Author-in-the-Loop Response Generation and Evaluation: Integrating Author Expertise and Intent in Responses to Peer Review

Anonymous ACL submission

Abstract

Author response (rebuttal) writing is a critical stage of scientific peer review that demands substantial author effort. Recent work frames this task as automatic text generation, underusing author expertise and intent. In practice, authors possess domain expertise, author-only information, revision and response strategies–concrete forms of author expertise and intent–to address reviewer concerns, and seek NLP assistance that integrates these signals to support effective response writing in peer review. We reformulate author response generation as an author-in-the-loop task and introduce *REspGen*, a generation framework that integrates explicit author input, multi-attribute control, and evaluation-guided refinement, together with *REspEval*, a comprehensive evaluation suite with 20+ metrics covering input utilization, controllability, response quality, and discourse. To support this formulation, we construct *Re*[3]*Align*, the first large-scale dataset of aligned review–response–revision triplets, where revisions provide signals of author expertise and intent. Experiments with state-of-the-art LLMs show the benefits of author input and evaluation-guided refinement, the impact of input design on response quality, and trade-offs between controllability and quality. We make our dataset, generation and evaluation tools publicly available.[1]

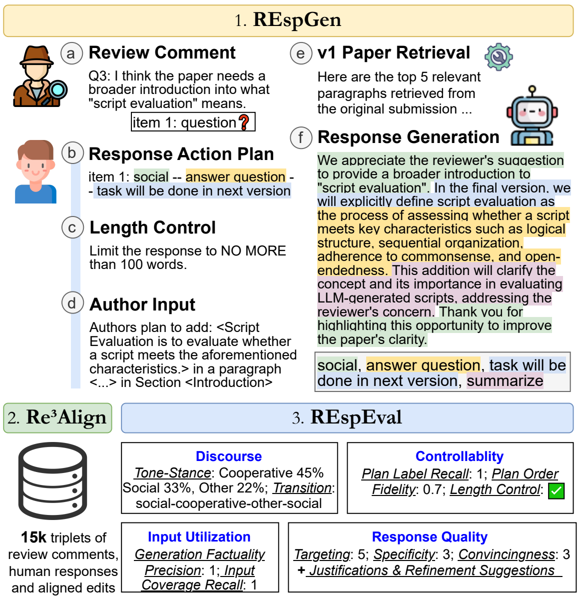


Figure 1: In this work, we contribute (1) *REspGen*, an author-in-the-loop ARG framework that integrates explicit author input (d), controllable planning and length (b–c), and additional paper context (e); (2) *Re*[3]*Align*, the first large-scale review–response–revision triplets dataset for modeling author signals; and (3) *REspEval*, a comprehensive response evaluation framework with over 20 metrics spanning four dimensions.

1 Introduction

Author response (rebuttal) writing is a critical stage of scientific peer review, where authors address reviewer concerns through clarifications, supporting evidence, and proposed revisions. Because acceptance decisions are at stake, this process requires substantial author effort, making response writing a promising yet challenging NLP task (Kuznetsov et al., 2024). However, existing work on **author response generation (ARG)** typically relies solely on reviewer comments, resulting in generic responses that fail to reflect concrete author expertise and intent. In practice, many reviewer concerns can only be adequately addressed using information known exclusively to the authors, such as planned clarifications (e.g., the precise definition of *script evaluation* in Figure 1), newly conducted experiments, or design rationales. Additionally, authors strategically choose response approaches (revising, justifying, deferring, or promising future work) while controlling response attributes such as length, tone, and discourse structure. Empirical studies of successful rebuttals further show that effective responses are improvement-oriented and highly specific, often citing concrete revisions, exact loca-

[1]Links removed for anonymity; added upon acceptance.

tions, or quantitative evidence (Noble, 2017; Gao et al., 2019; Huang et al., 2023). These observations underscore the importance of incorporating author expertise and intent in ARG, motivating an **author-in-the-loop ARG** framework that better aligns generated responses with author intent while synergizing human expertise with NLP assistance.

However, realizing this framework presents two key challenges: (i) **Data scarcity.** Author-in-the-loop ARG requires author expertise and intent (hereafter *author signals*) aligned with specific reviewer concerns. In conference settings, responses precede revision but often describe planned changes later appearing in revised papers, enabling post-hoc edit extraction to approximate author signals at response time. Journal responses instead document already-implemented revisions. Modeling either setting requires *complete paper records*—reviews, original submissions, revised papers, and authentic responses—available in only a few datasets (Dycke et al., 2023; Lin et al., 2023). More critically, existing resources lack fine-grained annotations needed to model author signals at the granularity of individual reviewer concerns, such as detailed edit analyses, review-response segment alignments, and mappings to concrete paper edits. (ii) **Response evaluation**. Prior ARG work relies primarily on surface similarity to human responses (Purkayastha et al., 2023; Tan et al., 2024), with only Zhang et al. (2025) using LLM-based evaluation on coarse quality criteria. Moreover, controllable ARG remains underexplored, and even in controllable generation broadly, rigorous evaluation of controllability and its trade-offs with generation quality remains an open challenge (Zhang et al., 2023; Liang et al., 2024).

To address the data challenge, we introduce ***Re^{3}Align*** (§3), the first large-scale dataset for author-in-the-loop ARG, comprising 3.4k complete paper records with 440k sentence-level edit annotations and 15k aligned review–response–edit triplets. We further propose ***REspGen*** (§4), an author-in-the-loop ARG framework supporting multiple forms of author input and enabling controllable generation over response planning and length, with modular components and iterative refinement guided by ***REspEval*** (§5). REspEval moves beyond similarity-based evaluation, providing over 20 novel metrics for controllability, input utilization (including factuality and coverage), response quality including targeting, specificity, and convincingness, and discourse characteristics such as tone–stance profiles and transitions. Finally, we experiment with five state-of-the-art (SOTA) LLMs across nine settings to systematically analyze effects of author signals, input design, attribute control, and evaluation-guided refinement (§6). Our work makes four key **contributions**:

- A first large dataset of review–response–edit triplets with rich annotations, enabling a new formulation of the ARG task;
- An author-in-the-loop ARG framework supporting flexible author input, multi-attribute control, and evaluation-guided refinement;
- A comprehensive evaluation framework with 20+ novel metrics for controllability, input utilization, response quality, and discourse;
- Extensive experiments across five LLMs and nine settings, yielding insights into ARG behavior under varied inputs and controls.

This work establishes a new paradigm of author-in-the-loop response generation and evaluation, bridging author expertise and intent with NLP assistance to support effective and efficient response writing while preserving the essential role of human engagement in scholarly communication.

2 Related Work

Author Response Generation has recently emerged as a challenging and underexplored task in NLP for scientific peer review (Kuznetsov et al., 2024; Staudinger et al., 2024). Early work on author response includes argument-pair extraction (Cheng et al., 2020) and response discourse analysis (Kennard et al., 2022). Empirical studies further identify key success factors for effective responses, including explicit revision statements, high specificity, concrete evidence, comprehensive coverage of reviewer concerns, and appropriate tone (Noble, 2017; Gao et al., 2019; Huang et al., 2023). Recent work shifts toward generation, with ARG studies focusing on attitude- and theme-guided generation (Purkayastha et al., 2023) and multi-turn review–rebuttal dialogue (Tan et al., 2024; Zhang et al., 2025). However, these generation approaches rely solely on reviewer comments, producing generic responses that lack concrete details, especially those requiring author expertise. Evaluation is limited to similarity metrics, overlooking response diversity and broader success factors from empirical work. As summarized in Table 1, we address these limitations in three ways by (i) introducing the first large-scale triplet dataset of reviews,

	Data				Generation		Evaluation	
	review	response	revision annotation	triplet alignment	author input	author control	dimension	metric
Jiu-Jitsu (2023)	✓seg	✓seg	✗	✗	✗	✗	Similarity	ROUGE, BERTScore
ReviewMT (2024)	✓doc	✓doc	✗	✗	✗	✗	Similarity	ROUGE, BLEU, METEOR
Re^2 (2025)	✓doc	✓doc	✗	✗	✗	✗	Similarity, Quality	ROUGE, BLEU, BERTScore
Ours	✓seg	✓seg	✓sent	✓seg	✓	✓	Similarity, Quality, Discourse, Input Utilization, Controllability	20+ novel metrics

Table 1: Comparison of related works on author response generation, including data, generation task formulations and evaluation dimensions and metrics. doc/seg/sent: document-/segment-/sentence-level alignments and annotations.

	#Paper	#Pair	#Edit	#Linked Edit	#Re^3 Triplet
EMNLP24	679	2,108	86,247	16,762	1,933
PeerJ	2,715	13,963	353,551	181,534	13,588
Total	3,394	16,071	439,798	198,296	15,521

Table 2: Re^{3}Align Dataset Statistics. Reported are the counts of papers, aligned review–response pairs, annotated sentence-level edits, edits linked to the pairs, and the final number of aligned triplets.

responses, and aligned sentence-level edits, treating revisions as explicit signals of author expertise and intent; (ii) formulating ARG as an author-in-the-loop task that integrates author expertise and intent through explicit input and controllable generation; (iii) proposing a comprehensive evaluation suite with 20+ metrics spanning four dimensions beyond similarity-based evaluation.

Controllable Text Generation and Evaluation aims to steer model outputs toward user-specified constraints (Zhang et al., 2023). Prior work primarily focuses on single-attribute control, including length (Kikuchi et al., 2016), topic (Wang et al., 2019), and sentiment (Firdaus et al., 2020), as well as content-based control such as query-focused (Xu and Lapata, 2021), entity-centric (Maddela et al., 2022), and aspect-based generation (Li et al., 2023). Recent surveys highlight persistent challenges in simultaneous multi-attribute control, trade-offs between controllability and quality, and the lack of rigorous evaluation methods (Zhang et al., 2023; Liang et al., 2024). In ARG, controllability remains unexplored despite authors' need to strategically control response construction while integrating their own content. We provide the first study of controllability in ARG, examining control over length, discourse planning, and content integration. We further introduce a comprehensive evaluation framework with novel fine-grained metrics assessing: (i) how well generations adhere to single- and simultaneous multi-attribute controls; (ii) how effectively author-provided content is incorporated; and (iii) how response quality is impacted.

3 Dataset Construction: *Re^{3}Align*

3.1 Data Collection and Preprocessing

Our framing of ARG requires raw data capturing the full review–revision–response (Re^3) process with authentic human texts. Only a few resources, such as the EMNLP24 subset of NLPEERv2 (Dycke et al., 2023) and MOPRD (Lin et al., 2023), provide peer reviews, author responses, original submissions, and revised papers. EMNLP24 provides peer reviews and rebuttal discussions from OpenReview,[2] which we organize into reviewer–author discussion chains, extracting and merging consecutive author replies into single responses. MOPRD offers data from PeerJ[3] across multiple scientific domains including computer science, chemistry, physics, and materials science. We retain only papers with a complete Re^3 record. The final corpus includes 679 EMNLP24 papers and 2,715 PeerJ papers (Table 2), covering both conference and journal workflows. We group each paper's versions, reviews, and responses under a unified identifier and convert them into intertextual graphs (ITGs) (Kuznetsov et al., 2022), augmented with sentence-level nodes (details in §A.1).

3.2 Review-Response Pair Alignment and Revision Annotation

Authors often quote review sentences to structure their replies. To extract review–response pairs, we match every review sentence to every response sentence using an assembled matching algorithm (§A.2) and merge the longest contiguous matches to identify quoted review spans. These spans are then used to segment the response, with each segment defined as the text following a quoted span and preceding the next one. An illustrative example is shown in Figure 4 in §A.2. After applying quality filtering strategies (§A.2), we obtain 2,108 and 13,963 review–response segment pairs from

[2] https://openreview.net/
[3] https://peerj.com/

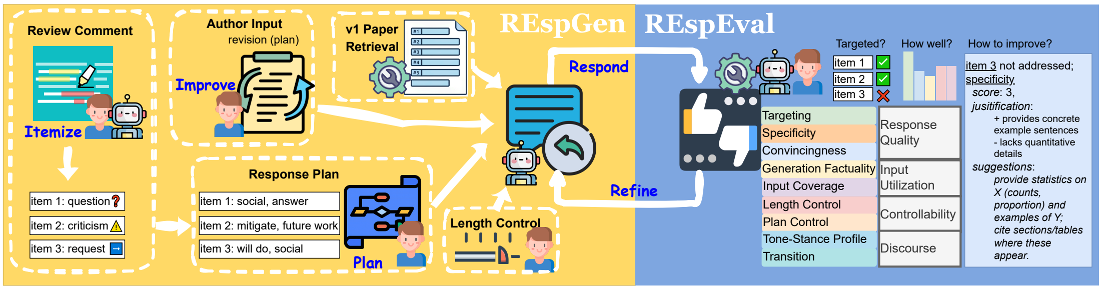


Figure 2: Frameworks: *REspGen* & *REspEval*.

EMNLP24 and PeerJ, respectively. Human verification of 100 pairs confirms a 98% alignment accuracy. We further apply SOTA revision analysis models (Ruan et al., 2024a,b) to align sentence-level edits across paper versions and label each with edit action and intent. These models achieve over 90 F1 for alignment and action labeling, and 84.3 F1/85.6% accuracy for edit intent classification. In total, this produces 439,798 edits.

3.3 Re^3 Triplet Alignment

For each submission, we have the original paper D^t, the revised paper D^{t+1}, and reviewer–author exchanges (C_k, A_k) for reviewer k. Sentences in D^t and D^{t+1} are denoted x_j^t and x_i^{t+1}. From earlier steps, we extract sentence-level edits $e_{ij} = e(x_i^{t+1}, x_j^t)$, with the full edit set denoted as E. We also obtain aligned review–response segments $p_{mn}^k = p(c_m^k, a_n^k)$ with $c_m^k \in C_k$ and $a_n^k \in A_k$, denoted collectively as P.

The task of triplet alignment is to determine whether an edit $e_{ij} \in E$ is relevant to a pair $p_{mn}^k \in P$. For each $p_{mn}^k = p(c_m^k, a_n^k)$, we use a two-way strategy: (i) align the review comment c_m^k to each edit $e(x_i^{t+1}, x_j^t)$ using a function set CE, and (ii) align the response segment a_n^k to the same edit using a function set AE. Each function set combines a fine-tuned SOTA LLM classifier (>90% accuracy) with a lightweight similarity component (details in §A.3). All positive alignments are aggregated as the edits linked to p_{mn}^k, denoted $[e_{\text{align}}]$, yielding triplets $t_{mn}^k = (c_m^k, a_n^k, [e_{\text{align}}])$. We obtain 15,521 triplets with non-empty aligned edits. Human evaluation of 125 aligned edits yields precision of 0.86 for EMNLP 2024 and 0.71 for PeerJ, with perfect recall in both cases. Errors primarily arise from aggregation, which may align lexically or semantically similar texts whose edits are not directly relevant to reviewer concerns. We adopt aggregation to prioritize coverage and minimize missed alignments.

4 Generation Framework: *REspGen*

REspGen is a modular framework (Figure 2, left) with response plan and length control (§4.1), configurable author input and paper context (§4.2), and evaluation-guided refinement (§4.3).

4.1 Response Attribute Control

Item-Based Response Planning. We adopt the *review action* taxonomy of Kennard et al. (2022) and derive three review item types[4], including *Criticism*, *Question*, and *Request*, to classify spans within each review segment (definitions in Table 5). To simulate realistic author planning, we prompt GPT-5 to jointly analyze each review–response pair, itemize the review, align spans from the human response, and assign *response action* labels (Kennard et al., 2022). Table 6 lists the 16 labels grouped into five stance classes: *Cooperative*, *Defensive*, *Hedge*, *Social*, and *Other*. Illustrative examples appear in Figure 1 and Figure 6, with the prompt in Figure 5 (§B.1). In *REspGen*, authors may specify a response plan for each review item by providing a sequence of *response action* labels (see (b) of Figure 1), which guides the tone, stance, and discourse flow of the generated response. In experiments, we simulate author control using the annotations above. Further details are provided in §B.1.

Length-Constrained Generation. Many peer-review venues impose strict length limits on author responses to encourage focused communication.[5] In *REspGen*, authors may specify an upper-bound word limit for generation. Since appropriate length depends on the complexity of the review concern, in our experiments we simulate realistic author-provided limits by setting them to n+50 where n is the human response length (Figure 1(c)).

[4]We retain types that Kennard et al. (2022) found to be commonly addressed in responses. Other types (e.g., strengths, summaries) typically require no response.

[5]For example, https://docs.openreview.net/reference/default-forms/default-rebuttal-form

4.2 Input Component Configuration

REspGen supports configurable input components, including flexible author input and optional paper context. We simulate author expertise and intent through aligned sentence-level edits $[e_{\text{align}}]$, where each edit can be supplied as either (i) an edited-sentence string (***S***) or (ii) the string with its paragraph ***context*** and section title, reflecting where the edit appears (Figure 1(d)). Beyond author input, *REspGen* supports an optional ***v1*** retrieval module that retrieves the top five relevant paragraphs from the original submission using a retrieval–reranking approach conditioned on the review segment (details in §B.2). This paper-level context provides additional topic grounding for response generation.

4.3 Evaluation-guided Refinement

REspGen includes an iterative refinement module that interfaces with the evaluation framework *REspEval*. Given a review segment, optional author input and *v1* retrieval, response plan and length control, the system first generates an initial draft. *REspEval* then evaluates this draft and returns evaluation metrics, justifications and refinement suggestions (§5). These results, together with the original inputs, controls, and initial draft, are fed back into *REspGen* to produce a refined response. This iterative process leverages *REspEval* feedback to progressively improve responses, helping them better reflect author intent, satisfy controls, and address reviewer concerns.

5 Evaluation Framework: *REspEval*

REspEval evaluates four dimensions: discourse (§5.1), controllability (§5.2), input utilization (§5.3), and quality (§5.4), with subdimensions and metrics color-coded for quick reference below.

5.1 Response Discourse Analysis

Following §4.1, we label response spans with actions, yielding two analyses: (i) tone–stance profile, obtained by mapping actions to the five stance classes and computing their word-weighted proportions *%Coop*, *%Defe*, *%Hed*, *%Soc*, *%Other* and *ArgLoad=%Coop+%Defe+%Hed*, which reflects overall argumentative load; and (ii) transition flow, capturing stance distributions across response positions and shifts between adjacent spans. These analyses characterize communicative attitude and discourse dynamics, enabling comparison of human and LLM responses.

5.2 Controllability Evaluation

For length control (lenC), we compute the difference between the upper bound limit and the generated length for each sample, where positive values indicate adherence to the constraint. As aggregate metrics, we report the percentage of generations that meet the constraint (*%met*) and the median length difference across all samples (*m.diff*). For response plan control (planC), we assess how closely generated response action labels and their ordering match the plan, reporting label precision (*P*), recall (*R*), and *F1*. To evaluate ordering, we compute order fidelity (*OF*), measuring how well correctly-produced actions preserve plan order. Let $\mathbf{m} = (m_1, \ldots, m_T)$ denote the indices of plan actions matched to the generated response in generation order, with $m_i = -1$ indicating no match. We then define $\mathbf{s} = \{\, m_i \mid m_i \geq 0 \,\}$ as the matched plan indices in generated order, and let $\mathbf{s}^*$ be the same elements sorted in ascending order (i.e., the plan order). OF is the longest common subsequence (LCS) between $\mathbf{s}$ and $\mathbf{s}^*$, normalized by $|\mathbf{s}|$:

$$\mathrm{OF}(\mathbf{m}) = \begin{cases} 0, & |\mathbf{s}| = 0, \\ \dfrac{\mathrm{LCS}(\mathbf{s}, \mathbf{s}^*)}{|\mathbf{s}|}, & \text{otherwise.} \end{cases}$$

5.3 Input Utilization Measures

We assess how generated responses use given inputs through two fact-based measures inspired by atomic fact-checking (Min et al., 2023), which decomposes text into atomic facts and verifies each against reference sources. We adapt their best GPT-based approach[6] and introduce: (i) Generation Factuality Precision (GFP): extract atomic facts from the generated response and verify each against all given inputs (edit strings, optional paragraph context, and v1 content). GFP is the proportion of generated facts supported by the inputs, indicating factual grounding. (ii) Input Coverage Recall (ICR): decompose the core author input (edit strings) into atomic facts and check whether each is expressed in the generated response. ICR measures how well the model prioritizes and incorporates the author's intended improvements. For both measures, we report the proportions of supported (*%sup*), unsupported (*%unsup*) and contradicted (*%con*) facts.

[6] Min et al. (2023) use GPT-3.5, the most advanced available at the time; we use GPT-5 with redesigned prompts and examples for improved scientific fact extraction.

5.4 Response Quality Evaluation

We assess response quality using three core criteria grounded in venue guidelines,[7] expert advice,[8] and empirical studies (Gao et al., 2019; Huang et al., 2023): targeting (directly addressing reviewer concerns), specificity (providing concrete evidence and details), and convincingness (presenting clear, persuasive justification), which emphasize substantive effectiveness beyond surface-level fluency.

We evaluate these three dimensions using GPT-5 as a judge with structured, rubric-grounded reasoning, an approach shown to improve interpretability, reliability, and alignment with human judgments (Li et al., 2024). Given the review, response, and review item–response alignments, GPT-5 assigns 5-point scores for targeting (*Targ*), specificity (*Spec*), and convincingness (*Conv*). For each dimension, it provides evidence-based justifications by listing strengths and weaknesses with concrete references to relevant items, as well as refinement suggestions. Prompts with scoring rubrics, example outputs, and additional details are provided in §C.1.

We validate our approach through comprehensive studies (§C.2), confirming assessments are (i) **consistent** across runs, (ii) **robust** to perturbations, distinguishing genuine from degraded responses, (iii) **interpretable**, and (iv) **reliable** via two human studies with 12 experienced researchers. In Study 1 (Figure 10(a)), annotators rate agreement (1–5) with GPT-5 scores, justifications, and suggestions per dimension. In Study 2 (Figure 10(b)), annotators judge which of two responses to the same review is superior (or tied) per dimension and overall. Results across 1,365 judgments show strong human-LLM alignment (agreement rating $> 4.17/5$, disagreement $< 5\%$) and substantial (Landis and Koch, 1977) inter-annotator agreement on win/loss comparisons (Krippendorff's $\alpha = 0.81$–0.89).

6 Experiments

6.1 Experimental Setup

We evaluate ARG with five state-of-the-art LLMs across nine *REspGen* settings using *REspEval*. Tables 3, 12-14 present results addressing eight research questions detailed in §6.2 and §D.2. We evaluate five leading LLMs, both open-source and proprietary, with strong generation and reasoning capabilities: Phi-4-Reasoning (Abdin et al., 2025) (Phi-4); Qwen3-32B (Yang et al., 2025) (Qwen3); Llama-3.3-70B-Instruct[9] (Llama-3.3); DeepSeek-R1 (DeepSeek-AI et al., 2025) (DeepSeek); and GPT-4o.[10] We select EMNLP24 cases where reviewers explicitly note a score increase (e.g., "I have improved the score"), yielding 48 instances. This ensures that the human responses used as baselines are verifiably effective.[11] We evaluate nine settings that systematically ablate and tune components of *REspGen*. Starting from review-only generation (Setting 1), we add author input as edit strings (Setting 2), with paragraph context (Setting 3), and further augment with v1 retrieval (Setting 4). We then examine impacts of length control (Setting 5), response plan control (Setting 7), and both controls combined (Setting 6). Finally, we evaluate evaluation-guided refinement applied to outputs from Settings 6 and 7, yielding Settings 8 and 9. Detailed descriptions and prompts are provided in Table 11 (§D.1).

6.2 Results and Discussion

6.2.1 Why Author Input Matters

To address ***RQ1: Are LLMs aware of missing information?***, we instruct models to insert placeholders (e.g., *[author info: <description>]*) when author-only information is needed. Table 12 (§D.2) shows all models except Phi-4 frequently use placeholders under review-only generation (54.2–95.8%), confirming awareness of missing information. Once any author input is provided (Settings 2–4), placeholder usage drops sharply (0–25%), demonstrating models recognize and leverage supplied information. For ***RQ2: Does author input improve response quality?***, Table 3 shows consistent quality improvements across all models and metrics when author input is added (Settings 2–4 vs. 1), with most gains statistically significant (Table 13, §D.2). While review-only responses (Setting 1) underperform human baselines, all models surpass them in most author-input settings. Together, these results demonstrate that author input is both necessary and effective for ARG, motivating our author-in-the-loop framework.

[7] https://aclrollingreview.org/authors; https://peerj.com/benefits/academic-rebuttal-letters/

[8] https://deviparikh.medium.com/how-we-write-rebuttals-dc84742fece1

[9] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[10] https://openai.com/index/gpt-4o-system-card/

[11] All responses are effective in a broad sense (papers accepted), we focus on cases where effectiveness is explicitly confirmed. PeerJ lacks score changes, and EMNLP24 releases only final scores; thus, we identify clear effectiveness through explicit reviewer statements.

Metric		GFP			ICR			lenC			PlanC				Quality		
		%sup	%unsup	%con	%sup	%unsup	%con	#w	%met	m.diff	P	R	F1	OF	Targ	Spec	Conv
Human		.458	.453	.089	.200	.756	.044	115	/	/	/	/	/	/	.788	.575	.575
Phi-4	1.noAIx	.362	.542	.096	.300	.926	.044	161	/	/	/	/	/	/	.775	.446	.483
	2.wAIx_①S	.575	.374	**.051**	**.509**	.450	.042	127	/	/	/	/	/	/	.821	.563	.579
	3.wAIx_②+context	.577	.364	.059	.470	.494	**.036**	428	/	/	/	/	/	/	.783	.583	.592
	4.wAIx_③+v1	.705	.236	.059	.358	.592	.050	343	/	/	/	/	/	/	.771	.579	.579
	5.+Cont._①lenC	**.748**	.200	.052	/	/	/	343	**.458**	-27	/	/	/	/	.779	.583	.579
	6.+Cont._②lenC&planC	.673	.263	.064	/	/	/	284	.250	-55	.471	.642	.497	.755	.829	.600	.613
	7.+Cont._③planC	.680	.253	.067	/	/	/	284	/	/	**.485**	.644	**.504**	**.791**	.821	.583	.596
	8.+Refine_Cont.②	.489	.442	.069	/	/	/	312	.104	-128	.387	.671	.444	.790	**.929**	**.733**	**.725**
	9.+Refine_Cont.③	.490	.443	.068	/	/	/	368	/	/	.368	**.691**	.434	.729	**.929**	.713	**.725**
Qwen3	1.noAIx	.308	.566	.126	.024	.930	.046	123	/	/	/	/	/	/	.808	.508	.554
	2.wAIx_①S	.618	.342	.040	**.628**	.343	.029	127	/	/	/	/	/	/	.892	.638	.654
	3.wAIx_②+context	.643	.317	.040	.572	.404	**.025**	164	/	/	/	/	/	/	.875	.683	.683
	4.wAIx_③+v1	**.744**	.214	.042	.496	.463	.041	205	/	/	/	/	/	/	.913	.721	.717
	5.+Cont._①lenC	.734	.223	.044	/	/	/	125	**1.00**	38	/	/	/	/	.904	.700	.700
	6.+Cont._②lenC&planC	.724	.224	.052	/	/	/	130	.958	33	.498	.696	.534	.842	.913	.700	.700
	7.+Cont._③planC	.719	.252	**.028**	/	/	/	216	/	/	.429	**.793**	.522	.826	.938	.725	.725
	8.+Refine_Cont.②	.576	.373	.050	/	/	/	142	.896	21	**.506**	.678	**.544**	.807	.938	.771	.758
	9.+Refine_Cont.③	.586	.380	.035	/	/	/	290	/	/	.385	.752	.459	**.847**	**.983**	**.842**	**.800**
Llama-3.3	1.noAIx	.483	.414	.103	.054	.912	.034	126	/	/	/	/	/	/	.763	.396	.438
	2.wAIx_①S	.766	.215	**.019**	**.664**	.319	**.017**	169	/	/	/	/	/	/	.800	.550	.567
	3.wAIx_②+context	.760	.217	.023	.534	.426	.040	183	/	/	/	/	/	/	.850	.608	.608
	4.wAIx_③+v1	.771	.173	.056	.420	.542	.039	198	/	/	/	/	/	/	.829	.588	.575
	5.+Cont._①lenC	**.820**	.142	.038	/	/	/	82	**1.00**	82	/	/	/	/	.829	.513	.517
	6.+Cont._②lenC&planC	.788	.157	.055	/	/	/	82	**1.00**	84	**.619**	.470	.490	.728	.804	.467	.504
	7.+Cont._③planC	.770	.183	.047	/	/	/	214	/	/	.486	.705	**.533**	**.825**	.850	.575	.592
	8.+Refine_Cont.②	.657	.261	.082	/	/	/	125	.875	52	.589	.545	.508	.718	**.892**	.667	.638
	9.+Refine_Cont.③	.647	.319	.034	/	/	/	304	/	/	.372	**.707**	.444	.806	.888	**.750**	**.700**
DeepSeek	1.noAIx	.412	.491	.097	.046	.913	.041	113	/	/	/	/	/	/	.771	.433	.496
	2.wAIx_①S	.720	.273	**.007**	**.695**	.272	**.033**	154	/	/	/	/	/	/	.850	.600	.621
	3.wAIx_②+context	.702	.279	.019	.584	.374	.041	172	/	/	/	/	/	/	.817	.608	.617
	4.wAIx_③+v1	.738	.232	.031	.452	.514	.035	194	/	/	/	/	/	/	.904	.692	.700
	5.+Cont._①lenC	**.815**	.144	.042	/	/	/	96	**1.00**	64	/	/	/	/	.879	.642	.638
	6.+Cont._②lenC&planC	.762	.203	.035	/	/	/	93	**1.00**	63	.626	.582	.563	.779	.867	.588	.625
	7.+Cont._③planC	.754	.218	.028	/	/	/	179	/	/	.554	**.710**	.577	.823	.888	.663	.671
	8.+Refine_Cont.②	.728	.231	.041	/	/	/	97	**1.00**	63	**.661**	.585	**.587**	.852	.913	.704	.704
	9.+Refine_Cont.③	.734	.238	.028	/	/	/	194	/	/	.585	.709	.585	**.861**	**.925**	**.746**	**.742**
GPT-4o	1.noAIx	.443	.467	.090	.033	.928	.039	247	/	/	/	/	/	/	.825	.479	.547
	2.wAIx_①S	.689	.284	.027	**.668**	.301	.031	265	/	/	/	/	/	/	.821	.600	.625
	3.wAIx_②+context	.708	.282	**.010**	.571	.400	**.029**	311	/	/	/	/	/	/	.817	.629	.629
	4.wAIx_③+v1	**.781**	.192	.028	.432	.532	.036	339	/	/	/	/	/	/	**.929**	.688	.708
	5.+Cont._①lenC	.774	.198	.028	/	/	/	158	.854	11	/	/	/	/	.867	.633	.633
	6.+Cont._②lenC&planC	.744	.238	.019	/	/	/	156	**.917**	10	.506	.744	**.567**	.834	.879	.596	.621
	7.+Cont._③planC	.762	.215	.023	/	/	/	336	/	/	.386	.784	.477	.794	.913	.692	.700
	8.+Refine_Cont.②	.715	.268	.017	/	/	/	163	.792	5	**.507**	.688	.554	.772	.900	.675	.675
	9.+Refine_Cont.③	.695	.277	.029	/	/	/	367	/	/	.373	**.790**	.470	**.880**	.925	**.721**	**.721**

Table 3: Evaluation results across five LLMs and nine settings. Metrics cover input utilization (GFP, ICR), controllability (lenC, planC), and response quality (*Targ*, *Spec*, *Conv*) (§5). Scores are normalized to [0,1]. Best results per LLM per metric are bolded, and top-three results across LLMs are marked in green.

6.2.2 Input Design and Its Impact

To address ***RQ3: How do input format and detail affect input utilization?***, we analyze Settings 2 (edit string), 3 (plus paragraph context), and 4 (add v1 context) using *GFP* and *ICR* (Table 3). As input context increases, all LLMs incorporate more input-supported facts, reflected by rising GFP support rates from Setting 2 to 4. In Setting 4, all LLMs achieve high GFP support (70.5–78.1%) with low contradiction rates (2.8–5.9%), indicating strong factual grounding and limited hallucination. In contrast, ICR support, which measures coverage of core author improvements, decreases with richer inputs, suggesting that additional context can dilute focus on the most critical information. Importantly, ICR contradiction rates remain low across all settings, indicating no increased hallucination. Together, these results reveal a trade-off: richer input improves overall factual grounding but may reduce emphasis on core information. For ***RQ4: How does input design affect response quality?***, Table 3 shows that, among Settings 2-4, Qwen3, DeepSeek, and GPT-4o achieve best quality metrics under Setting 4, while Llama-3.3 performs best under Setting 3 and Phi-4 shows only marginal differences. Overall, richer input generally improves quality, though the optimal setting is model-dependent.

6.2.3 Controllability and Its Impact

To answer ***RQ5: How well can models satisfy single- and multi-attribute controls in ARG?***, we evaluate length control (Setting 5), response plan control (Setting 7), and both controls combined (Setting 6) using *lenC* and *planC* metrics (Table 3). Length control (Setting 5) is effective for Qwen3, Llama-3.3, and DeepSeek (100% met), moderately so for GPT-4o (85.4%), but weak for Phi-4 (45.8%). Under plan control (Setting 7), all models achieve high label recall and order fidelity, indicating general adherence to the prescribed structure, though occasional extra actions reduce label precision. With multi-attribute control (Setting 6), Qwen3 and Phi-4 degrade notably in length control, while Llama-3.3 and DeepSeek maintain length adherence but exhibit reduced plan recall, F1, and order fidelity. GPT-4o is the only model that improves under joint control. Qwen3, Llama-3.3, and DeepSeek handle single-attribute control well, whereas Phi-4 consistently struggles with length control. For ***RQ6: How do different controls affect response quality?***, we compare Settings 5–7 against the unconstrained Setting 4. Length control alone (Setting 5) substantially degrades quality for all models except Phi-4, likely by limiting space for detailed argumentation. Adding plan control (Setting 6) improves Phi-4, minimally affects Qwen3, and slightly degrades Llama-3.3, DeepSeek, and GPT-4o. Among controlled settings, plan-only control (Setting 7) yields the best quality for all models except Phi-4 and matches Setting 4 performance. Overall, LLMs handle single-attribute control well, but joint multi-attribute control remains challenging and often degrades response quality, with length constraints as the primary bottleneck.

6.2.4 Refinement Effectiveness

To address ***RQ7: How does REspEval feedback improve responses?***, we analyze refinement applied to outputs from Settings 6 and 7, yielding Settings 8 and 9. Table 14 (§D.2) shows refinement produces statistically significant gains across all LLMs, settings, and quality metrics,[12] confirming the effectiveness of evaluation-guided refinement. Improvements are most pronounced for initially weak responses (scores < 3) and diminish as initial quality increases. Consistently, Phi-4 and Llama-3.3 show the highest improvement rates in *Specificity* (Figure 3) and *Convincingness* (Figure 11, §D.2), as they start with more weak responses that offer greater refinement potential. Refined responses are generally longer (Table 14), reflecting added detail for improvement. While length adherence slightly decreases after refinement in Setting 8 (except DeepSeek, which remains perfect), response plan controllability is largely preserved (Table 3). Across Settings 8 and 9, Qwen3 achieves the highest quality metrics. Overall, the proposed evaluation-guided refinement substantially improves response quality while largely maintaining controllability, demonstrating practical value for iterative ARG systems. Further discourse analysis and discussion are presented in §D.2.1.

Figure 3: Changes in *Specificity* after refinement across five LLMs. Colors indicate increase (green), no change (yellow), or decrease (red); the first bar shows overall proportions, followed by distributions by initial score.

7 Conclusion

We introduced *REspGen*, an author-in-the-loop ARG framework with explicit author input, controllable generation, and evaluation-guided refinement; *Re*^{3}*Align*, the first dataset enabling this new formulation; and *REspEval*, a comprehensive evaluation suite spanning controllability, input utilization, discourse, and response quality. Experiments reveal several key insights, including the necessity and effectiveness of author-in-the-loop ARG; the benefits of richer input contexts for improved factual grounding alongside a dilution of focus on core information; trade-offs between controllability and quality, especially under length limits; and the effectiveness of evaluation-guided refinement in improving response quality while preserving controllability. Our dataset, generation framework, and evaluation tools provide foundational resources for future research in NLP for peer review, controllable generation, and human–AI collaboration.

[12]The only exception is *Targeting* in Setting 9 with GPT-4o, where initial scores are already strong.

Limitations

This study has several limitations that should be considered when interpreting the results. From a data perspective, our study is restricted to English-language scientific publications, reflecting the limited availability of openly licensed source data. Examining the transferability of our findings to additional languages, domains, and application settings remains an important direction for future work and can be supported by our publicly released annotation models and analysis and evaluation tools. From a modeling perspective, the implementations and empirical results presented in this study are intended to illustrate the proposed task, generation, and evaluation frameworks. Their primary purpose is to establish technical feasibility and to lay the groundwork for the development of future NLP systems for collaborative author response writing that integrate human expertise and intent with AI assistance. Consequently, the provided implementations have inherent limitations. For example, our approach selectively employs state-of-the-art LLMs and does not systematically evaluate alternative architectures, fine-tuning-based methods, or smaller models. A comprehensive exploration of modeling approaches for the proposed task lies beyond the scope of this work and is left for future research, which can build on the publicly released dataset.

Ethical Considerations

All source data used in this work are publicly available under Creative Commons licenses (CC BY 4.0 and CC BY-NC 4.0). Data collection in the original sources followed ethical guidelines, and our dataset construction and redistribution adhere to the original licensing terms. Our *Re^{3}Align* dataset is released under a CC BY-NC 4.0 license. Human annotation was conducted by experienced researchers who participated voluntarily and without financial compensation. Annotators were informed of the study's purpose and provided consent for the use and publication of their annotations. This study does not involve the collection or processing of personal or sensitive information. To protect privacy, author, reviewer, and annotator identities have been excluded from the analysis and data release.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning technical report. *Preprint*, arXiv:2504.21318.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *ArXiv*, cs.CL/2306.12587.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*, PP:1–1.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(*Long and Short Papers*), pages 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Junjie Huang, Win bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. 2023. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. What can natural language processing do for peer review? *Preprint*, arXiv:2405.06563.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review. *Computational Linguistics*, 48(4):949–986.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. Aspect-aware unsupervised extractive opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Moprd: A multidisciplinary open peer review dataset. *Neural Comput. Appl.*, 35(34):24191–24206.

Guillermo Macbeth, Eugenia Razumiejczyk, and Rubén Ledesma. 2011. Cliff's delta calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica*, 10:545–555.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

William Noble. 2017. Ten simple rules for writing a response to reviewers. *PLOS Computational Biology*, 13:e1005730.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024a. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15049–15067, Miami, Florida, USA. Association for Computational Linguistics.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024b. Re3: A holistic framework and dataset for modeling collaborative document revision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4635–4655, Bangkok, Thailand. Association for Computational Linguistics.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86 2:420–8.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as a multi-turn and long-context dialogue with role-based interactions. *Preprint*, arXiv:2406.05688.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *Preprint*, arXiv:2505.07920.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

A Dataset Construction

A.1 Preprocessing

We preprocess each paper by first grouping its submission versions (v1 and v2), peer reviews, and author responses under a unified paper identifier and preserving all associated metadata if present (e.g., timestamps, reviewer scores, decision information). We then convert every document into the intertextual graph (ITG) representation (Kuznetsov et al., 2022), which encodes hierarchical structure and text order through node–edge relationships. Following Ruan et al. (2024b), we augment each ITG with sentence-level nodes produced via an assembled sentence-segmentation model, ensuring consistent granularity across documents and facilitating downstream alignment between review comments, responses, and revision edits at the sentence and segment levels.

A.2 Matching Algorithm

Given a review sentence s_1 and a response sentence s_2, we treat them as a match if any of the following conditions holds: (1) s_1 contains s_2; (2) s_2 contains s_1; (3) their SBERT similarity (Reimers and Gurevych, 2019) exceeds a threshold t_0; or (4) their partial-string fuzzy-matching score[13] exceeds a threshold t_1. The thresholds t_0 and t_1 were selected based on a pilot study over 20 examples,

[13] `fuzzywuzzy.partial_ratio`

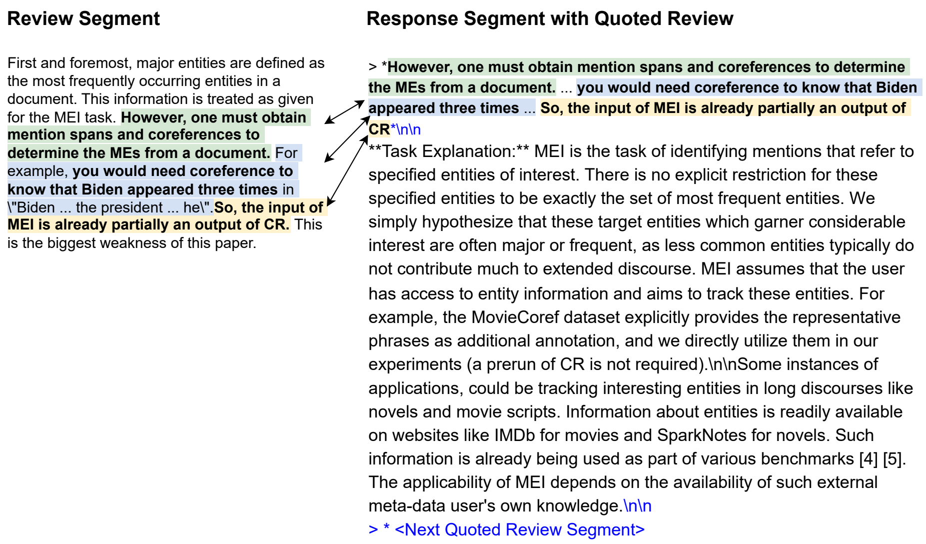


Figure 4: An illustrative example of segment-level review-response pair matching.

with the optimal configuration determined to be $t_0 = 85$ and $t_1 = 85$. Figure 4 presents an illustrative example of a matched review–response segment pair identified through the longest contiguous sentence-matching span.

After constructing initial review–response segment pairs, we apply several quality filtering steps to remove problematic cases. Specifically, we discard response segments that are too short (fewer than 2 sentences), which typically arise from noisy quotation matching or incomplete segmentation, and segments that are excessively long (more than 15 sentences), which usually indicate structural inconsistencies such as merged replies or missing quotation boundaries. After filtering, we retain 2,108 review–response segment pairs from EMNLP24 and 13,963 from PeerJ (Table 2), providing a high-quality basis for subsequent alignment with revision edits.

A.3 Re^3 Triplet Alignment

The task of Re^3 triplet alignment is to determine whether an annotated edit $e_{ij} \in E$ is related to a given review–response segment pair $p^k_{mn} \in P$. We use a two-way alignment strategy: (i) aligning the review segment c^k_m with each sentence edit $e(x^{t+1}_i, x^t_j)$ via a function set CE, and (ii) aligning the response segment a^k_n with the same edit via a function set AE. Each function set combines a fine-tuned state-of-the-art LLM classifier to capture semantic relations (i.e., CE_{llm}, AE_{llm}), together with a lightweight similarity-based component (i.e., CE_{sim}, AE_{sim}) for efficiency.

To maintain matching granularity, we split each review and response segment into sentences, denoted by $c^k_{mp} \in c^k_m$ and $a^k_{nq} \in a^k_n$. For review comment–edit alignment (CE), we enumerate all textual pairs $(s_1 = c^k_{mp}, s_2 = x^t_j + \text{“\textbackslash n”} + x^{t+1}_i)$. For author response–edit alignment (AE), we analogously consider all pairs $(s_1 = a^k_{nq}, s_2 = x^t_j + \text{“\textbackslash n”} + x^{t+1}_i)$

Similarity Matching. Given a review or response sentence (s_1) and the combined sentence edit representation (s_2), CE_{sim} and AE_{sim} identify a pair as a positive match if *all* of the following conditions hold: (i) their partial-string fuzzy matching score is ≥ 60, (ii) their SBERT similarity is ≥ 20, and (iii) their bigram overlap score is ≥ 10. These conditions and thresholds were optimized in a pilot study to ensure high precision. All positive matches are aggregated into the set of aligned sentence-level edits for the segment. This lightweight component serves to efficiently identify edits with high lexical or semantic similarity to the review or response text.

LLM Classifier. Edits related to a review–response pair do not necessarily exhibit high surface similarity to the review or response text. To capture such cases and reduce missed alignments, we fine-tune two LLM classifiers, CE_{llm} and AE_{llm}, following prior work (D'Arcy et al., 2023; Ruan et al., 2024a,b). Ruan et al. (2024a) show that LLMs can be fine-tuned using their SeqC approach to achieve state-of-the-art performance on several classification tasks, including edit intent classification, a closely related task requiring paired inputs and fine-grained understanding of sentence edits. Following this approach, we concatenate s_1 and s_2 as the model input and fine-tune a set of base LLMs to perform binary classification of positive vs. negative alignment. The base models include Llama 2-13B (Touvron et al., 2023) and Llama 3-8B (Grattafiori et al., 2024), the top-performing models identified by Ruan et al. (2024a), as well as the newly released Llama 3.2-3B model.[14]

We use existing human-annotated data available from prior work. For fine-tuning the review comment–edit alignment classifier ($\mathrm{CE_{llm}}$), we combine 466 positive sentence-level samples from Re3-Sci (Ruan et al., 2024b) with paragraph-level samples from ARIES (D'Arcy et al., 2023) that we decompose into 213 sentence-level instances, and generate 2,737 negative samples from Re3-Sci. The final dataset is split into train/validation/test sets (1,989/372/1,055). For the author response–edit alignment classifier ($\mathrm{AE_{llm}}$), we use 1,364 positive samples from Re3-Sci and create 4,092 negative samples, split into train/validation/test sets (3,819/818/819).

Table 4 reports the performance of the fine-tuned LLMs, from which we select the best-performing classifiers for annotation. We use the fine-tuned Llama 2–13B model as $\mathrm{CE_{llm}}$ (96.3% accuracy, 83.4 F1) and the fine-tuned Llama 3–8B–Instruct model as $\mathrm{AE_{llm}}$ (93.7% accuracy, 92.0 F1). Similar to the similarity-based approach, we enumerate all pairs of (s_1, s_2), and aggregate all samples receiving positive alignment predictions into the set of aligned edits for the segment.

Finally, all aligned edits identified by $\mathrm{CE_{sim}}$, $\mathrm{CE_{llm}}$, $\mathrm{AE_{sim}}$, and $\mathrm{AE_{llm}}$ are aggregated into the set $[e_{\text{align}}]$, which is then used to construct each triplet sample $t^k_{mn} = \left(c^k_m, a^k_n, [e_{\text{align}}]\right)$.

Base LLM	Accuracy	F1
Llama2-13b	**96.3**	**83.4**
Llama2-13b-chat	94.9	79.6
Llama3-8b	94.4	74.1
Llama3-8b-instruct	94.8	76.3
Llama3.2-3b	96.1	80.8
Llama3.2-3b-chat	95.6	80.9

(a) LLM classifier performance on review comment-edit alignment. The best-performing classifier (bolded) is used as $\mathrm{CE_{llm}}$.

Base LLM	Accuracy	F1
Llama2-13b	92.3	90.5
Llama2-13b-chat	93.3	91.5
Llama3-8b	93.1	91.8
Llama3-8b-instruct	**93.7**	**92.0**
Llama3.2-3b	92.9	91.1
Llama3.2-3b-chat	93.0	90.9

(b) LLM classifier performance on author response-edit alignment. The best-performing classifier (bolded) is used as $\mathrm{AE_{llm}}$.

Table 4: Fine-tuned LLM classifier performance.

B REspGen

B.1 Item-Based Response Planning

Table 5 provides definitions of our review item types and their mappings to the taxonomy of Kennard et al. (2022). Following their empirical findings that author responses most directly and explicitly address review actions of type *Request* and *Evaluative*, we rename *Evaluative* as *Criticism* and split *Request* into two categories: *Question* (requests for information) and *Request* (requests for changes). Table 6 presents our response action taxonomy and the corresponding stance classes, largely adapted from Table 3 of Kennard et al. (2022) with minor adjustments to align with our item types.

Figure 5 shows the optimized GPT-5 prompts for review and response analysis. Given a pair of review and response segments, the model identifies and categorizes review items, extracts the corresponding response spans to each review item, and assigns response action labels. The output is a structured JSON, with example results illustrated in Figure 6. Human verification indicates that GPT-5 consistently preserves the JSON format, and analyzing the review and response jointly produces better results than separate analysis, likely due to the additional cross-segment context enabling more reliable linking and reasoning.

B.2 Input Component Configuration

To provide the generator with additional paper-level context, we retrieve the most relevant paragraphs from the original submission using a hybrid retrieval–reranking pipeline. Each paragraph is prepended with its corresponding section title. Given a review segment as the query, we apply a two-stage retrieval procedure: (1) Hybrid first-stage retrieval combining a sparse BM25 retriever (Robertson and Zaragoza, 2009) with a dense retriever built on science-tuned SPECTER2 embeddings (Singh et al., 2022). BM25 captures exact

[14] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

DISAPERE (Kennard et al., 2022)		**REspGen (ours)**	
Label	Definition	Label	Definition
Evaluative	A subjective judgement of an aspect of the paper	Criticism	A subjective judgement of an aspect of the paper
Request	A request for information or change in regards to the paper	Request	A request for change in regards to the paper
		Question	A request for information that requires an explicit answer

Table 5: Review item types.

Stance Class	Response Action Label	Definition
Cooperative	answer question	answer a question
	task has been done	claim that a requested task has been completed
	task will be done in next version	claim that a requested task will be completed in resubmission
	accept for future work	express approval for a suggestion, but for future work
	concede criticism	accept a criticism
Defensive	refute question	reject the validity of a question
	reject criticism	reject the validity of a criticism
	contradict assertion	contradict a statement presented as a fact
	reject request	reject a request from a reviewer
Hedge	mitigate importance of the question	mitigate the importance of a question
	mitigate criticism	mitigate the importance of a criticism
Social	social	non-substantive social text
Other (NonArg)	follow-up question	clarification question addressed to the reviewer
	structure	text used to organize sections of the response
	summarize	summary of the response text
	other	all other sentences

Table 6: Response action labels and the corresponding stance classes. The labels are largely adopted from Table 3 of DISAPERE (Kennard et al., 2022), with minimal adjustments to align with the item types used in our framework.

lexical overlap, while SPECTER2 captures semantic similarity between scientific texts; we combine them to improve recall and robustness, especially when reviewer terminology differs from the paper wording. Scores from these two retrievers are fused using reciprocal rank fusion (Cormack et al., 2009), yielding a robust initial candidate set. (2) Reranking of the top candidates using the BAAI/bge-reranker-v2-m3 cross-encoder (Chen et al., 2024), which provides fine-grained semantic relevance scores. The final top-k paragraphs (five by default) constitute the retrieved paper context (***v1***). This module enriches the generator with topic-relevant scientific background while remaining lightweight and domain-agnostic, and is easily applicable across review–response scenarios.

C REspEval

C.1 Quality Evaluation Procedure

Given the review segment, response, and alignments of review items and response spans, we prompt GPT-5 to assign 5-point scores for targeting, specificity, and convincingness, and to provide both justifications and refinement suggestions for each criterion. The evaluation prompt (Figure 7), specifies explicit scoring rubrics. The evaluation output (Figure 8) includes (i) the three scores (ii) justifications per metric, expressed as concise, evidence-grounded bullet points that separate strengths(+) and weaknesses(-) and reference item IDs when relevant; and (iii) suggestions per metric, consisting of one or two actionable advices that, if implemented, would plausibly raise the score to five.

	mean std	median std	p90 std	max std	ICC
Targeting	.07	0	.58	.58	.94
Specificity	.09	0	.58	.58	.93
Convincingness	.05	0	0	.58	.95

Table 7: Consistency verification. Reported are the mean, median, 90th-percentile, and maximum standard deviation, as well as the Intra-class Correlation Coefficients (ICC), computed across repeated evaluation runs for the three scoring dimensions.

C.2 Quality Evaluation Verification

We conduct comprehensive studies to assess the consistency, robustness, interpretability, and reliability of the evaluation procedure.

Consistency. To assess consistency, we run the evaluation three times on the 48 experimental samples (§6) and report standard-deviation–based (std) statistics together with Intra-class Correlation Coefficients (ICC, Shrout and Fleiss, 1979). As shown in Table 7, both the median and mean std are zero, indicating that, on average, the model assigns iden-

Review and Response Analysis Prompt

```
Input: a peer review comment and an author response. \n
Tasks: Extract questions (Q), weakness criticisms (C), and requests (R) from the review. \n
For each Q/C/R item, find all related response sentences targeting it (can be none).\n
Response Labels ∈ ['answer question','refute question','mitigate importance of the question',\n
    'concede criticism','reject criticism','mitigate criticism','contradict assertion',\n
    'reject request','task has been done','task will be done in next version','accept for future work', \n
    'social', 'structure','summarize','follow-up question', 'other'] \n
Finally, list the rest response sentences that do not target any Q/C/R item
and label them (consider the last 5 non-argumentative labels). \n
Important: Output JSON only, no prose, no reasoning.
Keep field names exact. Empty arrays are allowed. Preserve the review’s original order of first appearance. \n
                        {
                        "questions": [
                            {
                            "review_text": [<list of review sentences about the same point>],
                            "response": [
                                {"text": <response text>, "labels": [<Q labels>]}
                            ],
                            }
                        ],
                        "criticisms": [
                            {
                            "review_text": [<list of review sentences about the same point>],
                            "response": [
                                {"text": <response text>, "labels": [<C labels>]}
                            ],
                            }
                        ],
                        "requests": [
                            {
                            "review_text": [<list of review sentences about the same point>],
                            "response": [
                                {"text": <response text>, "labels": [<R labels>]}
                            ],
                            }
                        ],
                        "other_responses": [
                            {"text": <response text>,
                            "labels": [<labels>]}
                        ]
                        }
```

Figure 5: Optimized prompt to itemize review segments and label response actions.

#involvements	0	1-3	4-6	7-10	>10
(i) as author	0%	16.7%	33.3%	25.0%	25.0%
(ii) as reviewer	8.3%	25.0%	16.7%	16.7%	33.3%

Table 8: Peer-review experience and expertise of the 12 human annotators. Each annotator had substantial peer-review experience, indicated by the number of full review cycles they had participated in as (i) an author or co-author and (ii) a reviewer.

tical scores across runs. For more than half of the samples, all three runs yield exactly the same score. Even at the 90th percentile, the std corresponds to at most a 1-point difference, and no sample exhibits a spread larger than one point (max. std. = 0.58). ICC values further demonstrate excellent reliability (Koo and Li, 2016) across all three evaluation axes. Taken together, these results show that the evaluation procedure is highly stable, with nearly all samples receiving identical or near-identical scores across repeated runs.

Robustness. To verify robustness, we evaluate the experimental samples under three conditions: original, mismatched (paired with a random review), and rewritten by GPT-5 to be less target, specific and convincing. As expected, scores dropped sharply from the original condition to the mismatch condition, with the rewrites showing intermediate degradation (Figure 9). For example, Targeting declined by 1.46 points under rewrites with a large Cliff’s δ (Macbeth et al., 2011) of 0.81, and by 2.50 points under mismatches with an even larger effect ($\delta = 0.96$). Specificity and Convincingness also decreased significantly, with large effect sizes (δ) ranging from 0.63–0.84. All reductions compared to the original group are statistically significant ($p < 10^{-9}$, paired t-tests (Student, 1908)). These results demonstrate that the scoring is robust to perturbations, which reliably distinguishes genuine responses from degraded or irrelevant ones.

Interpretability and Reliability. To assess the interpretability and reliability of our evaluation procedure and its alignment with human judgments,

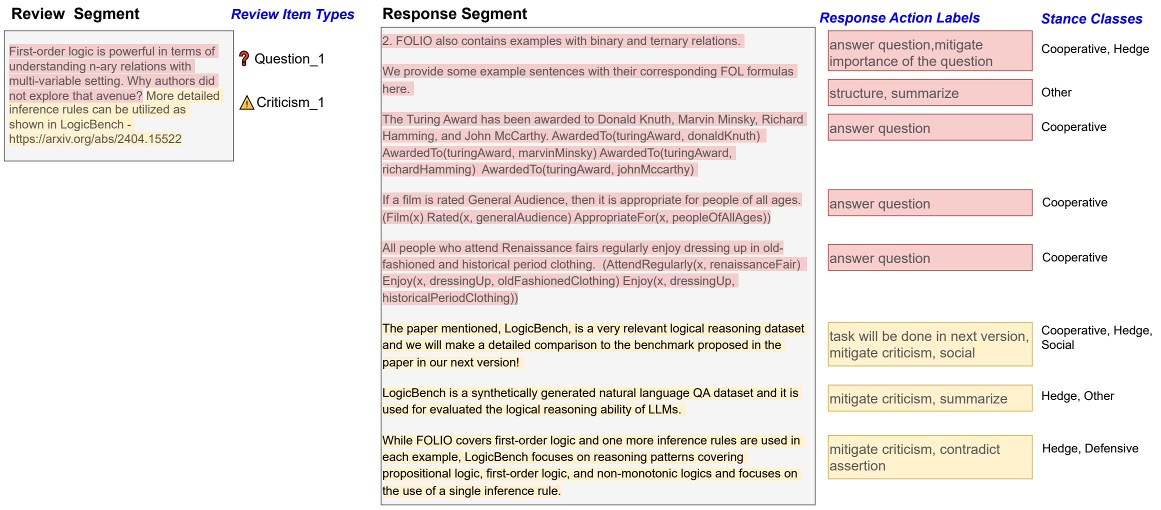


Figure 6: An illustrative example annotated review items, response action labels, and the stance classes.

	%Agree	%Not Disagree	Mean Agreement Rating
Targeting			
- score	86.7%	95.3%	4.38
- justifications	88.6%	97.2%	4.38
- suggestions	81.0%	97.2%	4.22
Specificity			
- score	85.2%	95.2%	4.18
- justifications	84.8%	97.2%	4.17
- suggestions	85.7%	97.2%	4.18
Convincingness			
- score	85.7%	96.2%	4.23
- justifications	83.8%	97.2%	4.21
- suggestions	80.0%	97.2%	4.18

Table 9: Results from Human Study 1: Response Evaluation Verification. Scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree. Reported are the percentages of human agreement (ratings of 4–5), non-disagreement (ratings of 3–5), and the mean agreement ratings for quality scores, justifications, and suggestions across the three scoring dimensions.

	%Agreement	Krippendorff's α
Overall	91.4%	.87
Targeting	83.8%	.81
Specificity	91.9%	.87
Convincingness	90.0%	.89

Table 10: Results from Human Study 2: Author Response Comparison. Reported are the percentages of human–LLM agreement on win/loss decisions across the three quality dimensions and overall, and the inter-annotator agreement measured by Krippendorff's α.

we conduct two human studies with 12 experienced researchers. The annotator pool includes 2 postdoctoral researchers, 8 PhD students, 1 research intern, and 1 professor, all of whom have participated in multiple full peer-review cycles as authors and/or reviewers (Table 8), demonstrating strong domain expertise for the task. Both studies were conducted after a joint review of the annotation guidelines and preparatory discussions.

In **Study 1** (Figure 10(a)), annotators rate their agreement (1–5) with the GPT-5's scores, justifications, and refinement suggestions for *Targeting*, *Specificity*, and *Convincingness*. In **Study 2** (Figure 10(b)), annotators compare two responses to the same review and judge which is better (or a tie) on the three quality dimensions and overall. From these two studies, we collect 1,365 human judges.

Results from **Study 1** (Table 9) show high human agreement with the GPT-5's scores (>85.2%), justifications (>83.8%), and suggestions (>80%). Across all three aspects and all three quality dimensions, mean agreement ratings exceed 4.17 (on a 1–5 scale, where 4 = agree and 5 = strongly agree), and disagreement rates remain below 5%. These findings indicate that the GPT-5's scores and suggestions are highly reliable and that its justifications are generally well-aligned with human interpretability. **Study 2** (Table 10) further demonstrates strong human–LLM agreement on win/loss decisions (83.8%–91.9%) and substantial (Landis and Koch, 1977) inter-annotator agreement (Krippendorff's α = 0.81–0.89).

D Experiments

D.1 Experimental Setup

Table 11 presents detailed descriptions and prompts for each of the nine experimental settings. Experiments with Llama-3.3, Qwen-3, and Phi-4 were conducted on a NVIDIA A100 GPU with 80GB memory.

D.2 Results and Discussion

Table 3 and Table 12 report REspEval evaluation results across five LLMs and nine settings, covering all metrics defined in §5. Table 14, Figure 3 and Figure 11 provide detailed analyses of refinement effectiveness. Table 13 shows a consistent and significant positive impact of author input on response quality across all five LLMs.

D.2.1 Discourse Analysis

To address ***RQ8: How do LLM-generated responses differ from authentic human responses in discourse?***, we analyze tone–stance profiles in Figure 12 and transition patterns in Figure 13. Figure 12 shows that both human-authored responses and those generated by all LLMs across settings predominantly exhibit cooperative and socially friendly tones, with cooperative stances (*%Coop*) being the most prevalent and the combined proportion of cooperative and social stances (*%Coop+%Soc*) exceeding 0.5 (i.e., 50%) in most cases. Both human and LLM-generated responses exhibit limited use of defensive language, with defensive stances (*%Defe*) consistently being the least frequent and remaining below 0.07. In addition, both humans and LLMs frequently employ hedge strategies, with hedging (*%Hed*) being the second most common stance and exceeding 0.20 in most settings. However, when conditioned solely on review comments (Setting 1), LLMs tend to generate overly polite responses, with *%Coop+%Soc* ranging from 0.61 to 0.69, compared to 0.53 for human responses, while employing hedging less frequently than humans (0.16–0.21 vs. 0.28). When explicit author inputs are provided, the proportions of cooperative and social stances decrease substantially toward human levels, while the use of hedging increases markedly. These results highlight the effectiveness of incorporating author inputs in producing responses that are more human-like in tone–stance and less unnecessarily overpolite.

Figure 13 shows that human-authored responses primarily employ social stances (e.g., thanking reviewers) at early positions and rarely conclude with defensive language. The remaining three stance types are distributed more evenly across relative positions. LLaMA-3.3 and DeepSeek exhibit similar patterns, concentrating social language at early positions. In contrast, Phi-4, Qwen3, and GPT-4o place social language at both the beginning and the end of responses, and these patterns remain consistent across different generation settings. In addition, LLM responses tend to exhibit a relatively high proportion of the Other (*Oth*) category at the final position, which in most cases corresponds to summary statements concluding the response.

Response Quality Evaluation Prompt

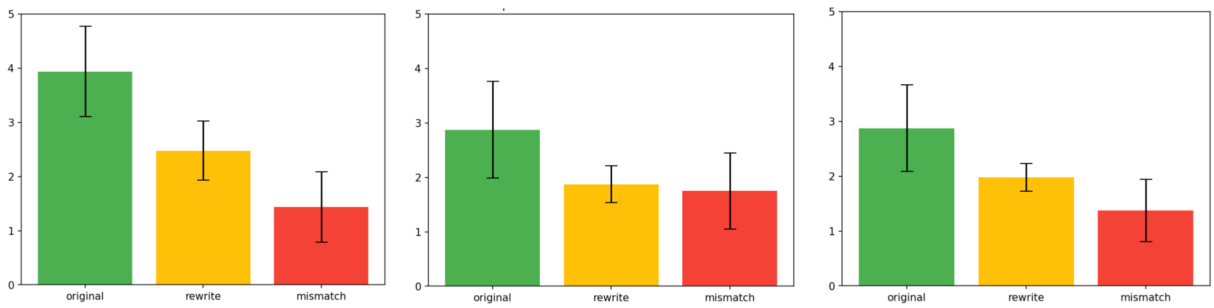
```
You are an impartial LLM judge. Your input is a JSON object with keys:
- "review comment": string (the reviewer's comment)
- "response": string (the author's reply to the comment)
- "paired items": array of objects, each with:
    - "id": string (e.g., "criticism_1"), the item type can be "question", "criticism", or "request"
    - "review_texts": array of strings
    - "response_spans": array of strings that the upstream system believes address the item
- "not linked response spans": array of strings (response parts not linked to any item)
    - "id": string (e.g., "unlinked_1")
    - "response_spans": string
Your task: produce a SINGLE overall evaluation (no per-item scores) of the author's response along three axes:
A) Directness — how clearly the response targets and engages with the reviewer's item(s).
B) Specificity — how much concrete detail and precision the response provides.
C) Convincingness — how persuasive and well-justified the response is.

Use the full "review comment" and "response" for context, but ground your reasoning primarily in the provided
"response_spans" when they are relevant. Do NOT invent facts, numbers, or references that are not present.
If you reference an item in your justifications or suggestions, include its item id in square brackets,
e.g., "[criticism_1]".

---
SCORING RUBRICS (integers 1–5 only):
A. Directness (targeting & alignment with reviewer's item)
- 5 – Very direct: Explicitly and fully engages with the reviewer's item; clear alignment between review
      comment and response.
- 4 – Direct: Addresses the item clearly, though may wander slightly or partially broaden the scope.
- 3 – Partly direct: Some engagement with the item, but diluted or mixed with unrelated content.
- 2 – Weakly direct: Minimal or tangential engagement with the item; mostly off-topic.
- 1 – Not direct: Does not engage with the reviewer's item at all.

B. Convincingness (persuasiveness & justification quality)
- 5 – Very convincing: Directly resolves the concern(s) with strong evidence (data, math, citations, explicit
      section/table/figure references) and clear logic; anticipates counterpoints where relevant.
- 4 – Strong: Substantively addresses the concern(s) with clear reasoning and at least one concrete support
      (e.g., section/table reference or quantitative detail). Minor gaps remain.
- 3 – Moderate: Engages the point(s) and offers some reasoning, but support is partial, qualitative, or
      incomplete; notable uncertainties remain.
- 2 – Weak: Acknowledges the point(s) but relies on assertion or vague justification; little to no concrete support.
- 1 – Not convincing: Ignores/deflects or contradicts without support; non-responsive or purely social niceties.

C. Specificity (precision & concreteness of detail)
- 5 – Very specific: Rich in precise details such as numbers, datasets, metrics, configurations, ablations,
      implementation details, and explicit section/table/figure pointers.
- 4 – High: Multiple concrete details (named components, explicit comparisons, at least one clear reference);
      some fine-grained details may be missing.
- 3 – Moderate: Some specific elements (e.g., naming components or methods) but limited detail; few or no
      numbers/references; scope partly vague.
- 2 – Low: Mostly general statements; promises to "clarify" without specifying where/how.
- 1 – Very vague: Generic acknowledgments; no concrete or actionable detail.
---
OUTPUT REQUIREMENTS:
- Return ONLY valid JSON matching the schema below. No extra prose, no backticks.
- Scores must be integers in [1, 5]. Do NOT output floats or 0.
- Justifications are the reasoning of the scores. + for strengths and - for weaknesses. Keep justifications concise
  (bullet-like strings), tie them to concrete evidence in the response/response_spans, and include ids when relevant.
- Suggestions must be actionable steps (1-2 per metric) that, if implemented, would plausibly raise the score to 5.
---
OUTPUT SCHEMA:
{
  "overall": {
    "directness": 1,
    "specificity": 1,
    "convincingness": 1,
    "justifications": {
      "directness": ["short bullet-like reasons, start with + or -"],
      "specificity": ["short bullet-like reasons, start with + or -"],
      "convincingness": ["short bullet-like reasons, start with + or -"],
    },

    "improve_suggestions_to_5":  {
      "directness": ["1-2 actionable suggestions to improve targeting"],
      "specificity": ["1-2 actionable suggestions to improve details"],
      "convincingness": ["1-2 actionable suggestions to improve persuasiveness"],
    },

  },
  "meta": {
    "confidence": 0.0,
    "judge_notes": "optional short note"
  }
}

FIELD RULES:
- "confidence" is a float in [0,1] reflecting your certainty in the overall assessment.
- If some paired items are not addressed by any meaningful span, reflect this in lower scores and mention their ids
  in justifications (e.g., "No direct engagement with [criticism_2]").
- Do not fabricate section/table/figure numbers. Only cite what appears in the response; if absent, penalize
  appropriately per rubric.

Produce only the JSON object described above.
```

Figure 7: Optimized prompt to evaluate response quality in terms of targeting (directness), specificity, and convincingness.

Response Quality Evaluation_Example Output

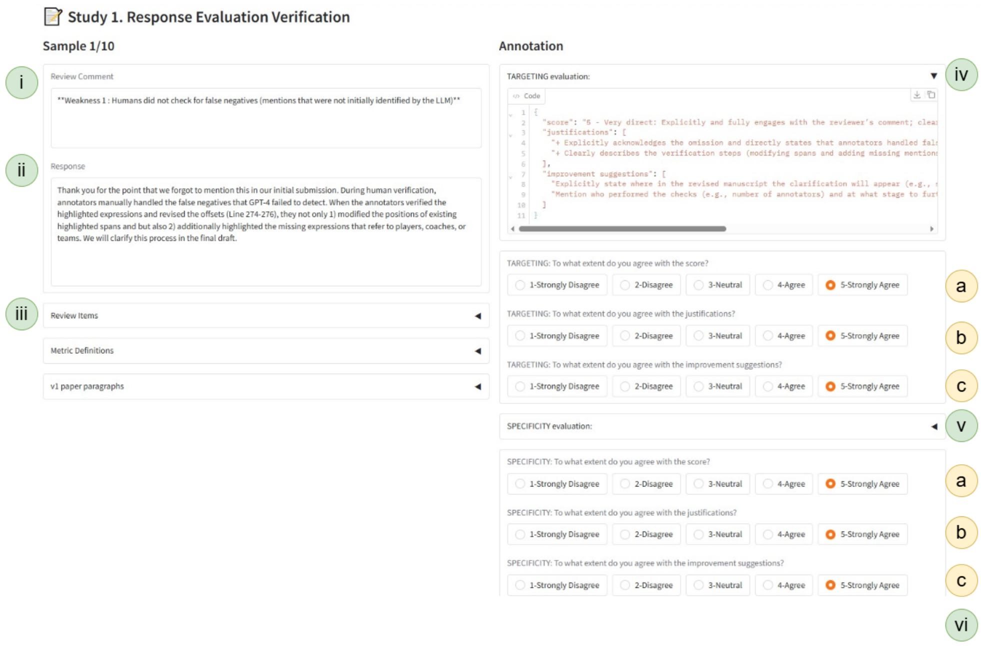

```
{
    "overall": {
        "directness": 4,
        "specificity": 3,
        "convincingness": 3,
        "justifications": {
            "directness": [
                "+ Directly addresses the question about n-ary relations by stating FOLIO includes binary and ternary
                   relations and giving examples [question_1].",
                "+ Responds to the LogicBench suggestion by acknowledging it and contrasting scope/inference-rule usage
                   between FOLIO and LogicBench [criticism_1].",
                "- Does not explicitly explain why higher-arity or more complex multi-variable settings beyond ternary
                   were not explored, leaving part of the \u201cWhy not explore that avenue?\u201d unanswered
                   [question_1]."
            ],
            "specificity": [
                "+ Provides concrete example sentences and corresponding FOL formulas illustrating binary and ternary
                   relations (e.g., AwardedTo, Rated/AppropriateFor, Enjoy with three arguments) [question_1].",
                "+ Gives a qualitative comparison that LogicBench focuses on a single inference rule and includes
                   propositional, FOL, and non-monotonic logics [criticism_1].",
                "- Lacks quantitative details (counts of n-ary instances, distribution, tasks, metrics) and no explicit
                   sections/tables referenced.",
                "- No concrete description of the \u201cmore detailed inference rules\u201d used in FOLIO (types, rule
                   sets, examples of inference chains) or how they compare empirically to LogicBench [criticism_1]."
            ],
            "convincingness": [
                "+ Examples demonstrate capability to handle n-ary relations, partially addressing the concern
                   [question_1].",
                "+ High-level differentiation from LogicBench suggests awareness of design choices [criticism_1].",
                "- Relies on assertions without evidence (no data, ablations, or evaluation showing effectiveness on
                   multi-variable/n-ary reasoning).",
                "- Promises future comparison without providing current justification; does not engage with the
                   suggestion to utilize \u201cmore detailed inference rules\u201d concretely (no list or rationale)
                   [criticism_1]."
            ]
        },
        "improve_suggestions_to_5": {
            "directness": [
                "Explicitly answer why higher-arity (beyond ternary) and broader multi-variable settings were not
                    explored (e.g., dataset scope or model limitations) and/or what is planned to address them.",
                "Directly state whether and how \u201cmore detailed inference rules\u201d from LogicBench will be
                    incorporated or why they are out of scope, tying back to the reviewer\u2019s suggestion."
            ],
            "specificity": [
                "Provide statistics on n-ary relation coverage (counts, proportion by arity) and examples of
                      multi-variable inference chains; cite sections/tables where these appear.",
                "Detail the inference rules used in FOLIO (names/types, number of rules per example, sample derivations)
                      and give a concrete mapping/contrast to LogicBench\u2019s rule patterns."
            ],
            "convincingness": [
                "Include empirical evidence (evaluation results or ablations) demonstrating performance on
                     n-ary/multi-variable reasoning and multi-rule inference, with references to figures/tables.",
                "Present a concrete comparative analysis with LogicBench (e.g., case studies or controlled experiments)
                     showing strengths/limitations of FOLIO\u2019s multi-rule setting versus LogicBench\u2019s
                     single-rule focus."
            ]
        }
    },
    "meta": {
        "confidence": 0.78,
        "judge_notes": "Good illustrative examples but lacks quantitative/comparative evidence and explicit rationale
                       for scope choices.",
    },
}
```


Figure 8: An illustrative output example of response quality evaluation, including scores, justifications and refinement suggestions per metric.

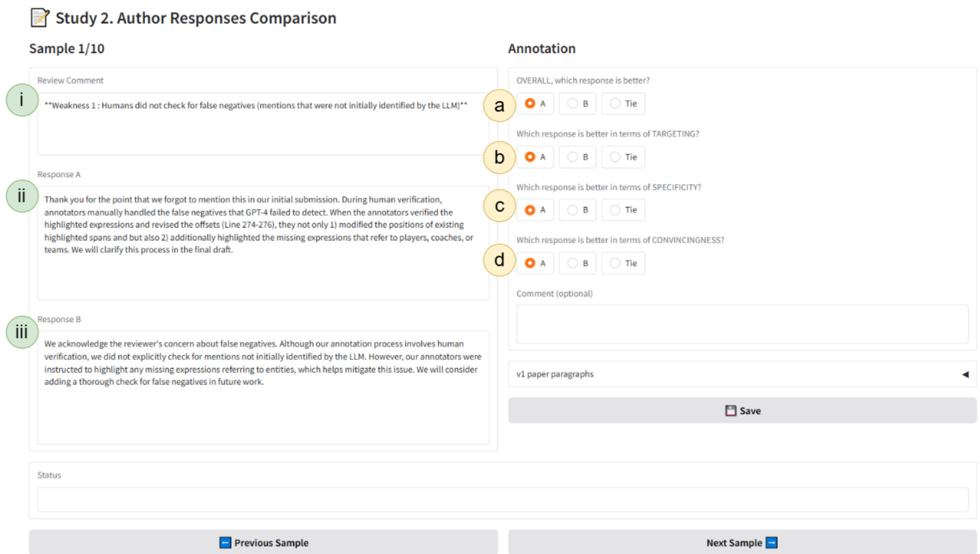


(a) Targeting (b) Specificity (c) Convincingness

Figure 9: Robustness verification. Mean scores and corresponding standard deviation error bars are presented for each scoring dimension across the three test conditions: original (green), rewritten (yellow), and mismatched (red).



(a) Human Study 1. Response Evaluation Verification. Shown on the right are: (i) a review segment, (ii) the corresponding response, and (iii) optional contextual materials available to annotators, including extracted review items, metric definitions and scoring rubrics, and the top five relevant paragraphs retrieved from the original paper for additional topic-specific information. On the left, annotators read three evaluation blocks, Targeting, Specificity, and Convincingness (iv–vi), each containing a score, justifications, and refinement suggestions. For each block, annotators indicate the extent to which they agree with (a) the score, (b) the justifications, and (c) the suggestions on a 5-point scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.



(b) Human Study 2. Author Response Comparison. Shown on the right are: (i) a review segment, (ii) a corresponding Response A, and (iii) another corresponding Response B. On the left, annotators indicate which response is better (or tie) with respect to (b) Targeting, (c) Specificity, and (d) Convincingness, as well as which response wins overall (a).

Figure 10: Annotation interface used in the human studies.

Exp. Setting	Description	Prompt
1. *noAIx*	ARG conditioned solely on the review segment without any additional inputs	L1.You are a research assistant helping authors prepare an author response for a paper under peer review. L2.You will receive: L3. - The reviewer's comment. L4.Your task is to write a specific and convincing response addressing the reviewer's comment. L5. - The review comment is: <review segment>. L6.Output the response only. Do not include any other text.
2. *wAIX* *_①S*	ARG conditioned on the review segment and author input provided as a list of simple edit strings	L1.You are a research assistant helping authors prepare an author response for a paper under peer review. L2.You will receive: L3. - The reviewer's comment. L4. - The author's additional input regarding the comment. L5.Your task is to write a specific and convincing response addressing the reviewer's comment. L6. - The review comment is: <review segment>. L7. - Refer to the author input below: [<edit string>] L8.Output the response only. Do not include any other text.
3. *wAIX* *_②* *+context*	ARG conditioned on the review segment and author input provided as a list of edit strings with paragraph context and section titles	Prompt from **Setting 2** with only modification in : L7.- Refer to the author input below: [«edit string> in <paragraph context> in Section <section title»]
4. *wAIX* *_③* *+v1*	**Setting 3** with additional ***v1*** paper content, retrieved as the five most relevant paragraphs from the original submission	L1.You are a research assistant helping authors prepare an author response for a paper under peer review. L2.You will receive: L3. - The reviewer's comment. L4. - The author's additional input regarding the comment. L5.Your task is to write a specific and convincing response addressing the reviewer's comment. L6. - The review comment is: <review segment>. L7. - Here are the top 5 paragraphs retrieved from the original paper: [<retrieved paragraph>] L8. - Refer to the author input below: [«edit string> in <paragraph context> in Section <section title»] L9.Output the response only. Do not include any other text.
5.+*Cont.* *_①lenC*	**Setting 4** with additional generation length control	Prompt from **Setting 4** adds: L10.Please limit the response to NO MORE than <*lenC*> words.
6. *+Cont.* *_②* *lenC* *&planC*	**Setting 4** with additional generation length control and response plan control	L1.You are a research assistant helping authors prepare an author response for a paper under peer review. L2.You will receive: L3. - The reviewer's comment. And extracted items from the review comment, including questions, criticisms and requests. L4. - The author's additional input regarding the comment. L5.Your task is to write a clear and convincing response addressing the reviewer's comment and the items. Make the response coherent, fluent and human-like, without necessarily listing the items. Write a response addressing the review comment and the items based on the given response action plan. L6. - The review comment is: <review segment>. - - The items extracted from the review comment are: - - - questions: [<#<id>: <question»] - - - criticisms: [<#<id>: <criticism»] - - - requests: [<#<id>: <request»] - - The response action plan is: - - - questions: [<#<id>: <*planC* for question»] - - - criticisms: [<#<id>: <*planC* for criticism»] - - - requests: [<#<id>: <*planC* for request»] L7. - Here are the top 5 paragraphs retrieved from the original paper: [<retrieved paragraph>] L8. - Refer to the author input below: [«edit string> in <paragraph context> in Section <section title»] L9.Output the response only. Do not include any other text. L10.Please limit the response to NO MORE than <*lenC*> words.
7.+*Cont.* *_③planC*	**Setting 4** with additional response plan control	Prompt from **Setting 6** removes L10.
8. *+Refine* *_Cont.②*	Refinement of the response generated under **Setting 6**, using its quality and factuality evaluations accompanied by justifications and improvement suggestions	Prompt from **Setting 6** adds: L11.Note: This is a refinement round to improve the quality of the previous generated response based on its evaluation results. L12. - The previous response generated is: <previous response from Setting 6>. L13. - The overall response scores (directness, specificity and convincingness, 5-point scale) and the respective justifications and improvement suggestions: <quality evaluation of previous response> L14. - Factuality score: <GFP score%> of the atomic facts in the previous response are supported by the provided inputs. L15.TASK: Please revise the previous response based on the review comment, the provided inputs and the requirements, as well as the evaluation results above to improve the directness, specificity, convincingness and the factuality of the response. Output the revised response only.
9. *+Refine* *_Cont.③*	Refinement of the response generated under **Setting 7**, using the same evaluation results as **Setting 8**	Prompt from **Setting 8** removes L10.

Table 11: Author response generation with REspGen. Shown are the nine experimental settings, their descriptions, and the prompts used. To explicitly capture the model's awareness of missing information, we also prompt: *Use placeholders like '[author info: <description>]' if you need extra information from the author to address the review comment.* For the EMNLP24 subset, we additionally clarify the review setting by appending the prompt: *This author response is prepared during the rebuttal phase, before submitting any revisions (like in ARR process). You should use the additional author input to address the review comment if they are useful, and may outline future planned changes in the final version if relevant but do not refer to completed revisions.*

Metric		Basic		Polite.	Meta		Tone-Stance Profile					
		RL	BS		#word	%Ph	%Coop	%Defe	%Hed	%Soc	%Other	ArgLoad
Human		/	/	.829	115	-	.454	.048	.277	.076	.145	.779
Phi-4	*1.noAIx*	.140	.834	**.850**	161	0	**.517**	.018	.194	**.115**	.136	**.728**
	2.wAIx_①S	**.158**	**.835**	.798	127	0	.443	.019	.245	.065	.227	.708
	3.wAIx_②+context	.136	.826	.772	**428**	0	.409	**.062**	.221	.077	.212	.691
	4.wAIx_③+v1	.144	.825	.761	343	0	.446	.037	.199	.034	**.284**	.682
	5.+Cont._①lenC	.144	.761	.761	343	0	.444	.034	.222	.054	.246	.699
	6.+Cont._②lenC&planC	.155	.829	.798	284	0	.445	.029	.220	.078	.228	.694
	7.+Cont._③planC	.155	.829	.798	284	0	.434	.026	**.247**	.077	.216	.708
	8.+Refine_Cont.②	.133	.820	.777	312	0	.484	.021	.197	.058	.240	.702
	9.+Refine_Cont.③	.131	.820	.772	368	0	.451	.017	.224	.058	.250	.692
Qwen3	*1.noAIx*	.149	.833	**.841**	123	**95.8**	.499	.035	.205	**.110**	.150	.740
	2.wAIx_①S	.158	.835	.798	127	0	.443	.019	.245	.065	**.227**	.708
	3.wAIx_②+context	.173	.840	.808	164	12.5	.492	.045	.237	.095	.131	.773
	4.wAIx_③+v1	.166	.839	.797	205	4.2	.468	.036	**.263**	.072	.161	.767
	5.+Cont._①lenC	.171	.843	.807	125	25.0	.440	**.054**	.258	.075	.173	.752
	6.+Cont._②lenC&planC	**.178**	**.846**	.825	130	4.2	.462	.040	.230	.082	.187	.732
	7.+Cont._③planC	.171	.842	.823	216	2.1	.448	.039	.251	.089	.174	.738
	8.+Refine_Cont.②	.163	.838	.797	142	2.1	**.520**	.046	.241	.059	.135	**.807**
	9.+Refine_Cont.③	.146	.829	.782	**290**	0	.508	.028	.207	.065	.192	.743
Llama-3.3	*1.noAIx*	.175	.839	.873	126	**83.3**	.524	.026	.159	**.115**	.176	.709
	2.wAIx_①S	.200	.847	**.876**	169	6.3	.466	.025	.208	.098	.203	.699
	3.wAIx_②+context	.206	.848	.863	183	4.2	.432	.026	.230	.084	.229	.688
	4.wAIx_③+v1	**.207**	.849	.853	198	0	.427	.026	**.241**	.085	.220	.695
	5.+Cont._①lenC	.203	**.853**	.843	82	12.5	.529	.034	.216	.075	.147	.779
	6.+Cont._②lenC&planC	.206	.852	.865	82	4.2	.567	**.050**	.220	.069	.094	**.837**
	7.+Cont._③planC	.196	.848	.860	214	0	.468	.026	.229	.081	.196	.723
	8.+Refine_Cont.②	.198	.848	.830	125	2.1	**.595**	.027	.198	.047	.132	.821
	9.+Refine_Cont.③	.166	.838	.830	**304**	2.1	.475	.018	.211	.053	**.243**	.704
DeepSeek	*1.noAIx*	.171	.838	**.870**	113	**83.3**	**.565**	.026	.179	**.126**	.104	.769
	2.wAIx_①S	.198	.845	.843	154	22.9	.449	.021	.219	.095	.218	.688
	3.wAIx_②+context	.199	.844	.855	172	25.0	.429	.018	.214	.090	**.249**	.661
	4.wAIx_③+v1	.192	.846	.835	194	14.6	.439	.043	.232	.077	.210	.714
	5.+Cont._①lenC	**.202**	.850	.824	96	14.6	.428	.051	.222	.081	.219	.700
	6.+Cont._②lenC&planC	.201	**.852**	.863	93	2.1	.459	.050	**.263**	.099	.129	.772
	7.+Cont._③planC	.197	.850	.863	179	2.1	.437	.048	.245	.100	.171	.730
	8.+Refine_Cont.②	.190	.847	.825	97	2.1	.529	**.068**	.253	.087	.063	**.850**
	9.+Refine_Cont.③	.185	.844	.831	**194**	8.3	.528	.057	.242	.085	.088	.827
GPT-4o	*1.noAIx*	.159	.833	**.878**	247	**54.2**	.499	.016	.213	**.148**	.125	.728
	2.wAIx_①S	.172	.839	.859	265	14.6	.475	.019	.186	.117	.183	.680
	3.wAIx_②+context	.157	.836	.839	311	8.3	.478	.041	.143	.094	**.244**	.662
	4.wAIx_③+v1	.162	.837	.829	339	4.2	**.514**	**.061**	.186	.107	.132	**.761**
	5.+Cont._①lenC	.200	.852	.864	158	0	.482	.011	.211	.118	.179	.704
	6.+Cont._②lenC&planC	**.204**	**.854**	.871	156	0	.446	.042	.227	.127	.159	.715
	7.+Cont._③planC	.166	.840	.848	336	0	.473	.027	.221	.111	.168	.721
	8.+Refine_Cont.②	.192	.850	.849	163	16.7	.473	.024	**.229**	.133	.142	.725
	9.+Refine_Cont.③	.155	.837	.825	**367**	8.3	.509	.030	.205	.079	.178	.744

Table 12: Evaluation results for the ARG task across five LLMs and nine settings. Reported metrics include basic similarity measures (Rouge-L (RL), BERTScore(BS)), average sentence politeness[15], metadata including word count (#word) and the proportion of samples with placeholders (%Ph), as well as the Tone-Stance Profile (*%Coop*, *%Defe*, *%Hed*, *%Soc*, *%Other*, *ArgLoad*), see §5 for definitions. All scores and proportions are normalized to [0,1]. For each LLM, the largest value per metric is bolded, and the top three values across LLMs are highlighted in green.

Metric		Targeting		Specificity		Convincingness	
LLM/Setting		t-test	Wilco	t-test	Wilco	t-test	Wilco
Phi-4	1 vs. 2	✗	✓	✓	✓	✓	✓
	1 vs. 3	✗	✗	✓	✓	✓	✓
	1 vs. 4	✗	✗	✓	✓	✓	✓
Qwen3	1 vs. 2	✓	✓	✓	✓	✓	✓
	1 vs. 3	✓	✓	✓	✓	✓	✓
	1 vs. 4	✓	✓	✓	✓	✓	✓
Llama-3.3	1 vs. 2	✗	✗	✓	✓	✓	✓
	1 vs. 3	✓	✓	✓	✓	✓	✓
	1 vs. 4	✓	✓	✓	✓	✓	✓
DeepSeek	1 vs. 2	✓	✓	✓	✓	✓	✓
	1 vs. 3	✗	✓	✓	✓	✓	✓
	1 vs. 4	✓	✓	✓	✓	✓	✓
GPT-4o	1 vs. 2	✗	✗	✓	✓	✓	✓
	1 vs. 3	✗	✗	✓	✓	✓	✓
	1 vs. 4	✓	✓	✓	✓	✓	✓

Table 13: Significance tests for quality improvements comparing Setting 1 with Settings 2–4. Significance is indicated by ✓ ($p < 0.05$) under paired one-sided t-tests (t-test) (Student, 1908) and Wilcoxon tests (Wilco) (Wilcoxon, 1945). Author input, regardless of format or level of detail, yields consistent and significant gains in *Specificity* and *Convincingness*. Only *Targeting* improvements are occasionally non-significant, primarily because baseline scores are already high: LLMs can reliably identify and stay focused on the reviewer's discussion points.

Metric					Quality		
LLM/Setting		len	len↑	eD	Targ	Spec	Conv
Phi-4	6.	284			.829	.600	.613
	8.	312	95.8	.714	.929	.733	.725
	7.	284			.821	.583	.596
	9.	368	93.8	.722	.929	.713	.725
Qwen3	6.	130			.913	.700	.700
	8.	142	85.4	.570	.938	.771	.758
	7.	216			.938	.725	.725
	9.	290	97.9	.549	.983	.842	.800
Llama-3.3	6.	82			.804	.467	.504
	8.	125	89.6	.506	.892	.667	.638
	7.	214			.850	.575	.592
	9.	304	95.8	.499	.888	.750	.700
DeepSeek	6.	93			.867	.588	.625
	8.	97	66.7	.542	.913	.704	.704
	7.	179			.888	.663	.671
	9.	194	68.8	.547	.925	.746	.742
GPT-4o	6.	156			.879	.596	.621
	8.	163	79.2	.409	.900	.675	.675
	7.	336			.913	.692	.700
	9.	367	72.9	.413	.925	.721	.721

Table 14: Refinement effects. Reported are response length before and after refinement (len), the proportion of longer responses after refinement (len↑), edit distance (eD), and mean quality scores for *Targeting*, *Specificity*, and *Convincingness*. All quality metrics improve across all LLMs and settings, with statistically significant gains under both paired one-sided t-tests and Wilcoxon signed-rank tests (marked in green). The only non-significant case is *Targeting* in Setting 9 with GPT-4o, as the initial targeting is already strong.

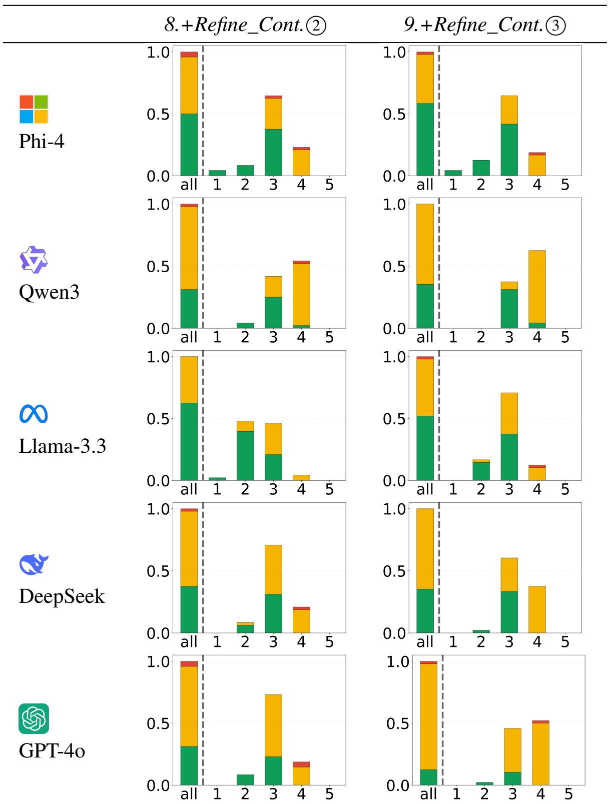


Figure 11: Changes in *Convincingness* after refinement across five LLMs. Colors indicate increase (green), no change (yellow), or decrease (red); the first bar shows overall proportions, followed by distributions by initial score.

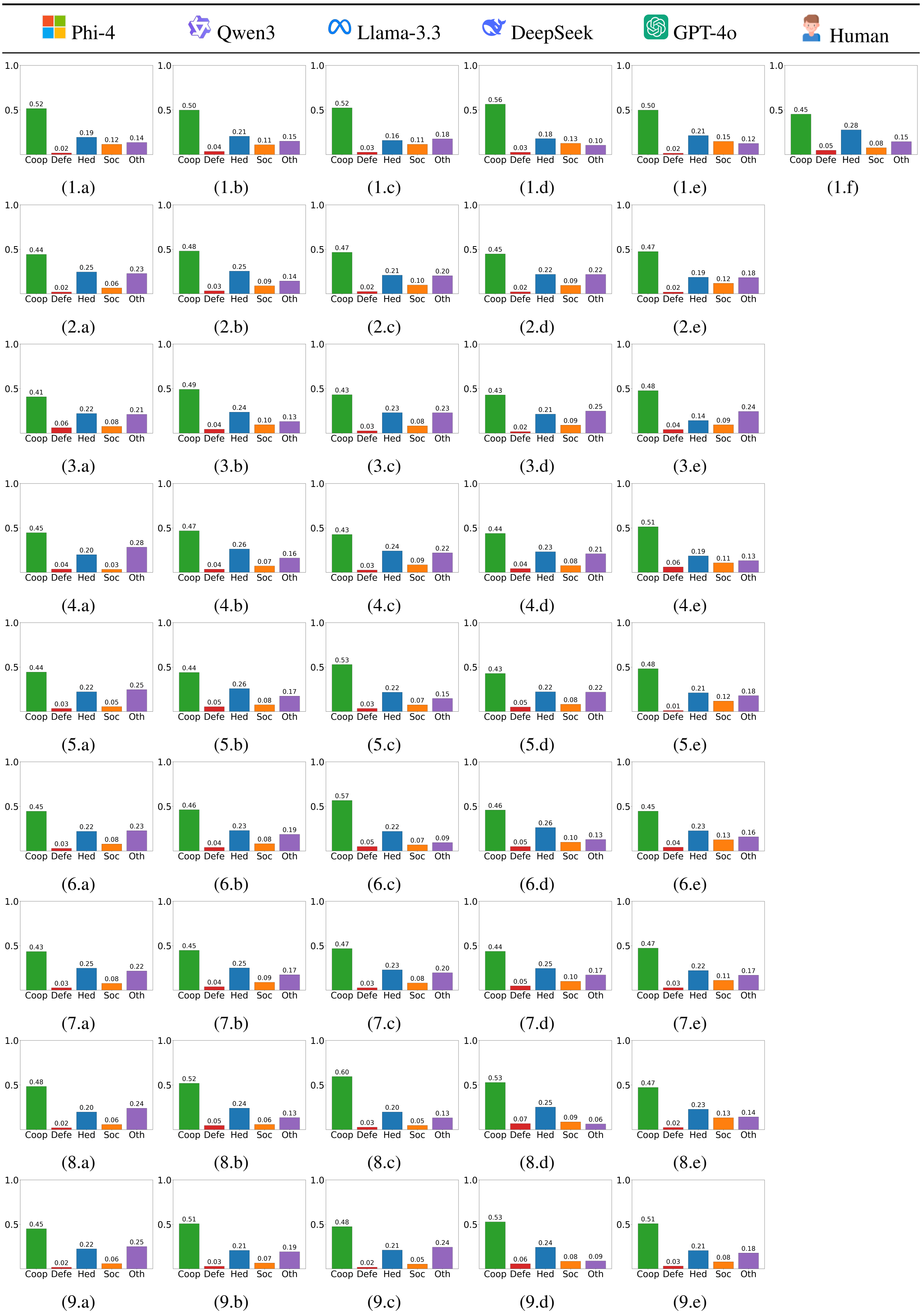


Figure 12: Tone–stance profiles of author responses generated by five LLMs (a–e) under nine settings (1–9), together with authentic human responses (1.f). The figure shows average word-weighted percentages (normalized to [0,1]) of cooperative (Coop), defensive (Defe), hedge (Hed), social (Soc), and other (Oth) stances; detailed definitions are provided in §5.1.

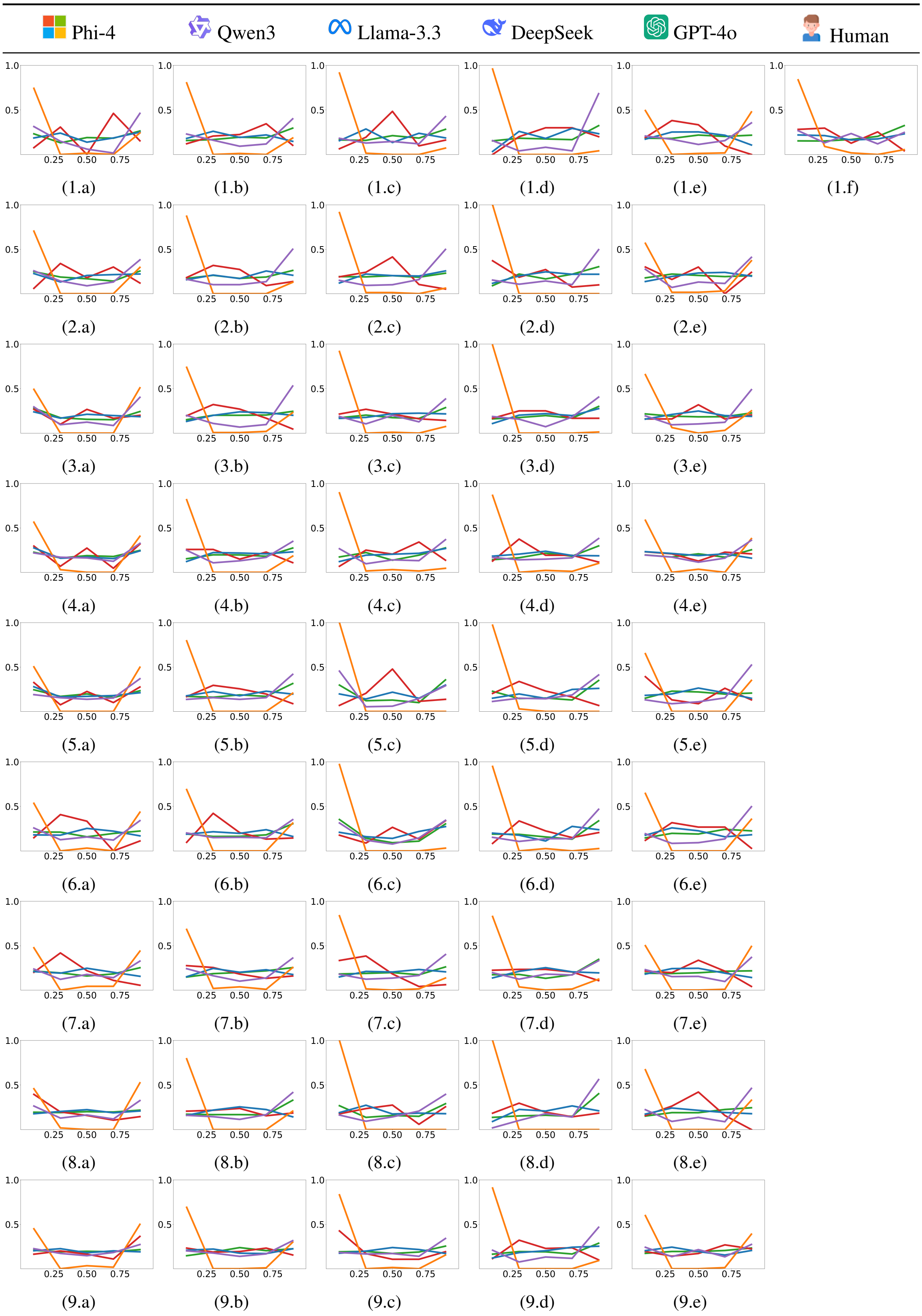


Figure 13: Distributions of stance types across relative positions in author responses generated by five LLMs (a–e) under nine settings (1–9), together with authentic human responses (1.f). The x-axis indicates the relative position within a response, and the y-axis shows the proportion of each stance type at that position. Stance categories include cooperative (green), defensive (red), hedge (blue), social (orange), and other (purple).