLM infers a profile that captures their interests. (2) When the researcher asks a query, MySQA proposes a list of actions that could alter the report, tailored to the researcher's profile. (3) The system generates a report that answers the query and executes these actions through a multi-stage retrieval and LLM generation pipeline. To improve transparency and control, users can edit profiles, adjust actions, and view highlights in the report where MySQA personalizes content.

profiles accurately cite papers for inferred user interests, personalized actions more closely match user profiles versus generic ones, and our reports have higher quality and adherence to actions versus open-source and commercial DR systems (§3.4). To contrast this with online evaluations, we draw on human-computer interaction and use MySQA as a probe (Hutchinson et al., 2003) to unveil real users' needs in personalized DR. In 90-minute interviews, 21 DR users build profiles, pick query actions, and rate reports via a MySQA-backed interface (§4.1).

Participants perceive MySQA as helpful—73% of its profiles, actions, and reports are satisfactory—but uncover nine personalization flaws our offline metrics miss: e.g., profiles overstating user expertise and tailored actions drifting from query intent (§4.2). To test if we could have flagged these flaws offline, we use LLM judges to predict user satisfaction ratings (§4.3) with validation sets derived from interviews: they never beat majority class baselines, failing to capture user needs. Beyond finding metrics to advance, user feedback also yields lessons to inform personalized DR design: make control easy, let users digest personalization, use content beyond papers, and evaluate via mixed study designs (§5).

We aim to reinforce a pillar of personalization research: developing personalized tools needs feedback from real users. Synthetic data is an alluring crutch and LLM judges seem reliable (Jiang et al., 2025a), but these evaluations can miss what users actually value. Thus, we urge their adoption as necessary but insufficient checks and instead advocate for user-centered evaluations to inform the design of personalized NLP systems.

Our contributions are:

1. MyScholarQA, the first personalized Deep Research (DR) system with an online demo.[2]
2. A synthetic benchmark and LLM judge metrics as offline evaluations of personalized DR.
3. Discovery that LLM judges fail to predict nine failure modes of personalized DR, advocating for user-centered personalization evaluations.
4. The first formative and usability studies of personalized DR with 26 active DR users to inform future work on personalization design choices.

2 MySQA: Personalized Deep Research

Drawing on Brusilovsky (1996)'s adaptive hypermedia, personalized Deep Research tools build a persistent user model of the researcher, then apply this model to adapt reports for user queries. Our design goals are: 1) build a user model to capture research interests; and 2) let users interpret and control how the system personalizes (Liu et al., 2024a), helping us learn the best ways to adapt outputs.

We realize this in MyScholarQA (MySQA), a DR tool (Fig 1) that infers profiles of users' interests via papers (§2.1), plans actions for user queries tailored to their profile (§2.2), and executes actions to adapt reports (§2.3). We now describe each step and evaluate its design in a formative study (§2.4).

2.1 Inferring Researcher Profiles

MySQA first infers a profile $\mathcal{P}$ from a user's papers $\mathcal{D}$ (Fig 2)—forming a persistent user model (Brusilovsky, 1996); we use papers, since Lin et al.

[2] https://personalized-scholarqa.apps.allenai.org/

(2024) show they capture research interests. Profiles can be biographies (Gao et al., 2024) or keywords (Mysore et al., 2023), but we use sentence-level inferences $\mathcal{P} = \{\mathcal{I}_1, \ldots, \mathcal{I}_{n_1}\}$ about the user (e.g. "*Your papers argue evaluations should move beyond metrics to online studies.*"), similar to prior research in computer agents and writing assistance (Shaikh et al., 2025; Garbacea and Tan, 2025).

$\mathcal{P}$ has n_1 inferences evenly split over five aspects inspired by Tang et al. (2024): knowledge (what they know), research style (how they do research), writing style (how they write), audience (whom they impact), and positions (what they believe). Users find this structure organized (§2.4). $\mathcal{P}$ transparently cites snippets from the user's papers $\mathcal{D}$ for each inference $\mathcal{I}$ with an explanation. We prompt LLMs to create $\mathcal{P}$ from $\mathcal{D}$ (Appendix A.11). Users can edit/disable any $\mathcal{I}$ to form a better profile $\mathcal{P}^*$.

2.2 Proposing Actions to Take

Equipped with a profile $\mathcal{P}^*$, users can ask a query q to get a report $\mathcal{R}$ adapted to $\mathcal{P}^*$ (Brusilovsky, 1996). Directly producing $\mathcal{R}$ limits customization; instead, we draw on query clarification (Zhang et al., 2025) and first return a list of actions $\mathcal{A} = \{a_1, \ldots, a_{n_2}\}$ (Fig 3) that MySQA could take when answering q, such as "*find trivia QA papers*" or "*add section on metrics*" (Srikanth et al., 2026). Exposing $\mathcal{A}$ lets users steer execution and tell us how to adapt $\mathcal{R}$—validated in a formative study (§2.4).

$\mathcal{A}$ has n_2 actions split over four categories, based on how they will adapt $\mathcal{R}$: content (what $\mathcal{R}$ covers), style (how $\mathcal{R}$ explains), specificity (how $\mathcal{R}$ interprets q), and research ideas ($\mathcal{R}$'s proposed ideas the user can incorporate). Actions can be *personalized* $\mathcal{A}_{\text{person}}$ (conditioned on q and $\mathcal{P}^*$) or *generic* $\mathcal{A}_{\text{gen}}$ (conditioned on q); $\mathcal{A}_{\text{gen}}$ ensures a is useful when $\mathcal{P}^*$ is unlike q (e.g. if users ask about a new field).

We prompt LLMs to create $\mathcal{A}_{\text{gen}}$ and $\mathcal{A}_{\text{person}}$ separately, merging them to form $\mathcal{A}$ (Appendix A.11); we add ways the LLM can adapt to $\mathcal{P}^*$ for $\mathcal{A}_{\text{person}}$'s prompt (e.g., skip basic terms for experts) but no strict rules, as we want to learn how to personalize reports from users (§4). Users can edit/disable any action $a \in \mathcal{A}$ to form custom actions $\mathcal{A}^* \subseteq \mathcal{A}$.

2.3 Synthesizing a Personalized Report

After the user submits actions $\mathcal{A}^*$, MySQA writes a multi-section report $\mathcal{R} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{n_3}\}$ to answer query q while executing each $a \in \mathcal{A}^*$, personalizing the report via actions which were shaped by the profile. We build on ScholarQA (Singh et al., 2025, SQA)—a DR system with high-quality reports. SQA chains LLMs to retrieve papers from Semantic Scholar,[3] cluster them into sections, and iteratively generate well-cited sections to form R.

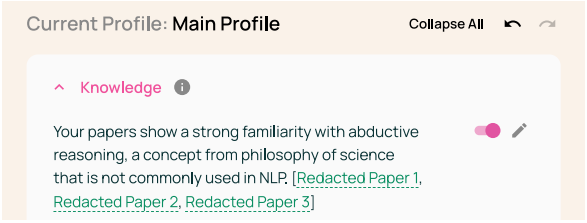

Figure 2: After a user selects research papers, MySQA infers an editable profile with inferences about the user, capturing their interests for personalizing reports.

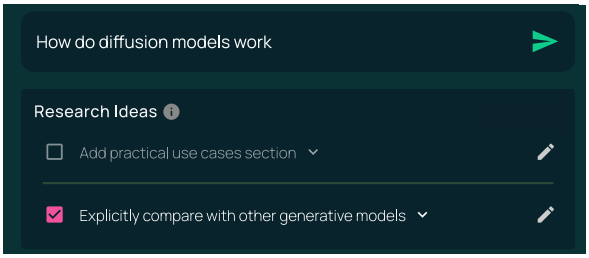

Figure 3: Before answering a user's query, MySQA proposes actions that change how the report could be created, which users can adjust to customize their report.

To equip MySQA to use $\mathcal{A}^*$, we tweak prompts in SQA's execution to also use $\mathcal{A}^*$ as input. For example, the first step is a prompt converting q to search terms for Semantic Scholar (i.e. "convert q to search terms"); we minimally modify it with instructions on how to also use $\mathcal{A}$ to create search terms (i.e. "convert q to search terms while following these actions: $\mathcal{A}^*$"). Most notably, we change: 1) the prompt for search terms—generating multiple search terms based on q and $\mathcal{A}^*$, while SQA originally generates just one; and 2) the report generation prompt—instructing the LLM to execute all $a \in \mathcal{A}^*$ and highlight parts of the text that relate to any action—one color per action (Fig 4). (1) tailors retrieval to $\mathcal{A}^*$, while (2) helps users see where MySQA personalizes $\mathcal{R}$ (Kim et al., 2025b, §5.2). Claude-4 Sonnet backs MySQA,[4] matching SQA.

2.4 A Formative Study with MySQA

To ensure our method matches users' expectations before investing in large-scale studies, we run a formative study (Nielsen, 1993) via an early version of MySQA as a technology probe (Hutchinson et al.,

[3] https://www.semanticscholar.org/product/api
[4] https://www.anthropic.com/news/claude-4

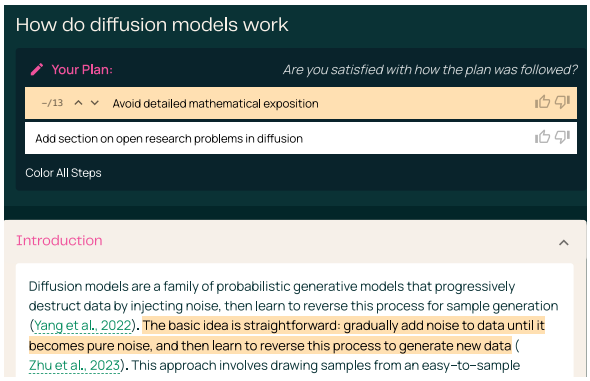

Figure 4: Reports in MySQA highlight personalized content, helping users navigate each action they select.

2003).[5] Five DR users/CS students bring papers of their interest and two queries they asked DR before. We first discuss their DR usage/personalization needs (~15 min). Then they use MySQA—creating and editing profiles/actions to get personalized reports ($35/hr; ~45 min); later over email, they rate reports for their queries (detailed in A.7).

Participants found using papers to infer personalized profiles/actions intuitive for personalized DR; out of all potential user context, P1 noted "*papers matter most*". They also found our profile/action categories useful and comprehensive (§2.1, §2.2). Many desired transparency, wanting it "*as transparent as possible*" (P4) and to "*know how they made inferences about me*" (P3).

Participants then rated their profiles, actions, and reports from MySQA. Profiles had surprising nuance and detail—"*It captures exactly what I have in mind but haven't expressed*" (P2)—but were incomplete: "*a really good starting point to review and refine*" (P5). Personalized actions were promising; some felt like "*speaking with a colleague who knew my work*" (P2). Still, participants liked generic/personalized ones similarly (~60%), so we do not yet know when personalized DR helps. Highlighting where reports tailored to actions made it "*much easier to see custom parts at a glance*" (P3) but such text could be "*a bit general*" (P5). Lastly, they compared generic SQA reports to personalized MySQA reports for their queries; all favored the latter, showing personalized DR's utility. In all, our study confirms MySQA is a solid basis for finding more aspects of personalized DR users value.

Post-study, we use participant feedback to refine MySQA. We tweak prompts to try and fix common errors (e.g. "make profiles specific", §2.1) and host a UI for further online use, with a: 1) profile page to find papers and toggle/edit LLM inferences (Fig 2); 2) home page to ask queries and toggle/edit LLM actions (Fig 3); and 3) report page with highlights showing where each action was executed (Fig 4). After adding these updates, we continue to larger-scale offline (§3) and online (§4) evaluations.

[5]Here, we use Gemini-2.5 Pro for MySQA profiles/plans.

3 Offline Evaluation

Our formative study showed MySQA's promise (§2.4) but did not thoroughly evaluate DR output quality. As a next step, we adopt NLP practices and test MySQA offline via simulated users (§3.1) and LLM judges (§3.2). These analyses are cheap and popular (Jiang et al., 2025a), but may not capture users' personalization needs (Venkit et al., 2025b), so we use them as necessary but insufficient checks to inform online, user-centered evaluations (§3.4).

3.1 Dataset Collection

No personalized DR datasets exist of user queries q and papers $\mathcal{D}$, so we make a synthetic one. We collect q from ScholarQA-CS2 (Bragg et al., 2026)—a research agent benchmark with 200 DR q (100 dev/100 test). We attach synthetic users to each q with low, medium, and high expertise for q. We simulate synthetic users based on papers in CS-PaperSum (Liu et al., 2025b): CS conference papers grouped by first author (with three or more papers). We compute expertise via cosine similarity[6] between GRIT-LM embeddings (Muennighoff et al., 2024) of q and user papers $\mathcal{D}$. We get 281 and 291 $(q, \mathcal{D})$ pairs for dev and test splits, respectively.

3.2 Metric Implementation

On our dataset (§3.1), we now study MySQA profiles, actions, and reports with objective metrics for clearly undesired errors, deferring a study of subjective metrics (§4.3). Gemini-2.5 Flash (Comanici et al., 2025) is our LLM judge (Zheng et al., 2023) for metrics (human agreement in Appendix A.2).

For each **profile** inference $\mathcal{I}$, we evaluate: 1) category accuracy, if $\mathcal{I}$ is in the correct category (e.g. knowledge); 2) inference accuracy, if $\mathcal{I}$ contradicts cited papers, like summary faithfulness (Kryscinski et al., 2019); 3) citation relevance, the proportion of cited papers that support any part of $\mathcal{I}$ to penalize overciting; and 4) specificity, a 1–5 score for how

[6]$[0, 0.2]$ (low), $(0.2, 0.35]$ (medium), and $(0.35, 1.0]$ (high); dropping ranges if no $\mathcal{D}$ matches

Model	Inf. Acc	Cit. Rel.	Cat. Acc.	Spec.	# Cite	# Words
G-Pro	**97.1**	**97.4**	99.4	3.73	**2.45**	23.2
Sonnet	92.5	97.4	99.1	4.12	1.91	21.6
OAI-o3	88.6	91.8	**99.8**	**4.20**	2.03	18.9
DS-r1	77.8	80.7	97.2	3.56	1.89	9.9

Table 1: Profile ($n_1 = 25$) inference accuracy, relevance, and specificity across LLMs. Highest scores are **bold**. Reasoning LLMs infer accurate, relevant user profiles.

Model	W.R.	p_{gen} Coh.	p_{person} Coh.	p_{gen} Uniq.	p_{person} Uniq.
G-Flash	91.3	93.2	84.0	42.9	60.7
Sonnet	93.5	91.8	82.0	**50.4**	**68.2**
GPT-4.1	**94.7**	**93.5**	**84.1**	49.1	66.3
DS-V3	94.3	89.2	72.3	54.5	72.2

Table 2: Personalized/generic action ($n_2 = 16$) win rate, coherence, and uniqueness. The highest score is **bold**. LLM judges that condition on synthetic users' papers are more likely to select personalized actions, indicating that they can personalize actions based on user profiles.

Model	An. Cov.	An. Prec.	Cit. Prec.	Cit. Rec.	$\mathcal{A}$ Adh.
MySQA	**91.4**	89.9	**91.8**	**81.4**	83.2
SQA	88.9	89.1	90.5	76.9	81.3
OpenSc.	77.2	**97.4**	82.5	60.4	82.5
STORM	72.0	92.2	73.3	64.7	74.4
Sonar DR	81.0	82.9	64.3	46.3	75.0
o3 DR	89.1	90.2	79.2	56.7	**93.8**

Table 3: DR report quality and action adherence scores for query q and eight actions in $\mathcal{A}_{\text{gen}} \cup \mathcal{A}_{\text{person}}$. The best score is **bold**. MySQA surpasses every DR baseline in three out of five metrics. We only evaluate o3 DR using ten examples due to latency and cost limitations.

much $\mathcal{I}$ differs among researchers. As confounders, we show mean cited paper count and words in $\mathcal{I}$.

We compare personalized and generic **actions** with: 1) win rate, how often a judge picks $\mathcal{A}_{\text{person}}$ vs $\mathcal{A}_{\text{gen}}$ given $\mathcal{P}$, a consistency check from Balepur et al. (2025a); 2) coherence, how often each $a \in \mathcal{A}$ does not contradict q (e.g. q ="what is QA", a ="focus on NLI" is a conflict); and 3) uniqueness, the proportion of $a \in \mathcal{A}$ that a system without $\mathcal{A}$ would not already follow. We check (3) via how often SQA prompted with just q executes $a \in \mathcal{A}$ in $\mathcal{R}$, via the "action adherence" report metric below.

In **reports**, we assess how well $\mathcal{R}$ uses q and $\mathcal{A}$. We use four report quality metrics from ScholarQA-CS2 (Bragg et al., 2026): 1) answer coverage, how many elements a correct answer for q must include are covered in $\mathcal{R}$; 2) answer precision, how directly $\mathcal{R}$ answers q; 3) citation precision, $\mathcal{R}$'s citation accuracy; and 4) citation recall, how often $\mathcal{R}$'s claims are cited. We add action adherence—how often $\mathcal{R}$ follows actions in $\mathcal{A}$ at any point (Qin et al., 2024).

3.3 Baselines

For profiles/actions, we assess LLMs to pick one for MySQA. **Profiles** infer over long contexts (Kuratov et al., 2024) only once, so we evaluate reasoning LLMs: Claude-4 Sonnet+think, o3,[7] Gemini-2.5 Pro (Comanici et al., 2025), and DeepSeek-r1 (DeepSeek-AI et al., 2025). **Actions** are per-query, so we test fast LLMs: Gemini-2.5 Flash, GPT-4.1, Claude-4 Sonnet, and DeepSeek-V3 (DeepSeek-AI et al., 2024). In **reports**, we compare MySQA to open-source/commercial DR that take concatenations of $\langle q \cdot \mathcal{A} \rangle$ as prompts: ScholarQA (Singh et al., 2025), OpenScholar (Asai et al., 2024), STORM (Shao et al., 2024), Perplexity's Sonar[8] and OpenAI o3 DR[9] (details in Appendix A.3).

[7] https://openai.com/index/introducing-o3-and-o4-mini/
[8] https://sonar.perplexity.ai/
[9] https://openai.com/index/introducing-deep-research/

3.4 Offline Results

With our dataset (§3.1), metrics (§3.2), and baselines (§3.3) set, we now test each step in MySQA. In **profiles**, all models but DS-r1 create accurately cited inferences (Table 1) and all accurately categorize inferences. More citations reduces specificity—a common personalization/generalization tradeoff (Han et al., 2022). For MySQA's profiles, we use Gemini-Pro, scoring the highest in three metrics.

All models give personalized **actions** with higher win rates and uniqueness over generic ones, indicating such actions are tailored to profiles (Table 2). Yet, personalized ones conflict queries more; Appendix A.4 shows tension between tailored actions and answering queries, often for papers with low query similarity (§3.1), so personalization may not always help (Zhang et al., 2013). Actions generally had modest adoption rates in our formative (§2.4), so we rotate all 4 models in MySQA for diversity.

In **reports**, MySQA has the best or second-best scores 4/5 times—the most of any DR system (Table 10). It also always beats its base system SQA: our modifications (§2.3) improve its report quality.

Having shown MySQA's output quality offline, we can now deploy the tool online to find flaws and insights our simulation-based data may miss (§4).

Output	Aspect	Description	Freq.
Profile	DOMAIN	Uses terms, definitions, or details that do not capture the user's domain of research.	27.6%
	OVERCLAIM	Claims something applies to the user broadly, but only applies to some/parts of papers.	17.9%
	CONVENTION	Infers a generic convention of the user's field (e.g. "You enumerate contributions").	12.8%
	CONTRAST	Has a contrast that misrepresents the user (e.g. "You are X, not Y", but the user is Y).	12.2%
Action	NARROW	The action is too specific and would overly constrain the information coverage.	43.8%
	OFFTOPIC	The action deviates too far from the query, distracting from the user's goal/intent.	23.6%
Report	UNINFORM	The content is too vague/high-level to be informative; the user wanted more details.	38.0%
	PRESENT	The user wanted the content presented in a different style/format (e.g. bullet points).	25.3%
	IGNORE	One or more implicit/explicit requirements in the action was ignored.	22.8%

Table 4: The nine most common personalization errors in MySQA's outputs discovered from our interviews, missed in our offline evaluations. All twenty metrics we discover are in Table 9.

4 Moving MySQA Offline to Online

MySQA excels on synthetic datasets (§3), but this may not mirror real user needs (Saxon et al., 2024) and even 10+ metrics might not cover all aspects of personalized DR users value (Venkit et al., 2025a).

We thus run a user study (§4.1) to answer open questions: **RQ1**—What personalization errors do offline metrics neglect? (§4.2) and can off-the-shelf LLM judges evaluate them? (§4.3). **RQ2**—What lessons can guide future personalized DR? (§5)

4.1 Interviewing Active Deep Research Users

We interview[10] 21 active DR users (CS researchers on Upwork, $30-40/hr) for 90 minutes who show DR familiarity in a pilot survey; 19 use OpenAI DR (Appendix A.9). Each bring: 1) five papers of interest;[11] and 2) three queries to ask DR. While using MySQA, they screen-record and "think out loud" (Danks et al., 1984). We transcribe recordings for open thematic analysis (Boyatzis, 1998).

In MySQA's UI (§2.4), participants review their profiles (~45 min.) and disable/edit poor inferences. Next, they review every query's proposed actions (~5 min.), picking or editing which ones MySQA should follow. They submit actions to tailor reports, rating its quality and ability to follow actions while verbalizing reasoning for their ratings (~15 min.).

4.2 RQ1: What our Offline Evaluation Missed

Participants liked 73% of profiles, actions, and action executions in reports (Fig 5), so MySQA was useful. We now analyze the remaining 27% to uncover issues, guiding future evaluation: an author reviewed all 1044 judgments and qualitatively labeled where participants were dissatisfied; another verified themes on 60 rows (with 90% agreement).

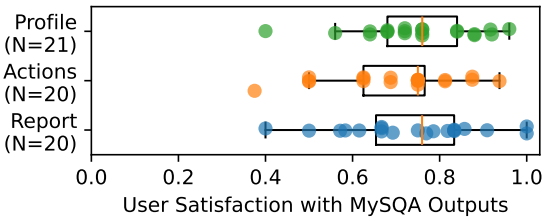

Figure 5: User satisfaction on MySQA profiles, actions, and reports. Users are satisfied with 73% of them.[12]

We find nine common issues (Table 4). Most often, profiles mistake technical terms (DOMAIN), actions are restrictive (NARROW), and reports lack detail (UNINFORM). Many are open NLP problems: factuality (Wang et al., 2024b, DOMAIN), style (Jin et al., 2022, PRESENT), or expertise calibration (Joshi et al., 2025, OVERCLAIM, CONVENTION). Some issues in the full list (Table 9) seem untraceable offline—e.g., accurate but unimportant profile inferences (UNIMPORT) and rejected actions due to mistrust in MySQA's ability (TRUST)—only detectable with online feedback.

Crucially, every issue is invisible to our offline metrics (§3.4)—despite covering 10+ reasonable aspects of personalization—so synthetic datasets can overlook what users value in personalization.

4.3 LLMs Don't Know What DR Users Want

Having found personalized DR aspects we missed offline (§4.2), we now see if we *could have* evaluated them via LLM judges. We answer this with a classifier: given an aspect, were users satisfied by profiles, actions, and action executions in reports for that aspect. For each aspect, we label outputs users dislike as $l = 1$ paired with two negatives ($l = 0$): a random liked output and a "hard" negative (the most stylistically-similar liked output)—

[10] Our internal review board approved our study (§9).

[11] Papers they have written, wish they had written, inspire them, or are relevant to a current project.

[12] U10 spent the interview just viewing the profile ($n = 20$).

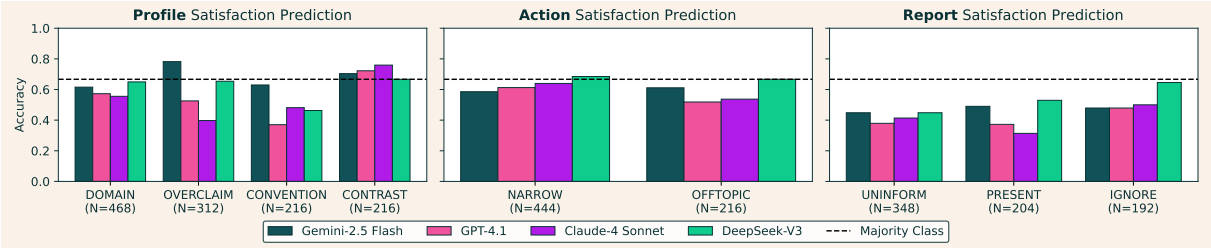

Figure 6: LLM judge accuracy for predicting users' satisfaction of personalized profiles, actions, and reports. No model ever beats a majority class baseline (Dror et al., 2018, $\alpha = 0.05$ Binomial test with Bonferroni correction).

both from the user. LLMs access the same context (e.g. papers, actions, highlights) as users, six few-shot examples, and definitions of each aspect[13] to predict $\hat{l}$: would the user like or dislike the output?

Our four LLM judges predict user satisfaction no better than majority-class baselines (Figure 6), so strong, off-the-shelf judges struggle to capture these issues (§3.4). As offline evaluations can miss what users value—metrics that LLM judges may not predict sans extensive engineering effort—we advocate shifting away from just simulation-based evaluations towards user feedback (Du et al., 2018).

5 RQ2: Lessons for Personalized DR

Beyond surfacing limitations in offline evaluations (§4.3), we now show how online user feedback also offers richer insights for DR. We distill these into four lessons to help NLP and HCI researchers design improved personalized DR tools for users.

Lesson 1: Balance User Control and Effort

Adding control to interactive tools helps, but such effort can exceed what users want to invest (Shneiderman, 1983; Jin et al., 2017). DR interactions are different: as reports take minutes to write, our pilot survey (Appendix A.9) reveals users are willing to dedicate more effort up front to avoid subpar outputs. Most DR tools add this control via clarification or follow-up questions (Jiang et al., 2024), but surveyed users disliked this feature: they must rearticulate their personal needs for every query.

Instead, MySQA creates a persistent profile for all queries—matching users who often "*want [the] system to know about me once then act accordingly*" (U21)—and has users pick among actions—easier than answering follow-up queries (Appendix A.9). Actions still had ample control; U1 noted actions "*are not complex but also very granular which I really like*" and U4 felt they let them "*do what I cannot express in a perfect way.*" U6 felt actions transparently showed how the system answered queries while U17 felt it kept them more actively engaged than existing tools—helping them "*brainstorm to get what I want*". Many believed this process could save time, for U18, "*on the order of days*", since it avoids the frustration of re-prompting/re-answering generic systems that "*keep missing the point*" (U3).

MySQA has more control than most DR tools, which made users realize they wanted more, even at the cost of more effort. Some wanted to see more actions per query (U13, U19), add new actions (U3, U16), or "emphasize" certain actions (U4). Others wanted to also steer DR via paper filters (U2, U11), multi-turn dialogue (U5, U6), and monitoring the tool as it learned their preferences over time (U15)—updating its memory of the user (Yuan et al., 2025).

While more personalized DR control seems helpful, it could risk reinforcing filter bubbles (Zhang et al., 2024a)—keeping users away from new ideas. Control also may not benefit all query types (Dou et al., 2007): U7 desired a toggle to control when MySQA proposes actions. Future work must study not just ways to control personalized DR, but *when* it should be controllable and *to what extent*.

Lesson 2: Make Personalization Easy to Digest

DR inputs and outputs can be hard to navigate, but MySQA's structure made content easy to digest. In most DR tools, users must engineer long queries to express personalized needs (U2, U12): a difficult process for DR users to manage alone (Zamfirescu-Pereira et al., 2023). Instead, MySQA decomposes tailored query writing via two structured phases—profiles and actions—which helped users efficiently organize their needs (U1, U5, U15); U3 noted: "*I have used 3-4 AI tools and none of them have such steps exactly like that in such a structured way.*"

While commercial DR masks if and how personalization occurs (§6.1), our report highlights kept personalization visible/easy to skim: they "*helped*

[13] e.g., for DOMAIN, we prompt "Would the user be satisfied with the technical terminology in this profile inference?"

focus on text" (U12) and were "*easy to cognitively process*" (U5). Still, highlights were tough to get right. Preferences on highlight frequency varied—some wanted full sections (U6), others only "*key numbers*" (U5) or "*evidence*" (U21)—and for more personalized reports with more actions, highlights were overwhelming (U6, U12). We also saw signs of over-trust (Lee and See, 2004); when no content was highlighted for an action, some assumed the content did not exist, rather than an error from our system. DR tools must offer trustworthy sensemaking aids (Zhang et al., 2008), simplifying personalization without overwhelming or misleading users.

Lesson 3: Dream Bigger than Just Papers

While MySQA infers interests just from chosen papers, participants were surprised by profiles' detail and nuance, often noting "*very important positions [they] would claim*" (U3) and "*digging stuff [they] didn't think it can detect*" (U16). This spurred users to ideate further signals for personalizing profiles, like active project materials and prior queries (Appendix A.9) and new personalization use cases, like collaborator search (U9), biography writing (U3), paper reading (U11), and programming (U12).

Participants also wanted modalities beyond text and citations in reports, pointing to a wide variety based on their personal needs, like code snippets for frequent coders (U1, U4) and LaTeX formulas for theory-focused researchers (U2, U4), while others generally desired tables/figures (U5, U7) and interactive visualizations (Mondal et al., 2024, U20). By capturing content preferences, participants felt reports would give a "*better view in a shorter time*" (U12) or a "*holistic view of entire papers*" (U17).

Going beyond papers in personalized DR is a clear win, but has challenges: reasoning over arbitrary user inputs to construct profiles (Zhao et al., 2025a; Shaikh et al., 2025) and routing to the best modality to convey content tailored to a user's background (Chen et al., 2025b). Future work should study not just how to implement these accurately, but to make them helpful for individual researchers.

Lesson 4: Evaluation Isn't One-Size-Fits-All

While we show offline metrics miss aspects of personalized DR (§4.2, §4.3), we caveat online studies are not perfect, final fixes (Hosking et al., 2024); for example, edits on action evince not just what users *want* DR to do but also what they *think* it can do—U6 skipped complex actions until they felt it "*understood the basics*." Similarly, users need to predict utility (Levy, 1992). For instance, participants noted MySQA "*can be a great timesaver*" (U17) and found points they "*miss while writing research papers*" (U7), but such benefits are hard to predict (Balepur et al., 2024, 2025c). Two directions can bolster personalized DR evaluation: longitudinal studies to account for learning effects (Jahani et al., 2024), and richer signals to evaluate downstream helpfulness (e.g., time taken, citations clicked).

We thus advocate for mixed evaluations in personalized DR. For MySQA, formative studies confirmed the workflow met user needs (§2.4) in early stages, while offline metrics formed a scalable way to check baseline report quality, like citation accuracy and action following (§3.4). Online evaluation then found what we missed offline—issues to fix in metrics (§4.3) and how to make DR more useful (§5). Overall, evaluations support distinct parts and stages of personalized system design, showing the need for NLP research in personalization to move beyond offline metrics towards real user feedback (Saxon et al., 2024; Balepur et al., 2025b).

6 Related Work

Given our paper's focus on personalized Deep Research (DR), we review personalization generally in NLP (§6.1) and the advent of DR tools (§6.2).

6.1 Personalization in NLP

Personalization tailors models to user-specific context (Zhang et al., 2024b; Sorensen et al., 2024), which can aid engagement (Kumar et al., 2019), satisfaction (Liang et al., 2006), and learning (Leong et al., 2024)—making its usefulness in DR clear.

Most personalized methods (Zhang et al., 2025): 1) gather data to form user models; and 2) adapt outputs to (1). User models have used demographics (Kirk et al., 2024), preferences (Ryan et al., 2025), and user history (Salemi et al., 2023), while adaptation has retrieved user contexts (Sun et al., 2025), prompted via personas (Lee et al., 2023), and tuned parameters or reward models (Tan et al., 2024; Chen et al., 2025a). MySQA infers user models from papers and adapts reports through prompting.

Our survey of 31 ACL'25 papers (Appendix B) shows NLP stays offline to test methods; 14 see if models match outputs from real users (Salemi et al., 2023), but this assumes LLMs cannot create outputs better than users. 17 works instead tailor models to simulated user chats (Liu et al., 2025a) scored by LLMs, but only ten check judges for reliability.

Even when reliable, the "user" does not exist (Kim et al., 2025a), so such judges may not reflect true user needs. We thus assess personalization online with real users—which only two ACL'25 papers do (Flicke et al., 2025)—revealing aspects of personalized DR LLM judges struggle to capture.

6.2 Deep Research Systems

Scientific Deep Research (DR) systems synthesize long-form reports via documents to aid scientific research (Java et al., 2025). Such systems can create Wikipedia articles (Sauper and Barzilay, 2009; Liu et al., 2018), expository text (Balepur et al., 2023; Jiang et al., 2025b), and surveys (Hu et al., 2024; Yan et al., 2025). DR often employs a mix of retrieval (Lewis et al., 2020) and AI agents (Yao et al., 2023), chaining these models to find, organize, and condense scientific papers (Wang et al., 2024a).

As users find DR useful (Shen et al., 2023), work has started deploying them online to aid researchers and crowdsource user feedback (Zhao et al., 2025b). Open-source UIs/systems include STORM (Shao et al., 2024), SCHOLARQA (Singh et al., 2025), and OPENSCHOLAR (Asai et al., 2024), but as our survey reveals (Appendix A.9), most exposure to DR is commercial (e.g. OpenAI DR, Perplexity).

Few DR tools are personalized. Exceptions are Co-STORM (Jiang et al., 2024) and products like OpenAI/Gemini's DR which ask users follow-up queries, but it is unclear if they tailor to persistent user models. Concurrently, Liang et al. (2025) construct a personalized DR benchmark with synthetic data and LLM judges like us (§3.1), but do not release a new system or run online studies. Instead, MYSQA is an open-source personalized DR tool based on adaptive hypermedia design (Brusilovsky, 1996): building a model of the user from their papers and using it to tailor actions for input queries.

7 Conclusion

Synthetic evaluations rule personalization in NLP, but our paper on personalized Deep Research (DR) shows how online studies with real users help: they reveal errors that LLM judges miss (§4.3) and guiding lessons for future systems (§5). Earlier search engine research used real users to push personalization (Joachims, 2002), but NLP has recently stopped at easy-to-use LLM judges (Zheng et al., 2023) claimed to simulate users (Binz et al., 2024), rarely adopting online studies. We urge readers to fight the temptation: simulated benchmarks are preliminary tests, but real progress in personalization requires real users (Seshadri et al., 2026).

As evidence, our online study points to new open questions in personalized DR, as MYSQA is imperfect. These include NLP modeling efforts in factuality (§4.2) and multi-modality (§5.3), and HCI studies for effortless control (§5.1) and digestible personalization (§5.2). Tackling these will be cyclical: new methods/evaluations will decide when DR meets basic needs, while online studies will reveal what lacks in features/design to truly help users.

Acknowledgments

We would like to thank the Allen Institute for Artificial Intelligence (Ai2) and the CLIP lab at the University of Maryland. In particular, we thank Connor Baumler and Paiheng Xu for reviewing earlier versions of our interface and study design. We appreciate discussions with Doug Downey, Dan Weld, Tal August, Rachel Rudinger, Shi Feng, Matt Latzke, Jonathan Bragg, and Jay DeYoung on earlier versions of our prototype and paper. We thank Dang Nguyen, Navita Goyal, Joy Wongkamjan, Connor Baumler, Paiheng Xu, Yapei Chang, Atrey Desai, and Hyojung Han for voting on candidate titles. Nishant is especially grateful to Yapei Chang, Hita Kambhamettu, Federica Bologna, Peiling Jiang, Xinran Zhao, Michael Noukhovitch, Anej Svete, Alexis Ross, Akhila Yerukola, and Amanda Bertsch for making the summer in Seattle memorable.

This material is based upon work supported by the National Science Foundation under Grant No. DGE-2236417 (Balepur) and IIS-2403436 (Boyd-Graber). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8 Limitations

Due to resource constraints, we could only deploy our MYSQA system online, so our findings on personalization are most directly tied to our setting of Deep Research and execution in MYSQA. However, many of our insights may apply generally; our uncovered issues in MYSQA's outputs (§4, e.g., do not over-claim what applies to a user) and suggestions for future work (§5, e.g., extend modalities) apply in general personalization settings. Further, we are not asserting that these are the only issues to focus on when building personalized systems; as argued in the paper, it is useful to study feedback

from end users of your system (Saxon et al., 2024).

We also note that our LLM judge experiments cannot cover all parameter configurations (§4.3); although we follow best practices in prompt engineering (Schulhoff et al., 2024)—such as adding few-shot examples and giving clear definitions for each metric—perhaps there is another prompt and LLM judge that leads to higher accuracy. In Appendix A.6, we show that changing the number of few-shot examples does not largely improve LLM prediction accuracy, meaning that it may escape off-the-self models, but could be attainable by training reward models on more data (Liu et al., 2024b).

MySQA can be slow—generating profiles takes 3 minutes and reports take 5 minutes—which can harm user experience (Shneiderman, 1984). While faster than some commercial DR tools (e.g., OpenAI DR took 8 hours for 10 queries), improving efficiency would make MySQA more useful. To manage delays, we let users view jokes, fun facts, a Chrome Dino Game,[14] and execution progress in our interface (Appendix A.10), which many users enjoyed. Future work can focus on reducing latency with smaller, efficient models (Gou et al., 2020) or better using other waiting time, like pre-computing reports while users spend time customizing actions.

9 Ethical Considerations

While the MySQA framework poses no risks in theory, we found that in the profile inference stage, even aligned LLMs (Bai et al., 2022) can generate potentially offensive or insensitive inferences. This did not happen with users, but while annotating profiles on our benchmark for agreement, we viewed one such inference: "*Your papers sometimes contain slightly awkward phrasing or non-standard terminology, suggesting a potential non-native English writing background or direct translation of concepts*". This shows that even strong LLMs are susceptible to harmful stereotyping in personalization (Kantharuban et al., 2024), motivating the need for future safeguards to defend against these issues.

We attended each interview with participants to mitigate any risks with our system. Further, our entire formative and interview study designs were approved by our organization's internal review board. We collect and release no PII in our data and ensure participants were fairly compensated between $30–$40/hr, well above our region's minimum wage. In Appendix A.8, we detail our recruitment protocol.

[14] https://nbalepur.github.io/ai2-trex-runner/

We use GenAI for UI design, revisions, and analysis; we detail this in Appendix C for transparency.

References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke S. Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hanna Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *ArXiv*, abs/2411.14199.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30(5):1–38.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025a. Whose boat does it float? improving personalization in preference tuning via inferred user personas. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3371–3393, Vienna, Austria. Association for Computational Linguistics.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025b. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418.

Nishant Balepur, Matthew Shu, Alexander Hoyle, Alison Robey, Shi Feng, Seraphina Goldfarb-Tarrant, and Jordan Lee Boyd-Graber. 2024. A SMART

mnemonic sounds like "glue tonic": Mixing LLMs with student feedback to make mnemonic learning stick. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14202–14225, Miami, Florida, USA. Association for Computational Linguistics.

Nishant Balepur, Matthew Shu, Yoo Yeon Sung, Seraphina Goldfarb-Tarrant, Shi Feng, Fumeng Yang, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025c. A good plan is hard to find: Aligning models with preferences is misaligned with what helps users. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11568–11595, Suzhou, China. Association for Computational Linguistics.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brandle, Frederick Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, No'emi 'EltetHo, Thomas L. Griffiths, Susanne Haridi, Akshay Kumar Jagadish, Ji-An Li, Alex Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo G. Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmuš, Evan M. Russek, Tankred Saanum, Natalia Scharfenberg, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert C. Wilson, Kristin Witte, Shuchen Wu, Dirk Wulff, Huadong Xiong, and Eric Schulz. 2024. Centaur: a foundation model of human cognition. *ArXiv*, abs/2410.20268.

Richard E. Boyatzis. 1998. Transforming qualitative information: Thematic analysis and code development.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S Weld. 2026. Astabench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*.

Peter Brusilovsky. 1996. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6:87–129.

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors. 2025. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. 2025a. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*.

Jiaqi Chen, Yanzhe Zhang, Yutong Zhang, Yijia Shao, and Diyi Yang. 2025b. Generative interfaces for language models.

Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

Joseph H. Danks, K. Anders Ericsson, and Herbert A. Simon. 1984. Protocol analysis: Verbal reports as data.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jiong Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, Ruiqi Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaokang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, Wangding Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xi aokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui

Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jun-Mei Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, Ruiqi Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shao-Ping Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, Wangding Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xuan Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang, Xiaoling Bi, Xiaodong Liu, Xiaohan Wang, Xi-Cheng Shen, Xi aokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yao Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yi-Bing Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxiang Ma, Yuting Yan, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *The Web Conference*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Fan Du, Sana Malik, Georgios Theocharous, and Eunyee Koh. 2018. Personalizable and interactive sequence recommender system. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA. Association for Computing Machinery.

Markus Flicke, Glenn Angrabeit, Madhav Iyengar, Vitalii Protsenko, Illia Shakun, Jovan Cicvaric, Bora Kargi, Haoyu He, Lukas Schuler, Lewin Scholz, et al. 2025. Scholar inbox: Personalized paper recommendations for scientists. *arXiv preprint arXiv:2504.08385*.

Zhaolin Gao, Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. End-to-end training for recommendation with language-based user profiles. *ArXiv*, abs/2410.18870.

Cristina Garbacea and Chenhao Tan. 2025. Hyperalign: Interpretable personalized llm alignment via hypothesis generation. *ArXiv*, abs/2505.00038.

Jianping Gou, B. Yu, Stephen J. Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789 – 1819.

Dong-Jun Han, Do-Yeon Kim, Minseok Choi, Christopher G. Brinton, and Jaekyun Moon. 2022. Splitgp: Achieving both generalization and personalization in federated learning. *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pages 1–10.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024. Taxonomy tree generation from citation graph. *arXiv preprint arXiv:2410.03761*.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, et al. 2025. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*.

Hilary Browne Hutchinson, Wendy E. Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane

Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *International Conference on Human Factors in Computing Systems*.

Eaman Jahani, Benjamin Manning, Joe Zhang, Hong Yi Tu Ye, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz. 2024. As generative models improve, people adapt their prompts. *People Adapt Their Prompts*(July 18, 2024)*.

Abhinav Java, Ashmit Khandelwal, Sukruta Prakash Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. 2025. Characterizing deep research: A benchmark and formal definition. *ArXiv*, abs/2508.04183.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. 2025a. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *ArXiv*, abs/2504.14225.

Junjie Jiang, Haodong Wu, Yongqi Zhang, Songyue Guo, Bingcen Liu, Caleb Chen Cao, Ruizhe Shao, Chao Guan, Peng Xu, and Lei Chen. 2025b. Archidocgen: Multi-agent framework for expository document generation in the architectural industry. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 605–618.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J Semnani, and Monica S Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv preprint arXiv:2408.15232*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Yucheng Jin, Bruno Cardoso, and Katrien Verbert. 2017. How do different levels of user control affect cognitive load and acceptance of recommendations? In *Jin, Y., Cardoso, B. and Verbert, K., 2017, August. How do different levels of user control affect cognitive load and acceptance of recommendations?. In Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2017)*, volume 1884, pages 35–42. CEUR Workshop Proceedings.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. 2025. Eli-why: Evaluating the pedagogical utility of language model explanations. In *Annual Meeting of the Association for Computational Linguistics*.

Anjali Kantharuban, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. 2024. Stereotype or personalization? user identity biases chatbot recommendations. *ArXiv*, abs/2410.05613.

Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. 2025a. Cupid: Evaluating personalized and contextualized alignment of llms from interactions. *arXiv preprint arXiv:2508.01674*.

Yoonsu Kim, Brandon Chin, Kihoon Son, Seoyoung Kim, and Juho Kim. 2025b. Intentflow: Interactive support for communicating intent with llms in writing tasks. *ArXiv*, abs/2507.22134.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Vipin Kumar, Bharath Rajan, Rajkumar Venkatesan, and Jim Lecinski. 2019. Understanding the role of artificial intelligence in personalized engagement marketing. *California management review*, 61(4):135–155.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Y. Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *ArXiv*, abs/2406.10149.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 46:50 – 80.

Joosung Lee, Minsik Oh, and Donghun Lee. 2023. P5: Plug-and-play persona prompting for personalized

response selection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16571–16582, Singapore. Association for Computational Linguistics.

Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting things into context: Generative ai-enabled context personalization for vocabulary learning improves learning motivation. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Jack S Levy. 1992. An introduction to prospect theory. *Political psychology*, pages 171–186.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Sheena Lewis, Mira Dontcheva, and Elizabeth Gerber. 2011. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 735–744.

Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. 2006. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, 23(3):45–70.

Yuan Liang, Jiaxian Li, Yuqing Wang, Piaohong Wang, Motong Tian, Pai Liu, Shuofei Qiao, Runnan Fang, He Zhu, Ge Zhang, et al. 2025. Towards personalized deep research: Benchmarks and evaluations. *arXiv preprint arXiv:2509.25106*.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*.

Guanyu Lin, Tao Feng, Pengrui Han, Ge Liu, and Jiaxuan You. 2024. Paper copilot: A self-evolving and efficient llm system for personalized academic assistance. *ArXiv*, abs/2409.04593.

Dongshuo Liu, Zhijing Wu, Dandan Song, and Heyan Huang. 2025a. A persona-aware LLM-enhanced framework for multi-session personalized dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 103–123, Vienna, Austria. Association for Computational Linguistics.

Javin Liu, Aryan Vats, and Zihao He. 2025b. Cs-papersum: A large-scale dataset of ai-generated summaries for scientific papers. *ArXiv*, abs/2502.20582.

Jiahao Liu, Yiyang Shao, Peng Zhang, Dongsheng Li, Hansu Gu, Chao Chen, Longzhi Du, Tun Lu, and Ning Gu. 2024a. Filtering discomforting recommendations with large language models. *Proceedings of the ACM on Web Conference 2025*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *ArXiv*, abs/2410.16184.

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2664–2684, St. Julian's, Malta. Association for Computational Linguistics.

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. 2025. The realhumaneval: Evaluating large language models' abilities to support programmers. *Transactions on Machine Learning Research*. Expert Certification.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *ArXiv*, abs/2402.09906.

Sheshera Mysore, Mahmood Jasim, Andrew Mccallum, and Hamed Zamani. 2023. Editable user profiles for controllable text recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 993–1003, New York, NY, USA. Association for Computing Machinery.

Jakob Nielsen. 1993. Usability engineering. In *The Computer Science and Engineering Handbook*.

Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A Huberman, Kimmo Kaski, and Santo Fortunato. 2015. Attention decay in science. *Journal of Informetrics*, 9(4):734–745.

Kevin Pu, KJ Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–31.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *ArXiv*, abs/2401.03601.

Michael J Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. Synthesizeme! inducing persona-guided prompts for personalized reward models in llms. *arXiv preprint arXiv:2506.05598*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *ArXiv*, abs/2304.11406.

Alireza Salemi and Hamed Zamani. 2025. Lamp-qa: A benchmark for personalized long-form question answering. *ArXiv*, abs/2506.00137.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216.

Michael Stephen Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. *ArXiv*, abs/2407.16711.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Minh Pham, Gerson C. Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Miserlis Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques. *ArXiv*, abs/2406.06608.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations. In *Algorithmic Fairness Across Alignment Procedures and Agentic Systems*.

Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S. Bernstein. 2025. Creating general user models from computer use. *ArXiv*, abs/2505.10831.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *North American Chapter of the Association for Computational Linguistics*.

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv:2304.02623*.

Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer*, 16:57–69.

Ben Shneiderman. 1984. Response time and display rate in human performance with computers. *ACM Comput. Surv.*, 16:265–285.

Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 scholar QA: Organized literature synthesis with attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 513–523, Vienna, Austria. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian R. Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *ArXiv*, abs/2402.05070.

Neha Srikanth, Jordan Boyd-Graber, and Rachel Rudinger. 2026. Discotrace: Representing and comparing answering strategies of humans and llms in information-seeking question answering.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.

Xiangru Tang, Xingyao Zhang, Yanjun Shao, Jie Wu, Yilun Zhao, Arman Cohan, Ming Gong, Dongmei Zhang, and Mark Gerstein. 2024. Step-back profiling: Distilling user history for personalized scientific writing. *ArXiv*, abs/2406.14275.

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. 2025a. Deeptrace: Auditing deep research ai systems for tracking reliability across citations and evidence.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025b. Search engines in the ai era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in llm-based search. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. 2024a. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi N. Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Conference on Empirical Methods in Natural Language Processing*.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Ruifeng Yuan, Shichao Sun, Yongqi Li, Zili Wang, Ziqiang Cao, and Wenjie Li. 2025. Personalized large language model assistant with evolving conditional memory. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3764–3777, Abu Dhabi, UAE. Association for Computational Linguistics.

J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Michael JQ Zhang, W. Bradley Knox, and Eunsol Choi. 2025. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations*.

Weinan Zhang, Jun Wang, Bowei Chen, and Xiaoxue Zhao. 2013. To personalize or not: a risk management perspective. *Proceedings of the 7th ACM conference on Recommender systems*.

Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. 2008. Citesense: supporting sensemaking of research literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 677–680, New York, NY, USA. Association for Computing Machinery.

Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024a. See widely, think wisely: Toward designing a generative multi-agent system to burst filter bubbles. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Ju-Ying Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2024b. Personalization of large language models: A survey. *ArXiv*, abs/2411.00027.

Dora Zhao, Diyi Yang, and Michael S. Bernstein. 2025a. Knoll: Creating a knowledge ecosystem for large language models. *ArXiv*, abs/2505.19335.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, et al. 2025b. Sciarena: An open evaluation platform for foundation models in scientific literature tasks. *arXiv preprint arXiv:2507.01001*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

A Appendix

A.1 Offline Dataset Details

The ASTABENCH, ScholarQA-CS2 (Bragg et al., 2026) and CS-PaperSum (Liu et al., 2025b) datasets we build off of on are publicly accessible and used within their intended use. To retrieve the full text of papers for offline experiments, we use an internal database to our organization that already collected them; in our interface to handle new papers, we use the Semantic Scholar Snippets API.[15] Our dataset is in English and has no PII that is not already publicly accessible (i.e. information present in a research paper). We summarize our dataset in Table 5.

A.2 Offline Metric Human Agreement

To develop offline metrics (§3.2) we created an initial version of our metrics and ran it on a subset of our dev set; one author then blindly gave ratings on 30–50 examples—half where the model predicted 1 and half where the model predicted 0—for agreement. For specificity, which uses a 1–5 Likert rating, we sample 10 examples over each predicted rating. If the metric had agreement < 0.7, we modify our prompts based on model errors, and repeat the process until we reach substantial agreement; this process typically took two rounds per metric.

For profiles, our agreements are: 88.3% for inference accuracy, 88.8% for relevance, 96.7% for category accuracy, and 0.66 Person's correlation for specificity. For actions, our agreements are: 93.3% for personalization win rate and 82.0% for relevance. For reports, our agreement is: 86% for action adherence; the other report metrics were already tested by Bragg et al. (2026). Appendix A.11 contains the prompts used for metrics.

A.3 Offline Experiment Setup Details

To create profiles and actions, we use 1.0 temperature and a max token length of 40960. All commercial LLMs are accessed via their original providers; we access DeepSeek models from TogetherAI.[16]

For generating reports, we use the original hyperparemeters for each model; MYSQA uses SCHOLARQA's hyperparameters (Singh et al., 2025). For a fairer comparison, we let OPENSCHOLAR (Asai et al., 2024) use Claude over its trained model, just like MYSQA, and let STORM (Shao et al., 2024) use Semantic Scholar as its retriever versus Google. Perplexity and OpenAI do not directly provide snippets of the sources in the web pages that they use, so we prompt these models to extract them from the web pages; these may be fabricated,[17] but we take them in good faith so models have the best possible chance. Report evaluation uses InspectAI[18] from ASTABENCH which needs a specific JSON format. We prompt Gemini to parse OpenAI's/Perplexity' reports into said format, like Bragg et al. (2026).

We allocate 72 CPU hours for each experiment run. All results are from a single run.

A.4 Queries versus Personalized Actions

In §3.4, we show personalized actions are less relevant to the query than generic ones. To learn why, we run the metric segmented by the similarity of the author to the query (i.e. low, medium, high). We consistently find personalized actions have more conflicts for low author similarities; for example, DeepSeek's relevance is 0.765, 0.72, 0.69 for high, medium, and low similarity authors, respectively. We believe such tension in queries and personalization is natural; for example, an author working on reasoning LLMs may see less benefit from personalization when asking about NLP+education.

Overall, as query conflicts are infrequent and our main goal was to see if this was truly a problem via online evaluations (§4), we proceeded anyway.

A.5 Model Ablations

We now ablate our design choices in MYSQA. We first consider the best place to show MYSQA the actions $\mathcal{A}$ in the execution prompts (Table 6): at all execution prompts, just the individual actions that are relevant (e.g. "Search for papers" actions only get shown in retrieval), the individual actions that are relevant and all prior actions, or ignoring multi-step execution entirely and generating the report in a single prompt; the former achieves the highest report quality and action adherence.

We then test LLM to power MYSQA (Table 7): Claude-4 Sonnet, Gemini 2.5 Flash, and GPT-4.1; all have similar scores, so we use Claude-4 to match Singh et al. (2025) in SCHOLARQA. Finally, we test how many actions MYSQA can use (Table 8); on 4–30 steps, MYSQA maintains report quality with only minor drops in action adherence scores.

[15] https://www.semanticscholar.org/product/api

[16] https://www.together.ai/

[17] A manual spot check showed models were pulling quotes from real web pages.

[18] https://github.com/UKGovernmentBEIS/inspect_ai

A.6 Simulation Prompt Variations

To further study how LLM judges predict user satisfaction, we test several prompt variations. We first decrease the number of few-shot examples from six (Figure 6) to three (Figure 7) and zero (Figure 8). LLM judges are still unreliable and adding exemplars does not boost accuracy, suggesting off-the-shelf LLM judges struggle with this task regardless of the number of few-shot examples; we suggest future work to train custom reward models for user satisfaction with real user data if they want to better simulate this offline (Lambert et al., 2025).

Lastly, we test if our metric definitions distract LLM judges (e.g. "Would the user be satisfied by the terms in this profile inference?") and would be much more accurate if they directly predict user satisfaction (i.e. "Would the user be satisfied with this profile inference?"). Figure 9 shows removing metric definitions does not boost LLM judge accuracy. Appendix A.11 has prompts used in the experiment.

A.7 Formative Study Details

We reach out to candidates for our formative study via email from an internal mailing list of DR users (more de-anonymized details will be released post-acceptance). We confirm they are active DR users over email. Three participants were M.S. students and two were Ph.D. students in North America, Europe, and Asia. Each interview was conducted in English and all participants had fluency in English. In Figure 10, we provide all questions we ask participants in the formative study (§2.4).

A.8 Interview Recruitment Protocol

This section provides more details on the protocol for our 90-minute usability studies, including annotator instructions (Appendix A.8.1), Recruitment (Appendix A.8.2), Consent (Appendix A.8.3), and Demographics (Appendix A.8.4).

A.8.1 Annotator Instructions

We first asked each participant to fill out a Google Form to gain insights on their knowledge and proficiency with deep research systems (§A.9). After reviewing their responses, we scheduled a 90 minute meeting with each participant to walk them through example annotations and understand their thought process. We asked participants to provide the following in advance of the meeting:

1. A set of 5 research papers on semantic scholar (the paper URLs) that you feel best represent your research interests. These are likely papers you have written, wish you had written, are relevant to a project you are working on, or have been highly influential to you as a researcher. Please make sure these are open-access PDFs! (i.e. the semantic scholar link will have the button "[PDF] Semantic Scholar" if it's valid). The should be in the form "url/-paper/title/id" on Semantic Scholar

2. Three research queries that you would be interested in asking our Deep Research System. These do not have to be about the papers you have selected.

3. Create a new Gmail account with the following credentials: a) Username: [provided email], b) Password: [anything], c) Name: [anything but your own name]. You do not need to use this email to join the Google Meet, but you will need it to log into our interface. You will use this account for both the pilot and any future tasks, and this will ensure no personally identifiable information is shared. You do not need to share your password, but please ensure that you use the username assigned to you.

We did not collect any PII from participants (all information collected is publicly available), and the study does not include any risks. Upon recruitment, we provided additional guidelines/task instructions for query selection and annotation (plan selection, rating responses, feedback, etc.).

A.8.2 Recruitment

We recruited participants via a job post on Upwork.[19] In the job post, we provided a description of the job, screening/pilot process, target domain (CS), estimated weekly time commitment (less than 30 hr/week for 1-3 months), compensation range (USD $30-40/hr), project agreement terms, and screening questions. The compensation range was determined based on previous projects involving annotators with similar backgrounds and which in turn is based on the typical hourly rate for annotation tasks requiring SME (>$30). We did not explicitly state our target education level (PhDs and PhD candidates) in the job posting, but asked for this as a screening question. Candidates submitted proposals (cover letter, answers to screening questions, and desired compensation) which we

[19] https://www.upwork.com/

reviewed, used to shortlist them, and send them an offer based on their proposal. After review, we scheduled paid 90-minute interviews for shortlisted candidates. Compensation was determined by the hours logged by each participant on a weekly basis — regardless of whether we decided to move forward with a candidate's application, we asked them to log hours for their time (form and interview) so that we could compensate them.

A.8.3 Consent

Our Upwork job description/offer provided a clear description of the task and what kind of data would be collected. Furthermore, the offer contained a project participation agreement and along with a statement that by accepting the offer, participants also implicitly accepted the terms of the attached agreement. The participation agreement is used for all of the organization's Upwork annotation projects and explains in detail what kind of data may be collected and how it may be used.

A.8.4 Demographics

Participants in our interviews were from varied countries in North America, Europe, and Asia with varying educational backgrounds—some were enrolled in MS/Ph.D. programs, others had already obtained their Ph.D., and some were already working in research-focused industry positions. All had expertise in various areas of computer science, including machine learning, computer vision, natural language processing, security, and AI+Society. Each interview was conducted in English and all participants had fluency in English. Our data will not release any protected demographic information.

A.9 Participant Survey Results

To ensure participants are active DR users, we first survey their DR usage; 56.6% use DR daily, 26.1% a few times a week, and others a few times a month. Most participants use OpenAI, Gemini, or Perplexity; only one used ScholarQA, reducing the risk of overly-positive feedback on MySQA. All but one participant valued DR having more knowledge about them, which they felt could best be uncovered via their authored/read papers and documentation for projects. Most wanted the system to adapt over time and tailor knowledge per query.

Participants wanted to control all execution steps of DR (paper search, section planning, generation), preferring picking among actions and flagging low-quality queries as forms of control. Less than half wanted to answer follow-up questions to control the system, despite their prominence in OpenAI/Gemini. Most DR use cases spanned learning, writing, experimentation, literature review, and ideation.

After acceptance, we will provide a link to the survey questions and responses (de-anonymized).

A.10 Full Interface Screenshots

In Figure 11, Figure 12, and Figure 13, we provide screenshots with more examples of profiles, actions and reports in MySQA, respectively. Since profile generation takes ~ 5 minutes, we provide various forms of entertainment as loading bars (Figure 14): jokes, trivia facts, and the Chrome dinosaur game. Many participants said that these were fun features.

A.11 Prompts

We now show most prompts we create, but we refer readers to our Github[20] to view them more easily.

For MySQA (§2), Prompts D.1 and D.3 contain the instructions for generating profiles and personalized actions; the prompt for generic actions is similar, but removes $\mathcal{P}$ and any mention of it. The prompts for generating reports mimic the ones used in ScholarQA (Singh et al., 2025), but with extra instructions showing the model where to use p; these are best viewed on our Github repository. Appendix A.10 has output examples from MySQA.

For offline evaluation (§3), Prompts D.7, D.6, D.8, and D.9 contain instructions for evaluating profiles. Prompts D.10 and D.11 contain instructions for evaluating actions. Prompt D.12 contains instructions for evaluating action adherence in the report, which is also used for action uniqueness.

For the user satisfaction/simulation experiments (§4.3), Prompts D.13, D.14, D.15 contain instructions for testing if LLM judges can predict if users would be satisfied with profile inferences, actions, and their execution in reports, respectively; the examples are for OVERCLAIM, OFFTOPIC, and UNINFORM metrics (Table 9), respectively.

B Surveying ACL Personalization Work

To ground our claim that online studies are neglected in NLP for personalization evaluation, we survey work published in the main conference and findings of ACL'25 (Che et al., 2025). We search through the full proceedings[21] for papers with titles

[20] This will be released upon acceptance

[21] https://aclanthology.org/events/acl-2025/

containing the terms "personalization" or "personalized", yielding 43 in total. One author reads each paper and removes 12 papers that do not introduce a new method or evaluation to improve model personalization, leaving 31 in total. In each paper, the same author labels: 1) whether the paper primarily uses real user data or synthetic data (often generated by an LLM); 2) whether the paper uses LLM judges for evaluation; 3) whether outputs are validated by humans, but not the same human related to the personalization context; and 4) whether the paper runs an online study with real users providing inputs and feedback on the personalized outputs.

As summarized in §6.1, 17 papers primarily use real user data, like in LaMP (Salemi et al., 2023; Salemi and Zamani, 2025), while 14 rely mainly on synthetic data—either by prompting LLMs or by mixing attributes (e.g. demographics). 18 papers use LLM-as-a-judge (Zheng et al., 2023) as a metric, but shockingly, only 10 of these papers validate the reliability of the judge. Finally, only two papers run online studies for personalization; Balepur et al. (2025a) asks eight users to provide personas for input queries and rate how well a query was answered and the personalization of the final response for their persona, while Flicke et al. (2025) build ScholarInbox—ironically also a tool for scientific literature recommendation—and check 1233 participants' user satisfaction with the system via 1–5 Likert ratings. The above online studies are also not that rigorous in terms of discovering aspects of personalization that matter to users, so we hope future work keeps assessing personalization online.

C Generative AI Usage Statement

Generative AI (GenAI) was used in several stages of this project. We use Cursor[22] to rapidly prototype our UI, Gemini-2.5 Pro to parse our interview transcripts, GPT-5 to modify our plots and refine paper writing for brevity, and MySQA for paper search in the related work. We check GenAI outputs before adopting them (e.g. dog-fooding any Cursor-generated UI, qualitative validation for transcript parsing). We never use GenAI for writing experimentation code, qualitatively coding data, or writing text from scratch. We take full responsibility for any issues stemming from GenAI errors.

By explicitly discussing GenAI usage here, we aim to encourage other researchers to do the same.

[22] https://cursor.com/agents

D Personalize the Title of this Paper

We were torn between several titles for this paper and spent many hours considering alternatives, so we wanted to discuss other options here. We always wanted the second half of the title to be "Evaluating Personalization in Deep Research Needs Real Users", but for the first half, we also considered:

1. LLMs Don't Know You
2. LLMs Shallowly Know You
3. LLMs Are Shallow Simulators
4. LLM Judges Swim in Shallow Waters
5. LLM Judges Stay in the Shallow End
6. LLM Judges Don't Take it Personally

By displaying these candidate titles (actions), we hope you the reader can also meta-personalize this DR paper (report) for your specific preferences.

Split	# Queries	Avg # Papers / Query	Avg Query Length	Avg Paper Length	Total Instances
Dev	281	2.81	17.42	5855.51	281
Test	291	2.91	13.10	5742.30	291

Table 5: Details of our collected synthetic dataset

Model	Ingredient Recall	Answer Precision	Citation Accuracy	Citation Recall	$\mathcal{A}$ Adherence
See All Actions	0.901	0.917	0.918	0.806	0.851
See Incremental Action	0.892	0.896	0.905	0.786	0.868
See Current Action	0.891	0.915	0.912	0.782	0.798
One-Shot Prompt	0.803	1.00	0.844	0.77	0.648

Table 6: Ablation of MYSQA across different strategies of where to include the actions in prompts. The dominant strategy is always giving MYSQA the actions at each execution step.

Model	Ingredient Recall	Answer Precision	Citation Accuracy	Citation Recall	$\mathcal{A}$ Adherence
MYSQA - Claude Sonnet	0.901	0.917	0.918	0.806	0.851
MYSQA - Gemini Flash	0.87	0.989	0.953	0.846	0.823
MYSQA - GPT-4.1	0.911	0.97	0.856	0.617	0.921

Table 7: Ablation of different LLMs to power MYSQA. LLMs have similar scores, so we choose Claude to match Singh et al. (2025).

Model	Ingredient Recall	Answer Precision	Citation Accuracy	Citation Recall	$\mathcal{A}$ Adherence
4 actions	0.901	0.917	0.918	0.806	0.851
8 actions	0.914	0.899	0.918	0.814	0.832
12 actions	0.927	0.896	0.91	0.81	0.812
24 actions	0.938	0.885	0.915	0.816	0.788
30 actions	0.947	0.898	0.916	0.807	0.777

Table 8: Ablation of MYSQA across various number of actions. The model preserves report quality, but the proportion of actions followed drops as the number of actions increase.

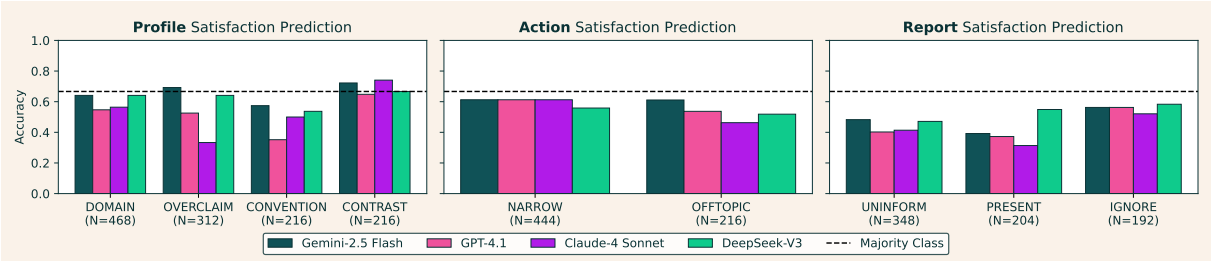

Figure 7: Accuracy of LLM judges for predicting user satisfaction in personalization aspects over profiles, actions, and reports with **one few-shot example**. LLM judges are still worse than the majority class baseline.

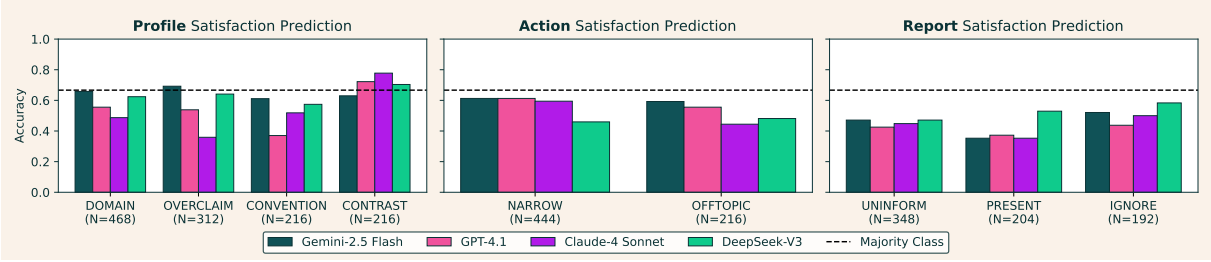

Figure 8: Accuracy of LLM judges for predicting user satisfaction in personalization aspects over profiles, actions, and reports with **no examples**. LLM judges are still worse than the majority class baseline.

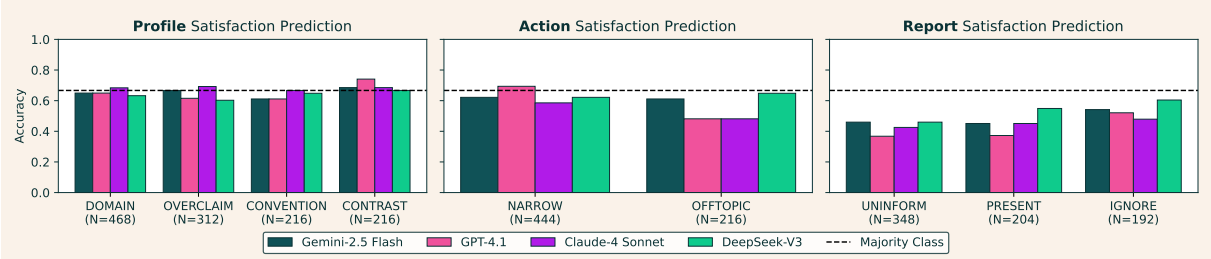

Figure 9: Accuracy of LLM judges for predicting user satisfaction in personalization aspects over profiles, actions, and reports with **no metric definitions**. LLM judges are still worse than the majority class baseline.

Output Type	Aspect	Description	Freq
Profile	**DOMAIN**	Uses terms, definitions, or details that do not capture the user's domain of research	27.6%
	OVERCLAIM	Claims something applies to the user broadly, but only applies to some/parts of papers	17.9%
	CONVENTION	Infers a generic convention of the user's field (e.g. "You enumerate contributions")	12.8%
	CONTRAST	Has a contrast that misrepresents the user (e.g. "You are X, not Y", but the user is Y)	12.2%
	UNIMPORT	The inference is true from their papers, but not a part of their papers they care about	8.3%
	STRENGTH	States something too strongly about the user (e.g. "You are a deep expert in X")	7.1%
	STYLE	The user wanted the inference to use a different tone or style (e.g. formality)	5.8%
	IMPOSSIBLE	States something likely untrue for anyone (e.g. "You can prove P=NP")	5.1%
	GLOBAL	The current inference contradicts or repeats another inference in the profile	3.2%
Action	**NARROW**	The action is too specific and would overly constrain the coverage of information	43.8%
	OFFTOPIC	The action deviates too far from the query, distracting from the user's goal/intent	23.6%
	TRUST	The user does not trust the system could execute the action	13.5%
	VAGUE	The user does not understand how the action would alter the report	7.9%
	EXPERT	The user is an expert and does not want to see basic actions (e.g. define terns)	5.6%
	GLOBAL	The current action contradicts or repeats another action in the list	5.6%
Report	**UNINFORM**	The content is too vague/high-level to be useful, as the user wanted more details	38.0%
	PRESENT	The user wants the content presented in a different style/format (e.g. bullet points)	25.3%
	IGNORE	One or more implicit/explicit requirements in the action is ignored in execution	22.8%
	FACTUAL	The report hallucinates, mis-cites, or makes a factually incorrect statement	10.1%
	GLOBAL	The report contradicts or repeats itself across sections	3.8%

Table 9: Dimensions of personalization errors in MySQA's outputs uncovered in our interviews, missed by offline evaluation (§3.4). **Bold text** indicates aspects with sufficient data (50+ examples) for evaluating if LLM judges can simulate them (§4.3).

Beforehand ask them for:

1. Representative research papers
2. Some queries to ask the DR system (2 examples)
3. A tl;dr of their own biography

Preliminaries:

1. How do you currently use AI tools to support scientific research, like literature review and brainstorming?
2. Have you ever wished these tools could better understand your background or research style? If so, what would you want them to know?
3. How important is it for you to understand and control how a model personalizes to you?

Framework [walkthrough]:

1. Does this approach align with how you'd want personalization to work? Would it save you any time?
2. What kinds of research tasks would you want to use this for?
3. Why did you pick these papers?
4. Besides your research papers, is there any other information you think would help the system understand you better?

Profile Generation [walkthrough]:

1. Would you be open to opting-in for storing an editable version of this profile on MyScholarQA?
2. If improved, would you want to see this kind of profile for other researchers? Would you let other researchers see this profile of yourself?
3. Could you see this being useful in other applications beyond MyScholarQA?

Plan + Response Generation [walkthrough]:

1. Did you like having the option to view and select potential ways to personalize the query? Would it save you time?
2. Which of the requirements did MyScholarQA do a good and bad job at following? Did any deviate from what you were expecting?
3. Did the highlights make it easier to tell where the model attempted to personalize? Would you like this highlighting to be more or less clear if it was built into the prototype?

Conclusions:

1. After seeing this in practice, what do you like and dislike about the prototype?
2. Is there anything else we should note while building such a personalized system?

Figure 10: List of questions we ask participants in the formative. Sections marked with "[walkthrough]" indicates that participants also gave feedback on model outputs or high-level design in addition to answering the questions. During the study, we refer to the list of actions as "plans", which users found clear.

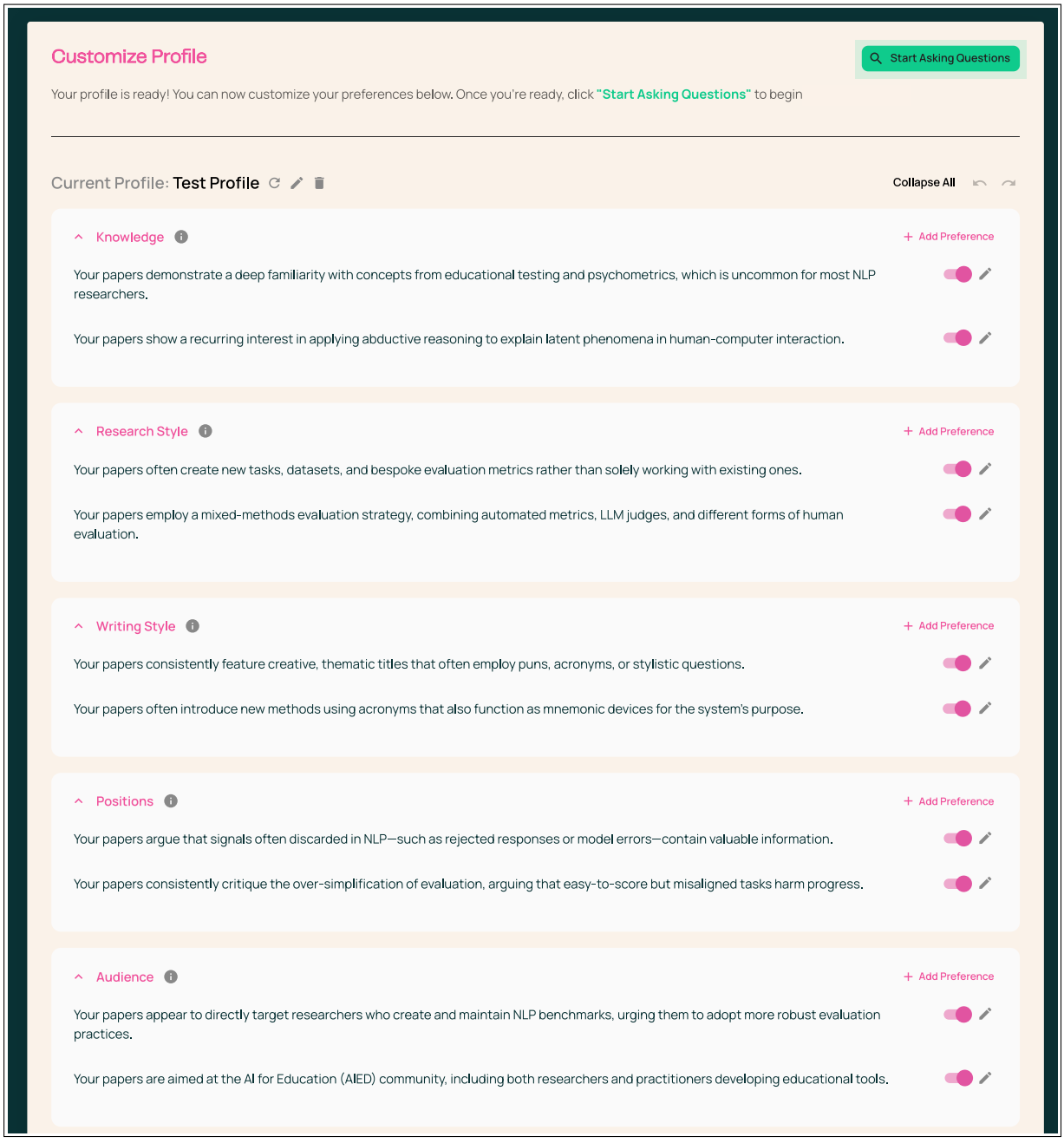

Figure 11: Full example of one of the author's profiles with two inferences per section in MYSQA. We remove the paper citations after each profile inference for anonymity. Users can view, toggle, and edit their profile inferences.

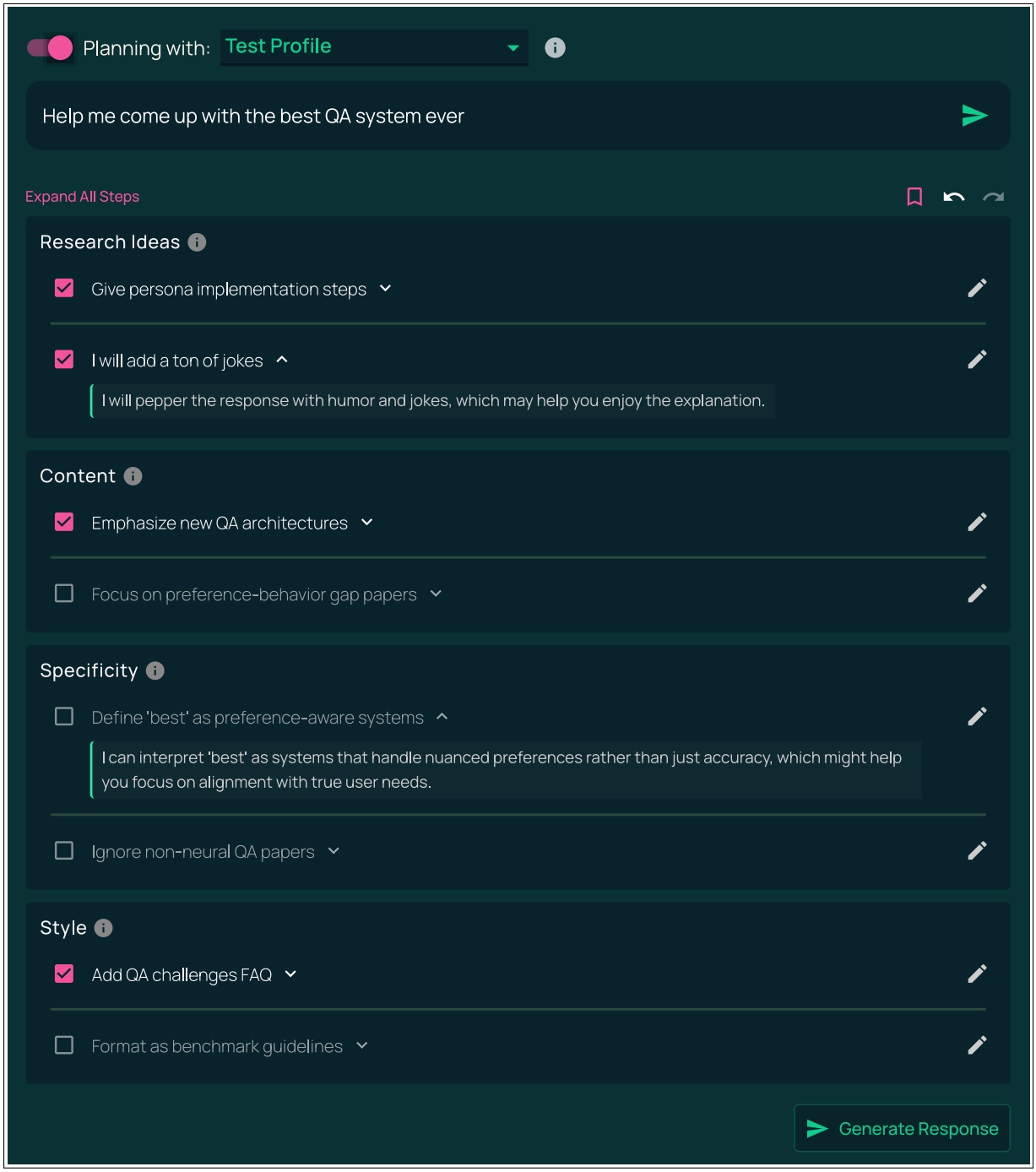

Figure 12: Full example of actions for a query in MySQA. Users can view, toggle, and edit actions in the list. Clicking on the dropdown arrow gives a more elaborate description of each action.

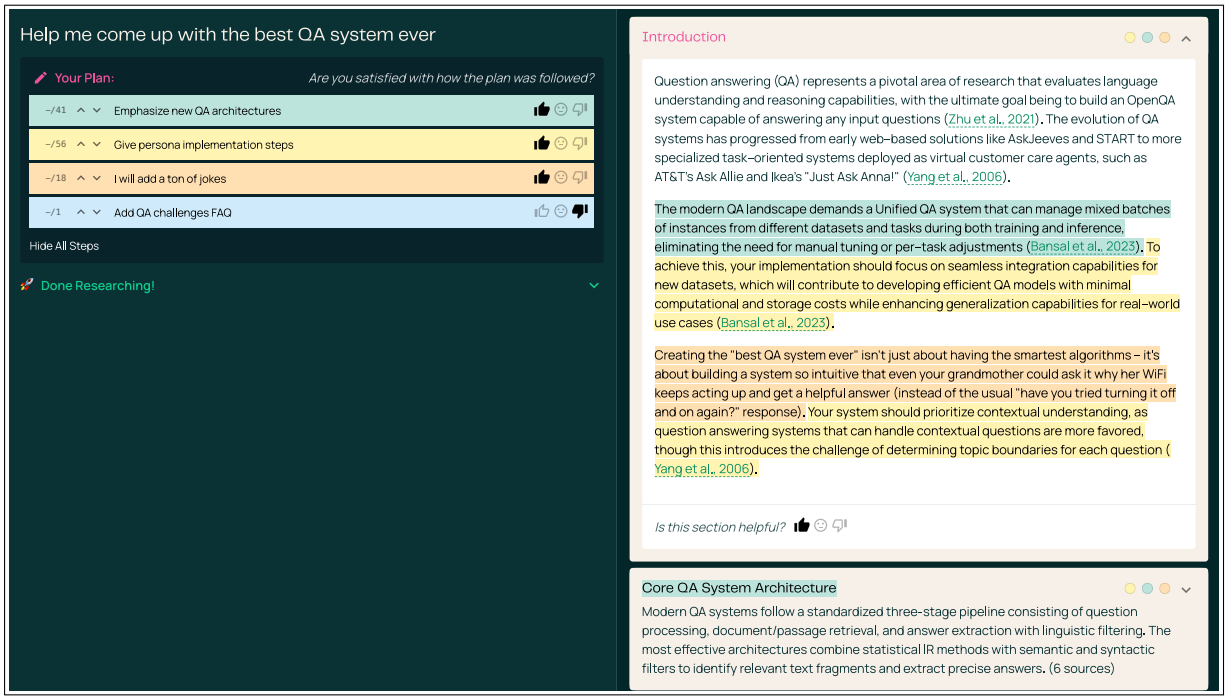

Figure 13: Full example of a report for a query and actions in MySQA. Users can view each section of the report, collapse them for just a TL;DR, and view highlights for where MySQA personalized. Clicking on an action (left) enables/disables the highlights, revealing an action bar with: 1) the number of highlighted spans in the report; and 2) navigation arrows to jump between highlighted spans. Each plan action and section has buttons which we use to collect feedback.

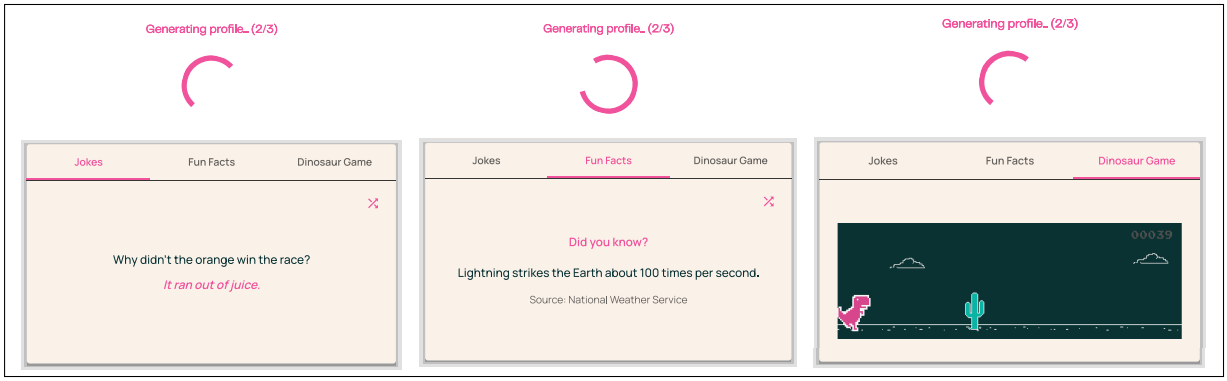

Figure 14: We provide activities to entertain users while they wait for their profiles to be generated. Users can view jokes, fun facts, or play the Chrome T-Rex game. Adding engaging activities before annotation tasks can have positive outcomes (Lewis et al., 2011).

Prompt D.1: Profile Generation Prompt (§2.1)

Here are a list of paragraphs from research papers that the author has selected (either written or read):
<papers> $\mathcal{P}$ </papers>

Based on the papers, generate a list of inferences about the author of the paper, categorized as:
<rubric> rubric </rubric>

<format instructions> Generate your output as a JSON object with 5 keys: "knowledge", "research_style", "writing_style", "positions", "audience"—where each key has a list of inference objects categorized under that type. An inference object should have the key "inference" with a string high-level inference of the author's preferences that spans across multiple papers and the key "atomic_inferences" with a list of strings specific to the paper with evidence that supports "inference". Each "inference" should be a single brief sentence with a hypothesis about the author's preferences in the form "Your papers..." that is more general and derived across multiple papers; it does not need an explanation. Each item in "atomic_inferences" should have three keys: 1) string "atomic_inference" with a single brief sentence conveying an explanation of how the paper relates to that inference in the form "One paper..." which will be more specific to the paper; 2) string "paper_title" which has the title of the paper from which the inference was derived; and 3) list of integer "paragraph_numbers" for the paragraphs (not sections) in "paper_title" from which the inference was derived. There should be five inference objects for each category, and each object should have a list of "atomic_inferences" that cite all of the author's papers that are relevant. No two atomic inferences under the same inference object should cite the same paper. Use each paper at least once. Do not cite section titles. </format instructions>
<personalization requirements>
- Inferences must be extremely personalized to the author; they should not apply to many authors in the field. For example, the inference "You know about issues in LLMs" under knowledge is too vague; you should point to a specific type of issue (e.g. memorization) that is more specific to just this author. The more specific the better.
- Do not make inferences that are conventions for certain fields in research papers.
* Limitation sections, listing contributions, using tables and figures, having related work, pointing out problems and proposing solutions, and using footnotes are all common conventions across conference papers for "writing style" and should thus NEVER be used as inferences. Instead, you should focus on aspects like tone, argumentation, and quirks that are specific to the author and would infrequently be found by other authors submitting similar papers.
* Similarly, when stating which audience an author writes for, do not stop at just the field or intersection of fields (e.g. computer vision for scientists). Drill down into this fields, like the type of computer vision researcher (e.g. image recognition) or type of scientist (e.g. doctor).
* Many researchers who work on the same topic often believe the same things that appear specific at first. For example, many researchers working on biases believe that biases are harmful. So instead of this, try to carve out a more unique preference for that author. For example, you might say that the researcher believes biases are harmful because they cause mistrust in high-stakes domains.
- Each high-level inference should not group multiple aspects together, and instead should focus on aspects that are most important for the researcher. For example, instead of saying the researcher works on training, evaluating, and deploying models, you could mention just one of the aspects the researcher works on that is unique compared to most researchers. </personalization requirements>
<inference requirements> - When generating a high-level inference, do not force atomic inferences from each of the papers. If a paper does not support a high-level inference, do not include it. For example, if you mention the researcher studies generalization, do not claim all of their papers relate to generalization if in reality they do not.
- When generating inferences, you should always briefly mention how this aspect of the researcher is different from most other researchers in their field if it is not obvious. If you say the author prefers X, explain that most other researchers prefer Y, which is different from X. For example, instead of just saying a researcher "prefers hyperparameter sweeps", you could mention "You prefer to test a larger space of hyperparameters compared to most researchers who typically test just a few ablations".
- Do not be excessively flattering. Use simple, clear, and easy-to-understand language. For example, instead of saying that the researcher "has a unique, nuanced skillset in machine learning and psychology" simply say they "are familiar with machine learning and psychology".
- You do not just have to say what the researcher wants to hear. Be honest. You are encouraged to include what the researcher does NOT know, does NOT do, or does NOT prefer. These can be seen as areas of growth.
- Your inferences should not imply that the author has done everything in those papers, as they may not have written them. For example, instead of saying "You use automated red-teaming" under research style if all papers discuss red-teaming, you should infer something like "Your papers use red-teaming". </inference requirements>

Prompt D.2: Profile Category Rubric

<Knowledge>
Definition: Inferences from the paper about what the researcher *knows or doesn't know*. This includes:
1. **Topic expertise or interest** - What domains the researcher seems fluent in or newly exploring.
2. **Familiarity with methods** - Implicit or explicit comfort with specific techniques or paradigms.
3. **Awareness of prior work** - The breadth and specificity of citations or conceptual framing.
Distinguishing Focus: What knowledge the researcher holds or lacks. </Knowledge>
<Research Style>
Definition: Inferences about *how the researcher prefers to conduct research*. This includes:
1. **Methodological preferences** - Use of qualitative, quantitative, computational, or hybrid approaches.
2. **Types of research questions** - Empirical, theoretical, exploratory, evaluative, etc.
3. **Study or experiment strategies** - How data is collected, analyzed, or operationalized.
Distinguishing Focus: How the researcher approaches doing research.
</Research Style>
<Writing Style>
Definition: Inferences about *how the researcher writes and explains ideas*. This includes:
1. **Argumentation and structure** - How ideas are developed, ordered, and emphasized.
2. **Tone and voice** - Formality, assertiveness, didacticism, etc.
3. **Explanation preferences** - Use of examples, metaphors, definitions, or technical language.
4. **Stylistic quirks** - Repetition, narrative devices, or particular rhetorical habits.
Distinguishing Focus: How the researcher communicates in writing.
</Writing Style>
<Positions>
Definition: Inferences about *what the researcher believes or argues*. This includes:
1. **Claims and conclusions** - What stances are taken or avoided.
2. **Normative views** - Ethical, political, or philosophical commitments evident in the writing.
3. **Arguments emphasized** - Which perspectives are advanced or critiqued.
Distinguishing Focus: What positions the researcher is taking or signaling.
</Positions>
<Audience>
Definition: Inferences about *who the researcher is writing for or trying to impact*. This includes:
1. **Assumed audience background** - What the researcher expects the reader to already know.
2. **Stakeholder relevance** - Who is likely to be affected by or benefit from the work.
3. **Audience alignment** - Whether the writing aligns with academic, practitioner, policy, or public communities.
Distinguishing Focus: Who the researcher is addressing or aiming to influence.
</Audience>

Prompt D.3: Action Generation Prompt (§2.2)

Here are a list of inferences about a user. The numbered inference is a high-level inference, while the sub bullet points provide evidence for these inferences: <profile> $\mathcal{P}$ </profile>
They are now asking: <query> q </query>
This query will eventually be fed into a system called PersonalizedQA that executes: 1. retrieval: searches for research papers
2. organization: outlines sections for the final response to include
3. generation: produces text for each of these sections

To help PersonalizedQA personalize responses based on the user's profile, come up with a list of personalization strategies that the system should follow. Each personalization strategy should specify two requirements:
1. What kind of response the user will experience (Qualitative Personalization)
2. How the system should behave at each step (Implementation Personalization)

The qualitative personalization label is based on how the response will be personalized to the user at a qualitative level: <qualitative personalization strategies> insert qualitative rubric </qualitative personalization strategies>

The implementation personalization label is based on the three-step execution of ScholarQA, categorized as: <implementation personalization strategies> insert implementation rubric </implementation personalization strategies>
<format instructions> Generate your output as a JSON object with 4 keys based on the implementation personalization strategies: "search_add", "search_refine", "organization", "generation"—where each key has a list of personalization strategy objects. Each strategy object should have four keys: 1) a string "strategy" as a brief high-level requirement for the final output that would lead to a more helpful response that is personalized to this user; 2) a string "tldr" with an extremely brief version of the "strategy" (less than fifteen words); 3) an integer "inference_number" which has the numbered inference from which the strategy was derived; and 4) a string "qualitative_strategy" categorizing how the output will be affected qualitatively (one of "content", "explanation_style", "specificity", "usefulness"). All requirements should be a concise sentence (<30 words) in the exact form "I can... [action to take], which might help you... [predicted help]". Include exactly four strategies for each implementation category (four personalization strategies for each of "search_add", "search_refine", "organization", "generation"). Make sure all of the strategies directly address the query. </format instructions>
<personalization instructions> - When designing a personalization strategy, do not just consider what the researcher knows or prefers, but also what the researcher does NOT know or does NOT prefer. For example, if a cybersecurity researcher asks for papers genetic sequencing, we likely need to add more background information for this user. This should involve adding a preliminary background section in "organization" or using simple terminology in "generation". On the converse, if the user is an expert in a topic, state that you will not add preliminary sections and avoid basic redefinitions to help save the user time.
- Do not force the personalization strategies to be specific. The specificity of each strategy should depend on how similar the query is to the user's profile. For example, if a user works on knowledge graphs and the query relates to knowledge graphs, the personalization strategies should be very specific based on the user's profile, outlining more concrete actions to take. However, if this same user with interests in knowledge graphs asks about computer vision, the actions to take in the personalization strategies should be more high-level.
- Do not always try to directly copy the user's profile when making requirements. For example, if a user's profile says they are interested in a specific psychological construct and you want to give a strategy involving this (e.g. I will connect the explanation to Ebbinghaus's learning curve), do not mention the specific construct. You should instead write more broadly (e.g. I will connect the explanations to memory constructs).
- If the query is very aligned with the user's profile, provide much more concrete suggestions for personalization. But if the query is quite dissimilar, keep the personalization suggestions very high-level and broad.
- Not every strategy should involve adding information. It is extremely important for you to also propose strategies so that the user can save time, like by ignoring papers in search_add, skipping sections in organization, and not redefining terms they already know in generation. Include at least one of these time-saving strategies, and add even more if the user is closely related to the information in the query.
- Ensure a considerable amount of the strategies (one third) involve giving suggestions for how the user could apply the information in the response for their own research—the qualitative strategy label "usefulness" </personalization instructions>
<action instructions> - The action X to take in each strategy should be specific to each category:
* search_add: I can also search for papers on X, I will add X to my list of search terms, I will expand the search to include X, etc.
* search_refine: I will interpret X in your query to mean Y, I will ignore papers that do X, I will narrow the domain/task to X, etc.
* organization: I can add/ignore a section on X, I can have a more/less detailed on section X, etc.
* generation: I can connect my explanation to concept X, I can add explain X by doing Y, I can use an X style, etc.
- If you are not confident the action is possible (e.g. if you do not know if there are papers that exist on a topic X in search_add or search_refine), use careful, hedged wording to avoid overclaiming, like "I can see if there are papers on X". Always hedge on search actions, but only when you are not fully confident on organization and generation.
- Please include several actions to take that involve saving time and generating shorter responses, like by ignoring papers in search_refine or search_add, skipping sections in organization, and not redefining terms they already know in generation.
</action instructions>

...[truncated]...

Prompt D.4: Action Qualitative Categories

<Content>
Definition: Specifies *what information* the response should include and how it should be conceptually framed. This can include:
1. **Conceptual scope** - Which concepts to emphasize, omit, or define.
2. **Depth of explanation** - Whether to provide brief overviews (if the user is knowledgeable on the area) or in-depth knowledge (if the user is new to the field).
3. **Terminology alignment** - Tailoring vocab to match the user's disciplinary conventions.
Distinguishing Focus: What content is covered and how deeply.
</Content>
<Explanation Style>
Definition: Specifies *how the explanation for the information is communicated*. This can include:
1. **Explanatory style** - Empirical, intuitive, formal/mathematical, or example-led.
2. **Cognitive structuring** - Layered explanations, definitions first vs. bottom-up learning.
3. **Framing mechanisms** - Use of analogies, metaphors, or domain-specific language aligned with the researcher's background. **Distinguishing Focus:** How the content is explained, formatted, and connected to other concepts.
</Explanation Style> <Specificity> **Definition:** Clarifies and narrows the scope of the response to better match the researcher's intended focus. This can include:
1. **Disambiguating vague inputs** - Interpreting terms like "methods, "frameworks", or "best" in the user's specific context.
2. **Focusing by domain/task** - Aligning content to a subfield, methodology, or research phase.
3. **Resolving underspecification** - Filling in implicit assumptions (e.g., assuming qualitative when not stated).
4. **Removing irrelevant scope** - Avoiding generalizations or adjacent topics not central to the task.
Distinguishing Focus: What exactly is meant or needed, and how to restrict the response to that.
</Specificity> <Usefulness> **Definition:** Shapes the response to be *actionable* or *instrumental* for the researcher's goals or workflow. This can include:
1. **Direct application** - Helping write a section, implement a method, interpret results, etc.
2. **Workflow integration** - Mapping content to stages of research or types of output.
3. **Next steps** - Suggesting what to do with the information (e.g., adapt, cite, reframe, test).
4. **Decision support** - Helping choose between options, methods, or framings based on task-fit.
Distinguishing Focus: How the information can be turned into research actions or outputs.
</Usefulness>

Prompt D.5: Action Implementation Categories

<Search Add>
Definition: Personalizes the search by **adding new terms or dimensions** to the original query. This includes:
- Introducing related subfields, topics, or concepts the user may not have explicitly mentioned
- Incorporating inferred preferences such as favored methods, datasets, evaluations, or research types
- Expanding the query scope to make responses more relevant or actionable in the user's workflow
- Suggesting new combinations of fields or terms to enhance discovery
</Search Add>
<Search Refine>
Definition: Personalizes the search by **revising or improving** the original query. This includes:
- Disambiguating unclear or subjective terms
- Making existing search terms more specific or technically precise
- Narrowing or clarifying the focus based on known user preferences or context
- Adjusting query language to better match terminology used in the literature
- Removing irrelevant, redundant, or low-value terms that may dilute the quality of results
</Search Refine>
<Organization>
Definition: Personalizes how the papers are grouped into sections for the final response. This includes:
- Which sections should be included or excluded (e.g. skipping intro sections if the user is an expert in the field, adding background sections if the user is a novice)
- High-level structure (e.g. organizing by themes, topic, methods, history, research questions, etc.)
- Additional sections to make the response more useful in the user's research workflow (e.g. a section on follow-up ideas, implementation steps, considerations, suggestions, etc.). This should be heavily prioritized.
</Organization>
<Generation>
Definition: Personalizes how certain sections are written and explained. This includes:
- What connections the responses should make (e.g. connections to the user's prior frameworks, methods, papers, etc.)
- The strategy of explanations (e.g. definitions-first, example-first, intuitive-first, etc.)
- The writing style and level of elaboration based on the user's expertise (e.g. brief overviews versus deep exposition)
</Generation>

Prompt D.6: Profile Inference Accuracy Prompt (§2.1)

<task> As an Attribution Validator, your task is to verify whether a given inference can be accurately derived from a list of references. A reference is a collection of snippets from a research paper. Specifically, your response should clearly indicate the relationship: Attributable or Contradictory. A contradictory error occurs when you can infer that the inference contradicts the information presented in the reference. If the inference appears true based on the papers, even if some of the papers are irrelevant (i.e. the model "over-cited" the papers), then the inference is Attributable. </task>

<inference> $\mathcal{I}$ </inference>

<references> $\mathcal{D}_{\text{cite}}$ </references>

<format> Output your response as a json with only a single key "output" and a value of one among - ("Attributable", "Contradictory"). </format>

Prompt D.7: Profile Inference Category Accuracy Prompt (§2.1)

<task> As a Category Validator, your task is to verify whether a given inference can be classified under the specified category. Specifically, your response should clearly indicate the relationship between the inference and category: Match or Mismatch. A mismatch occurs when you can infer that the inference does not relate at all to the category and its definition. A match occurs when you can infer they do relate to each other </task>

<category> [insert category] </category>

<category definition> [insert definition] </category definition>

<inference> $\mathcal{I}$ </inference>

<format> Output your response as a json with only a single key "output" and a value of one among - ("Match", "Mismatch"). </format>

Prompt D.8: Profile Inference Relevance Prompt (§2.1)

<task> As a Relevance Validator, your task is to determine whether a specific text from a paper provides support for and is relevant to a broader inference intended to span multiple papers. If the paper text provides support for at least one aspect of the inference, then it is relevant. If the paper text supports no part of the inference, then it is irrelevant. For example, if the inference claims "Your papers use the terms 'first' and 'novel'" and from the text we can infer that "The paper uses the term 'first'", the paper text is relevant since it relates to the claim about 'first', even though the word 'novel' is not discussed. Thus, for the paper text to be "Relevant", it only needs to support one aspect of the inference. </task>

Here is the paper text: <paper text> d_{cite} </paper text>

And here is the inference: <inference> $\mathcal{I}$ </inference>

<format> Output your response as a json with only a single key "output" and a value of one among - ("Relevant", "Irrelevant"). </format>

Prompt D.9: Profile Inference Specificity (§2.1)

<task> As a Specificity Validator, your task is to rate the specificity of a given inference about a computer science researcher from one to five. </task>

Use the following criteria:
<criteria>
Criteria: Personalization: How specifically tailored and insightful is the inference about the computer science researcher?
Score 1: Extremely vague or generic; the inference could apply to almost any researcher in computer science.
Score 2: Broad and minimally tailored; captures a common area or trait that applies to many researchers in computer science.
Score 3: Moderately specific; identifies a more refined topic or pattern but still describes a large population of computer science researchers.
Score 4: Specific and reasonably personalized; reflects a more distinctive sub-area, approach, or motivation of the researcher.
Score 5: Highly specific and personalized; demonstrates a deep, nuanced inference that could plausibly distinguish this researcher from almost every other researcher in their field.
</criteria>

Here is the inference you must rate: <inference> $\mathcal{I}$ </inference>

<format> Output your response as a json with only a single key "output" and an integer rating for Specificity from one to five. </format>

Prompt D.10: Action Personalization Win Rate Prompt (§2.2)

As a Plan Validator, your task is to determine which of two plans for how to tailor a response best matches a user's profile. The user profile will be a series of inferences about a user derived from their research papers, organized under various categories. The two plans will be labled as "Plan A" or "Plan B" and describe a list of suggestions an external question answering model could execute to generate a more personalized response. Your output should denote whether plan "A" or "B" is better aligned with suggestions that the user described in the profile would prefer.

Here is the profile: <profile> $\mathcal{P}$ </profile>

Here is Plan A: <plan A> p_{person} </plan A>

Here is Plan B: <plan B> p_{gen} </plan B>

<format> Output your response as a json with only a single key "output" and a value of one among - ("A", "B"). </format>

Prompt D.11: Action Relevance Prompt (§2.2)

As a Plan Contradiction Validator, your task is to determine if a plan step directly conflicts the instructions in the query. The query will be a question related to scientific research. The plan will describe a list of suggestions an external question answering model could execute to generate a better response.

Your output should denote whether the plan has a "CONFLICT" or "NO_CONFLICT" with the query. For example, if the query asks "What are the best question answering datasets?" and a plan step says "Focus search on summarization benchmarks", there would be a "CONFLICT", since the model cannot focus on summarization benchmarks without ignoring question answering datasets, and thus would have to ignore the query to follow the instruction. However, if a plan step for this query said "Focus on Extractive Question Answering" it would be "NO_CONFLICT", since the model could follow this step while still answering the query. Similarly, if the plan step said "Draw connections to summarization benchmarks" it would be "NO_CONFLICT", as drawing a connection does not mean ignoring the request in the query.
</task>

<query> q </query>

<plan step> a </plan step>

<format> Output your response as a json with two keys: 1) "output" with a value of one among - ("CONFLICT", "NO_CONFLICT"); and 2) "explanation" with a brief explanation as to why. </format>

Prompt D.12: ACtion Instruction-Following Prompt (§2.3)

You are given a query, an instruction, and a corresponding long answer. As an Instruction-Following Validator, your task is to determine whether the answer correctly follows a given instruction with the boolean flag "was_followed".

If the response directly follows the instruction, then "was_followed" is true. If the response does not directly adhere to the instruction, either failing to fulfill any of its requirements or failing to acknowledging it, then "was_followed" is False

For example, if the instruction states that the response should "Include a section on metrics" and the response has a section on metrics, then "was_followed" is true

Similarly, if the instruction states that the response should "Discuss future directions" and the response only reports on current trends, then "was_followed" is false

You should be strict with your judgments. If the instruction says the model should do something (e.g. add a section titled "X", add an analogy on "Y"), the model must follow it exactly for 'was_followed' to be true. If the model vaguely follows the instruction (e.g. adding sections related to "X" but not with the right title, using keywords linked to "Y" but not adding an analogy), 'was_followed' should be false

Return your result as a JSON object with the keys: 1) 'was_followed': a true/false boolean for whether the instruction was followed in the answer; and 2) 'reason': a string explanation behind your decision:
{{
"was_followed": boolean true/false for if the instruction is followed in the answer
"reason": string explanation for your decision in "was_followed"
}}

Question: q
Instruction: a
Answer: $\mathcal{R}$

Prompt D.13: Profile Satisfaction Prompt (§4.3)

You are an expert at evaluating generated text with respect to user satisfaction across specific metrics.

<task>
Given a set of research papers selected by a user, a model must generate a profile containing a series of inferences about the user, each of which cite the papers from which the inferences were derived. These inferences are supposed to capture information about the user that would help a question-answering system personalize its responses when the user asks questions. You will be given one of the model-generated profile inferences that the user reviewed, and will be asked to predict if the user was satisfied with the profile inference.

Here are the research papers the user selected to represent their profile:
<papers>
$\mathcal{P}$
</papers>

Here is an inference the model generated about the user that you must evaluate:
<profile inference> $\mathcal{I}$ </profile inference>
Here is the categorization of the above profile inference:
<profile inference category> [insert category] </profile inference category>
Your job is to evaluate if the user would be satisfied or dissatisfied with this inference in the profile. Satisfied means that the user believes the inference perfectly captures one part of their preferences and interests. If the user is satisfied with the inference they would leave it unaltered in their profile, with no desire for modifications or noting any issues (no matter how minor).

Specifically, evaluate the response for user satisfaction with the following criteria in mind:
<metric>
Metric criteria: Would the user be satisfied with how broadly the profile inference claims apply across their papers and in particular, the papers cited in the inference? Or is the profile inference overstating its scope?
- Set is_satisfied=true if the profile inference describes something that genuinely applies to a substantial portion of the user's papers, making it a meaningful part of their profile.
- Set is_satisfied=false if the profile inference is overstated, claiming to apply across the user's papers when in fact it only applies to a small subset or is not significant enough to represent the user's overall work.
</metric>
</task>

<format> Structure your output as a JSON with a boolean key "is_satisfied", which is set to true if the user would be fully satisfied and false otherwise, and "explanation", which provides a brief rationale as to why you picked the label in "is_satisfied".
</format>

Model	Answer Coverage	Answer Precision	Citation Precision	Citation Recall	Action Adherence
MySCHOLARQA	**91.4**	89.9	**91.8**	**81.4**	83.2
SCHOLARQA	88.9	89.1	90.5	76.9	81.3
OPENSCHOLAR	77.2	**97.4**	82.5	60.4	82.5
STORM	72.0	92.2	73.3	64.7	74.4
Perplexity Sonar DR	81.0	82.9	64.3	46.3	75.0
OpenAI DR (o3)	89.1	90.2	79.2	56.7	**93.8**

Table 10: Deep Research report quality and action instruction-following for query q and 8 actions in $\mathcal{A}_{\text{gen}} \cup \mathcal{A}_{\text{person}}$. MySQA **surpasses** all DR tools in 3/5 metrics. We run OpenAI DR on 10 examples due to latency and cost issues.

Prompt D.14: Action Satisfaction Prompt (§4.3)

You are an expert at evaluating generated text with respect to user satisfaction across specific metrics.
<task> Given a query asked by a user and a profile that captures that same user's preferences and interests, a model must generate suggested actions (which we refer to as plan steps) that a system could also perform when answering the question. The plan steps, when followed, are supposed to result in more useful information for the user in the final response. The usefulness of a response depends on the user's intent in the query, but is likely intended to help them learn new information, find relevant papers they can save, propose new ideas for them to explore, or give implementation advice. You will be given one of the model-generated plan steps that the user reviewed, and will be asked to predict if the user was satisfied with the plan step and wanted the model to execute it when answering the query.

Here is the query the user provided:
<query> q </query>

Here is the user's profile:
<profile> $\mathcal{P}$ </profile>

Here is one of the plan steps the model generated:
<plan step> a </plan step>

Here is the categorization of the above plan step:
<plan step category> [insert category] </plan step category>

Your job is to evaluate if the user would be satisfied or dissatisfied with the plan step that the model proposed. If the user is satisfied with the plan step, they would want a model to follow this extra request in addition to answering their query, with no desire for modifications or noting any issues (no matter how minor).

Specifically, evaluate the response for user satisfaction with the following criteria in mind:
<metric>
Metric criteria: Given their original query, would the user be satisfied with the information that this plan step would incorporate in the answer to the query? Or would this add information that is overly distracting?
- Set is_satisfied=true if the plan step stays aligned with the user's query and directs the response toward information that would be clearly useful for addressing their request.
- Set is_satisfied=false if the plan step shifts the focus away from the query, leading the response toward content that is irrelevant or distracting from what the user actually wants to know.
</metric>
</task>

<format> Structure your output as a JSON with a boolean key "is_satisfied", which is set to true if the user would be fully satisfied and false otherwise, and "explanation", which provides a brief rationale as to why you picked the label in "is_satisfied".
</format>

Prompt D.15: Report Satisfaction Prompt (§4.3)

You are an expert at evaluating generated text with respect to user satisfaction across specific metrics.
<task> Given a query asked by a user and a plan step containing additional instructions for the model to perform when answering the query, a model must generate a multi-section response that answers the query and follows the extra steps. The response is supposed to contain information related to the plan step that the user would find useful in the entire response, but particularly in the spans of highlighted text. The usefulness of a response depends on the user's intent in the query, but is likely intended to help them learn new information, find relevant papers they can save, propose new ideas for them to explore, or give implementation advice. You will be given one of the model-generated responses and the plan step that the user reviewed, and will be asked to predict if the user was satisfied with how the plan step was followed in the response.

Here is the query the user provided:
<query> q </query>
Here is the plan step the user asked the model to follow:
<plan step> a </plan step>

Here is the categorization of the above plan step:
<plan step category> [insert category] </plan step category>

Here is the response the model generated:
<response> $\mathcal{R}$ </response>

Your job is to evaluate if the user would be satisfied or dissatisfied with how the model followed the plan step in its response. If the user is satisfied with how a plan step was followed in the response, they would find that the information related to the plan step in the response is perfectly described and useful, with no desire for modifications or noting any issues (no matter how minor).

Specifically, evaluate the response for user satisfaction with the following criteria in mind:
<metric>
Metric criteria: Would the user be satisfied with the depth of information in this response related to the plan step? Or is the response content related to the plan step too vague, high-level, or general to be useful?
- Set is_satisfied=true if the response content related to the plan step provides concrete, detailed, and specific information tied to the plan step that adds meaningful value for the user.
- Set is_satisfied=false if the response content related to the plan step is vague, superficial, or generic, giving little more than high-level statements without useful depth or detail.
</metric>
</task>

<format> Structure your output as a JSON with a boolean key "is_satisfied", which is set to true if the user would be fully satisfied and false otherwise, and "explanation", which provides a brief rationale as to why you picked the label in "is_satisfied".
</format>