RAVEL: REASONING AGENTS FOR VALIDATING AND EVALUATING LLM TEXT SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models have evolved from single-round generators into long-horizon agents, capable of complex text synthesis scenarios. However, current evaluation frameworks lack the ability to assess the actual synthesis operations, such as outlining, drafting, and editing. Consequently, they fail to evaluate the actual and detailed capabilities of LLMs. To bridge this gap, we introduce **RAVEL**, an agentic framework that enables the LLM testers to autonomously plan and execute typical synthesis operations, including outlining, drafting, reviewing, and refining. Complementing this framework, we present C3EBENCH, a comprehensive benchmark comprising $1,258$ samples derived from professional human writings. We utilize a "reverse-engineering" pipeline to isolate specific capabilities across four tasks: CLOZE, EDIT, EXPAND, and END-TO-END. Through our analysis of 14 LLMs, we uncover that most LLMs struggle with tasks that demanding contextual understanding under limited or under-specified instructions. By augmenting **RAVEL** with SOTA LLMs as operators, we find that such agentic text synthesis is dominated by the LLM's *reasoning capability* rather than raw generative capacity. Furthermore, we find that a strong reasoner can guide a weaker generator to yield higher-quality results, whereas the inverse does not hold. Our code and data are available at https://anonymous.4open.science/r/ICLR2027-RAVEL-Reasoning-Agents-Text-Eval/.

1 INTRODUCTION

Large Language Models (LLMs) have evolved from simple monolithic auto-regressive generators (Vaswani et al., 2023; Brown et al., 2020; Ouyang et al., 2022) into long-horizon agents capable of handling complex, sophisticated long-form text synthesis (Yao et al., 2023; Bai et al., 2024; Park et al., 2023; Singh et al., 2025; Gemini Team, Google, 2025; Anthropic, 2025). Most recent frameworks involve LLMs to perform long-form text synthesis in an agentic approach (Yang et al., 2022; Madaan et al., 2023; Schick et al., 2022).

In practical text synthesis scenarios, LLMs have to handle distinct tasks, such as outline planning (Yang et al., 2022; Rashkin et al., 2020), drafting partial sections (Schick et al., 2022; Bai et al., 2024), content-specific editing (Ke et al., 2024; Lin et al., 2024), rather than a simple one-run generation (Wu et al., 2025; Liu et al., 2024). These tasks are usually concurrent and interleaved during the whole synthesis task. However, current evaluation protocols (Zheng et al., 2023; Fu et al., 2023; Kim et al., 2024b; Du et al., 2025) and benchmarks (Wu et al., 2025; Liu et al., 2024; Zheng et al., 2023; Kim et al., 2024a) treat the complex composition as a single run generation, neglecting the intricate capability differences. As a result, they fail to show the actual and detailed capabilities of LLMs during handling real text synthesis.

To align the evaluation framework with the actual LLM capability, we propose **RAVEL** (**R**easoning **A**gents for **V**alidating and **E**valuating **LL**M text synthesis), a framework that integrates the tested LLM to autonomously plan and execute the synthesis operations by itself, expanding the evaluation scope to more complex scenarios. This agentic setup allows us to evaluate the LLM based on its execution trajectory rather than just the final output. Specifically, we feature four representative synthesis operations (outlining, drafting, reviewing and editing) as the action space $\mathcal{A}$ and formalize the process as a sequential decision process (SDP), where the state $\mathcal{S}$ represents the evolving synthesized textual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

content. By reasoning and acting (Yao et al., 2023), the LLM evolves the text and terminates the synthesis until the text meets the quality threshold evaluated by itself.

To complement **RAVEL**, we present C3EBENCH (derived from **C**loze, **E**dit, **E**xpand, **E**nd2End), a comprehensive benchmark consisting of $1,258$ samples and expanding four distinct synthesis scenarios: CLOZE, EDIT, EXPAND, END-TO-END. Unlike existing datasets (Wu et al., 2025; Kim et al., 2024a) that pairing text references from elaborated instructions, C3EBENCH employs a "reverse-construction" pipeline to pair professional human references (golden truth) with instructions, in order to recover the complexity and reality of actual text synthesis.

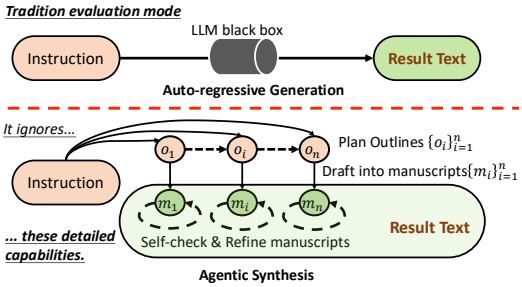


Figure 1: An Illustration of the evaluation gap.

We conduct an extensive empirical analysis of 14 mainstream LLMs, including proprietary models (Singh et al., 2025; Gemini Team, Google, 2025; Anthropic, 2025; Team et al., 2025; GLM-Team, 2025; xAI, 2025; DeepSeek-AI, 2024) and open-source counterparts (Grattafiori et al., 2024; Yang et al., 2025). Using **RAVEL** to roll out trajectories for the tested LLMs, we expose **behavioral divergences** hidden by static leaderboard rankings. Our results show that most LLMs can handle tasks with detailed, clear instructions, yet struggle in underspecified scenarios requiring substantial contextual understanding. We find a clear disconnect between refinement density and quality gains: while models like Claude-4.5 and Qwen3-32B exhibit high refinement density, they often fail to convert iterations into quality gains, contrasting with models such as Gemini-3 Pro that make fewer, more effective refinements. Further ablation studies with **RAVEL** demonstrate that synthesis success is dominated by the *LLM's reasoning ability* to plan and critique, rather than raw draft-refinement capacity. In particular, we find that a strong reasoner can effectively guide a weaker generator to success by 39%, whereas a strong generator cannot compensate for a weak reasoner.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION AND NOTATION

To construct realistic, long-horizon synthesis scenarios for evaluating the capabilities of LLMs, we must move beyond static input-output based evaluations. As discussed, text synthesis is inherently iterative and hierarchical. Consequently, we formalize the synthesis task not as single-step generation, but as a Sequential Decision Process (SDP), denoted by the tuple $\langle \mathcal{Q}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$:

- $\mathcal{Q}$ represents the **query space**. A specific query $q \in \mathcal{Q}$ is defined by a topic T and a style guide G.
- $\mathcal{S}$ is the **state space**. A state $s_t \in \mathcal{S}$ at step t is a pair $\langle \mathcal{O}, \mathcal{M} \rangle$, where $\mathcal{O} = \{o_1, \ldots, o_n\}$ is the set of planned outline nodes, and $\mathcal{M} = \{m_1, \ldots, m_n\}$ are the synthesized text nodes.
- $\mathcal{A}$ is the **action space** consisting of five operations: $\mathcal{A} = \{\text{outline}, \text{draft}, \text{review}, \text{refine}, \text{finish}\}$.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the **transition function**, representing the deterministic or stochastic update of the manuscript based on the execution of an action.
- $\mathcal{R} : \mathcal{S} \times \mathcal{Q} \to \mathbb{R}$ is the **intrinsic reward function**. Crucially, $\mathcal{R}$ is not provided by an external oracle; instead, the agent serves as its own *generative oracle* to evaluate s_t relative to q.

2.2 OPTIMIZATION OBJECTIVE

The policy $\pi_\theta(a_t|s_t, q)$ maps the state and query to action a_t. Because ground-truth feedback is inaccessible during inference for the LLM agent, the process is driven by the internal reward estimator $\mathcal{R}$. Formally, we seek the most efficient trajectory to reach a state s_T satisfying a quality threshold τ:

$$\min_{a_{0:T}} \quad T \quad \text{s.t.} \quad \mathcal{R}(s_T, q) \geq \tau \tag{1}$$

This formulation enforces a closed-loop refinement process where REVIEW and REFINE actions iteratively improve the manuscript until the quality constraint is met.

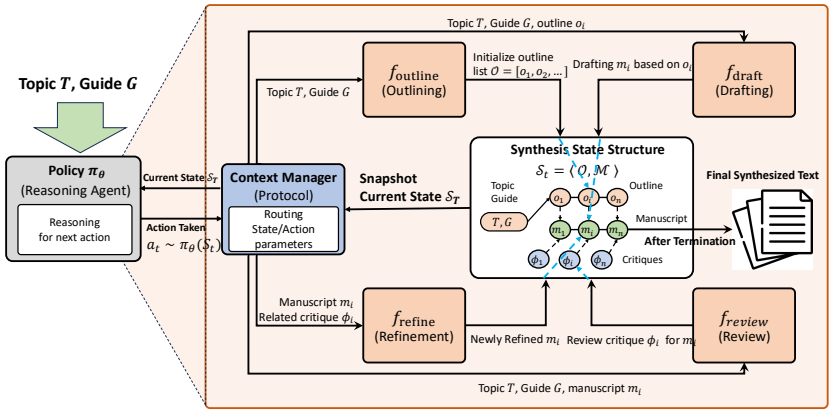


Figure 2: **RAVEL**: an evaluation framework that integrates the testing LLM in agentic text synthesis. After receiving the writing instruction, the testing LLM will autonomously reason the synthesis actions to take, and carry out the actions to update its state.

3 THE **RAVEL** FRAMEWORK

We instantiate the SDP formulation via **RAVEL**, where the LLM functions as an autonomous agent. Unlike static generation pipelines, **RAVEL** operates on a *Reasoning-Act* paradigm (Yao et al., 2023). At each step t, the agent observes the state s_t and samples an action a_t with execution parameters:

$$(a_t, \text{params}_t) \sim \pi_\theta(s_t, \text{prompt}_{sys}) \tag{2}$$

where prompt_{sys} encodes the state transition protocol and action primitives. After the agent executes action a_t, the state updates to $s_{t+1} \leftarrow \mathcal{P}(s_t, a_t, \text{params}_t)$. Therefore, the execution trajectory is denoted as $\mathcal{T} = \{(a_t, \text{params}_t), s_{t+1}\}_{t=0}^{T}$. The transition dynamics $\mathcal{P}$ are defined by four specialized primitives: *Outline, Draft, Review,* and *Refine*, detailed below.

3.1 OUTLINING

The outlining primitive $f_{\text{outline}} : (T, G) \rightarrow \mathcal{O}$ generates the initial structural backbone of the manuscript, where T is the writing topic and G is the genre specification. Each outline node $o_i \in \mathcal{O}$ is defined as a tuple containing textual specifications t_i and a status indicator σ_i:

$$\mathcal{O} = \{o_i\}_{i=1}^{n}, \quad \text{where } o_i = \langle t_i, \sigma_i \rangle \tag{3}$$

The indicator $\sigma_i \in \{\text{pending, drafted, revision_needed , completed}\}$ tracks the lifecycle of each node. Upon execution of f_{outline}, all σ_i are initialized to "pending". This primitive serves to establish global coherence and structural constraints prior to more specified content synthesis.

3.2 GUIDED SYNTHESIS: DRAFTING

The drafting primitive f_{draft} is responsible for generating content for a target outline node. To ensure narrative consistency, we define each manuscript node $m_i \in \mathcal{M}$ as a tuple $\langle t, \phi \rangle_i$, where t represents the textual content and ϕ denotes the associated critique (initially empty).

For a selected node o_i with $\sigma_i =$ "pending", the agent generates $m_i.t$ conditioned on the preceding context $m_{i-1}.t$, the topic T, and the style guide G:

$$m_i.t \leftarrow f_{\text{draft}}(T, G, o_i, m_{i-1}.t,) \tag{4}$$

Upon completion, the status of o_i is updated to $\sigma_i =$ "drafted". This sequential dependency ensures the local synthesis remains aligned with the global discourse flow established by previous nodes.

3.3 QUALITY ASSESSMENT: REVIEW

To simulate the self-checking scenario, the review primitive f_{review} evaluates the manuscript $m_i.t$ against the query constraints $\langle T, G \rangle$. It yields a score $r_i \in [1, 10]$ and a natural language critique ϕ_i:

$$(r_i, \phi_i) \leftarrow f_{\text{review}}(m_i.t, T, G) \tag{5}$$

where r_i and ϕ_i are then stored within m_i to guide subsequent refinement. The state transition for the corresponding outline node o_i is governed by a quality threshold τ:

$$\sigma_i \leftarrow \begin{cases} \text{completed}, & \text{if } r_i \geq \tau \\ \text{revision_needed}, & \text{if } r_i < \tau \end{cases} \tag{6}$$

This thresholding mechanism hints that the agent only proceeds to subsequent nodes or the terminal state once the current content satisfies the internal quality standards.

3.4 Refinement

The refinement primitive f_{refine} simulates the human editorial revision, reflecting the iterative nature of human writing. The agent invokes this primitive to update the manuscript content $m_i.t$ by incorporating the critiques $m_i.\phi$:

$$m_i.t \leftarrow f_{\text{refine}}(m_i.t, m_i.\phi, T, G) \tag{7}$$

where left hand side $m_i.t$ is the revised text intended to address the identified deficiencies. Following the execution of f_{refine}, the status indicator is reset: $\sigma_i \leftarrow$ "drafted" This transition effectively returns the node to the *Review* phase, establishing a non-linear optimization path.

Algorithm 1 RAVEL Agentic Synthesis Loop

Input: Topic T, Style Guide G, Threshold τ
Initialize: $s_0 \leftarrow \langle \emptyset, \emptyset \rangle$, $t \leftarrow 0, a_{-1} \leftarrow \emptyset$
$\mathcal{T} \leftarrow \{a_{-1}, s_0\}$
while $a_t \neq$ finish **and** $t < T_{max}$ **do**
  $a_t \leftarrow \pi_\theta(s_t)$
  if $a_t =$ outline **then**
    $\mathcal{O} \leftarrow f_{plan}(T, G)$
  else if $a_t =$ draft **then**
    $m_i.t \leftarrow f_{\text{draft}}(T, G, o_i, m_{i-1}.t,)$
  else if $a_t =$ review **then**
    $(r_i, \phi_i) \leftarrow f_{\text{review}}(m_i.t, T, G)$
  else if $a_t =$ refine **then**
    $m_i.t \leftarrow f_{\text{refine}}(m_i.t, m_i.\phi, T, G)$
  end if
  UpdateState : $s_{t+1} \leftarrow \mathcal{P}(s_t, a_t)$
  $\mathcal{T} = \mathcal{T} \bigcup \{a_t, s_{t+1}\}$
  $t \leftarrow t + 1$
end while
Return: $\mathcal{M}, \mathcal{T}$

3.5 Termination Criteria

The termination of the agentic workflow is controlled by a dual-criterion termination policy. The process reaches a terminal state s_T when the policy π_θ emits the terminal action $a_t =$ finish. Alternatively, we enforce that the environment halts the policy if the step t exceeds a maximum execution step T_{max}, saving the final manuscript and the trajectory $\mathcal{T}$. The orchestration of agent logic within **RAVEL** is described in Algorithm 1.

4 C3EBench: Benchmarking the Full Synthesis Lifecycle

To complement our evaluation framework, we introduce C3EBench, a comprehensive benchmark constructed to isolate and evaluate the distinct capabilities required for agentic text synthesis. Rather than treating generation as a monolithic task, C3EBench characterizes the synthesizing process with four distinct scenarios. In order to represent the high-quality demand of actual text synthesis demands, we use processed high-quality professional human writing as references and reverse-construct the instruction queries. C3EBench evaluates LLMs across varying levels of granularity, from end-to-end generation to specific editing capabilities, bridging the gap in task coverage during evaluation.

4.1 Task Formulation

We formalize text synthesis as $W = f(\mathcal{I}, \mathcal{G})$, where model f generates text W conditioned on instruction $\mathcal{I}$ and optional grounding inputs $\mathcal{G}$. We design one coarse-grained task (**End2End**) and three fine-grained subtasks (**Expand**, **Cloze**, **Edit**). Task formats are in Figure 3, statistics in Table 1, and examples in Table 6.

End2End Generation. This task evaluates the LLM's ability to synthesize coherent text from scratch. Given a high-level instruction $\mathcal{I}$ (specifying theme, style, and constraints), the model must generate a complete text W that satisfies all requirements without additional grounding. This tests the model's capacity for long-form text synthesis with autonomous planning and generation.

Expand (Outline-Conditioned Generation). This task assesses the ability to adhere to structural and outline constraints. The LLM is provided with an instruction $\mathcal{I}$ and a detailed outline $\mathcal{G}_{\text{outline}}$. The objective is to synthesize the full text W by expanding upon the provided outline points, requiring precise control over content logic.

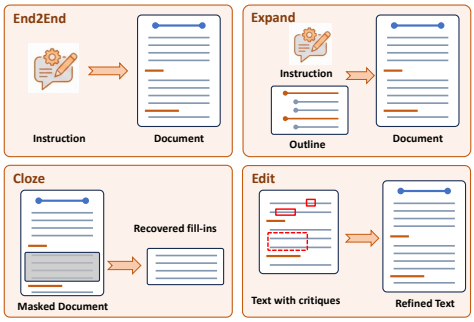


Figure 3: Overview of the task formulations.

Cloze (Contextual Infilling). This task tests local coherence and context awareness. Selected spans within a reference text are masked. Given the surrounding context $\mathcal{G}_{\text{context}}$ and the mask position, the LLM must reconstruct the missing content. Unlike standard pre-training objectives, our masks target semantically dense segments (e.g., rhetorical narrations or transitional clauses) rather than random tokens.

Edit (Feedback-Driven Revision). This task evaluates the ability to refine text based on critiques. The input consists of a draft text $\mathcal{G}_{\text{draft}}$ and specific revision instructions $\mathcal{I}_{\text{feedback}}$. The model must produce a polished version W that addresses the critique while maintaining the original intent. This mimics the human iterative writing process.

4.2 Data Construction Pipeline

Constructing a high-quality benchmark is challenging due to the difficulty of obtaining paired instructions and professional references. To address this, we adopt a *Reverse-Construction* paradigm (Thorne et al., 2018): we source high-quality references $\mathcal{R}$ first, then annotate the corresponding instructions $\mathcal{I}$ and inputs $\mathcal{G}$ with the help of automated LLM tools. The pipeline consists of 3 stages:

Table 1: Statistics of the Chinese (CN) and English (EN) datasets. Lengths are measured in tokens/characters.

Task	Lang.	Instruction			Input		Reference		Total
		Avg.	Max.	Min.	Avg.	Max.	Avg.	Max.	
Cloze	CN	44.40	60	33	2037.44	4795	422.04	1458	200
	EN	19.00	19	19	951.89	5179	408.55	2412	150
Expand	CN	109.94	183	64	359.57	750	1779.80	3894	200
	EN	246.72	387	141	239.23	342	1546.07	4863	99
Edit	CN	178.00	178	178	2434.84	4338	982.32	2159	200
	EN	85.00	85	85	405.72	1229	419.47	1904	110
End2E	CN	86.62	171	33	0.00	0	1856.66	3967	200
	EN	67.17	86	43	0.00	0	1624.09	5744	99

Stage 1: Sourcing and Filtering. We curated a corpus of professional writings from various specialized sources (fiction, essays, reports, contracts). To ensure reference quality ($\mathcal{R}$), we applied a strict two-step verification: 1. **Genre Classification:** GPT-5.2 classified texts by genre, with human verification correcting errors (1.4% error rate). 2. **Quality Filtering:** Using Claude-4.5-Sonnet with genre-specific rubrics (Appendix G.3), we discarded texts scoring below 4 out of 5. Figure 6 shows the taxonomy and distribution.

Stage 2: Instruction Construction via Reverse-Engineering. We employ specific strategies to reverse-engineer task inputs from the verified reference $\mathcal{R}$. Human annotators construct with the help of a pre-processing LLM. ***End2End & Expand:*** We model the construction as $(\mathcal{I}, \mathcal{G}) \leftarrow \text{BackConstruct}(\mathcal{R})$. A prompted LLM summarizes the content (S), extracts the topic (T), and compiles genre-specific constraints(G). For ***End2End***, $\mathcal{I}$ is derived from the topic T and genre constraints G. For ***Expand***, $\mathcal{G}$ incorporates the detailed outline S. ***Cloze:*** We identify spans in $\mathcal{R}$ that are crucial to the narrative flow. These spans are masked to create the input, ensuring the task requires semantic deduction rather than simple pattern matching. ***Edit:*** To create realistic revision scenarios, we first generate a draft with LLM (via the Cloze task) and compare it against the professional reference $\mathcal{R}$. Experts analyze the gap between the draft and $\mathcal{R}$ to craft specific critique instructions $\mathcal{I}_{\text{feedback}}$. This ensures the feedback creates a trajectory from the draft towards the human reference. The prompts for the assisting LLM (Gemini-3-Pro) are in Appendix C.1, C.2, C.3 and C.4. The reference instruction templates for human are in Appendix C.5.

Stage 3: Quality Assurance & Reference Optimization. We implemented a human-in-the-loop review to validate the $(\mathcal{I}, \mathcal{G}, \mathcal{R})$ triplets. Domain experts verified the naturalness of generated instructions. Crucially, we employed a **Best-of-N Reference Optimization**. We recognize that while our human sources are professional, SOTA LLMs can occasionally yield superior specific

outputs. We generated responses using GPT-4o, GLM-4.5-plus, and Gemini-2.5-Flash and performed blind pairwise comparisons against the human reference. We dropped 10.5% of cases where models outperformed the human text. Finally, we performed safety checks using GPT-5.2, followed by human verification to remove PII and sensitive content. Appendix D shows the guideline for human experts. We further verify that C3EBENCH is free from **pre-training contamination** via perplexity, Min-$k\%$ Prob, and greedy recall analyses (Appendix E).

Evaluation Setup. We use an LLM-as-a-judge approach (Zheng et al., 2023) for single-round evaluation. To align with human judgment, we develop prompts for four tasks (Appendix H).

5 EXPERIMENTS

We evaluate 14 LLMs on C3EBENCH and within **RAVEL**, examining writing quality, autonomous planning and refinement, and the relative contributions of reasoning and generation.

5.1 EXPERIMENTAL SETUP

We evaluate proprietary and open-source models with identical generation settings (temperature 0.7). **RAVEL** uses C3EBENCH End2End with $T_{max} = 50$, a shared, uncalibrated $\tau = 8.0$, and external Judge GPT-5.2-1211. Main results use each tested model as its own policy; external policy/refiner substitutions occur only in designated ablations (Section 5.4, Appendix K). Appendix J gives prompts; Appendix L examines self-scoring and protocol sensitivity.

5.2 AGENTIC EVALUATION METRICS

To rigorously quantify the reasoning efficacy and generation quality within **RAVEL**, we define four metrics. They capture task completion, structural overhead, and self-refinement loop effectiveness.

Task Success Rate ($\mathcal{S}$). Measures the agent's ability to converge within the budget T_{max}. A trial is successful if the policy π_θ triggers the terminal action "finish". For N samples:

$$\mathcal{S} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left(\exists t \leq T_{\text{max}} : a_t^{(j)} = \text{finish}\right) \tag{8}$$

Trajectory Efficiency (η_{traj}). Evaluates reasoning efficiency by calculating the ratio of total actions $|\mathcal{T}|$ to the planned outline size $n = |\mathcal{O}|$. This reflects the execution overhead: $\eta_{\text{traj}} = \frac{|\mathcal{T}|}{n}$. A theoretical lower bound of $\eta_{\text{traj}} \approx 2$ indicates optimal execution (draft $\rightarrow$ pass review). Higher values imply intensive revision cycles.

Refinement Density (ρ_{ref}). Quantifies the intensity of self-correction relative to the manuscript scale:

$$\rho_{\text{ref}} = \frac{1}{n} \sum_{t=1}^{|\mathcal{T}|} \mathbb{I}(a_t = \text{review}) \tag{9}$$

High ρ_{ref} suggests the LLM requires multiple iterations of scrutiny to meet the threshold τ.

Refinement Delta (δ_{gain}). Quantifies the magnitude of improvement achieved via the iterative loop:

$$\delta_{\text{gain}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} (r'_k - r_k) \tag{10}$$

A positive δ_{gain} indicates increased self-review scores; final quality is assessed by the external Judge.

Auxiliary statistics are **Action Failure Rate** (ERR_{π_θ}), the fraction of invalid action attempts; **Outline Length** ($|\mathcal{O}|$), the mean planned node count; and **Trajectory Length** ($|\mathcal{T}|$), the mean steps per episode.

5.3 MAIN RESULTS AND ANALYSIS

Table 2 presents the comprehensive benchmarking results. Our analysis reveals distinct behavioral patterns across model families, highlighting the gap between raw generation capability and agentic reasoning proficiency during text-synthesis.

Table 2: Main benchmarking results on C3EBENCH and **RAVEL**. Metrics are grouped into **Task Capabilities**, **Agentic Dynamics**, and **Overall Efficiency**. **Bold** indicates the best per category.

Model	C3EBENCH Tasks (Score ↑)				Agentic Dynamics					Execution Efficiency		
	Cloze	Expand	Edit	E2E	$\mathcal{S}\%$	η_{traj} % ↓	$\rho_{\text{ref}}\%$	δ_{gain}	Judge	ERR_{π_θ} % ↓	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{T}\rvert$
Proprietary LLMs												
GPT-5.2-2025-11-20 (Singh et al., 2025)	**4.53**	**7.89**	7.50	7.37	64.3	2.35	17.9	0.00	6.64	0.95	15.5	30.2
Gemini-3 Pro (Gemini Team, Google, 2025)	4.44	7.79	7.18	7.51	**95.1**	**2.32**	2.40	**27.96**	**7.05**	**0.00**	7.1	16.2
Claude-4.5 Sonnet (Anthropic, 2025)	4.44	7.79	7.42	7.52	68.1	2.95	100.5	19.55	6.68	**0.00**	10.3	28.0
Grok-4 (xAI, 2025)	4.42	7.76	7.12	7.39	69.0	2.50	38.2	3.52	6.31	0.03	7.4	18.4
GLM-4.7 (GLM-Team, 2025)	2.62	7.25	5.43	7.17	73.3	2.51	38.5	0.00	6.31	0.37	7.3	18.0
Qwen3-Max (Yang et al., 2025)	4.32	7.86	**7.71**	7.35	47.8	**1.92**	23.9	0.00	5.74	**0.00**	7.3	13.1
Kimi-K2-Thinking (Team et al., 2025)	3.99	7.75	7.05	**7.71**	58.3	2.81	55.4	0.00	6.48	1.32	7.9	21.6
DeepSeek-v3.2 (DeepSeek-AI et al., 2025)	4.22	7.33	7.21	7.45	**95.7**	2.59	26.2	2.18	6.64	0.02	7.0	17.9
Open Source LLMs												
Llama-3.1-405B (Grattafiori et al., 2024)	3.09	3.64	5.94	3.51	47.7	3.94	30.3	1.09	5.05	0.00	7.5	29.6
Hermes-3-Llama-70B (Grattafiori et al., 2024)	1.88	5.55	4.19	5.28	91.8	3.16	18.7	**11.39**	4.58	0.00	7.7	24.1
Llama-3.1-8B (Grattafiori et al., 2024)	2.44	5.89	5.85	5.33	0.0	8.59	18.3	4.60	1.66	0.04	5.8	51.0
Qwen3-235B-A22B (Yang et al., 2025)	**4.60**	**7.49**	**7.67**	**7.38**	89.4	**2.47**	20.5	0.00	**6.94**	**0.00**	8.3	20.1
Qwen3-32B (Yang et al., 2025)	3.74	6.56	7.05	6.07	58.5	3.02	105.8	5.54	5.54	0.00	7.2	21.2
Qwen3-8B (Yang et al., 2025)	3.14	6.14	6.99	5.65	**94.7**	2.43	**13.9**	2.57	5.51	0.02	7.5	**18.0**

Capability. LLMs perform poorly in under-specified tasks. C3EBENCH reveals a key vulnerability: reliance on explicit instruction scaffolding. All models exhibit significant degradation in CLOZE, with even the strongest proprietary models failing to exceed 5.0—here, LLMs receive only surrounding context without explicit requirements. A consistent gap exists between EXPAND and END2END, particularly in open-source models (Qwen3-235B-A22B: 7.49 vs. 7.38). While LLMs excel at following detailed instructions, they lack autonomous text synthesis capability, identifying a critical frontier for future work.

Proprietary models establish the state-of-the-art. GPT-5.2 achieves the highest Cloze (4.53) and Expand (7.89) scores. **Gemini-3 Pro**, while slightly trailing in raw scores, achieves the highest execution robustness ($\mathcal{S} = 95.1\%$), significantly outperforming GPT-5.2 (64.3%). **Open-source models are closing the gap**: **Qwen3-235B-A22B** matches or surpasses proprietary models in specific metrics (Judge score 6.94). However, **Llama-3.1-8B** fails completely ($\mathcal{S} = 0.0\%$), indicating that models below a certain scale lack the planning capability for the **RAVEL** workflow.

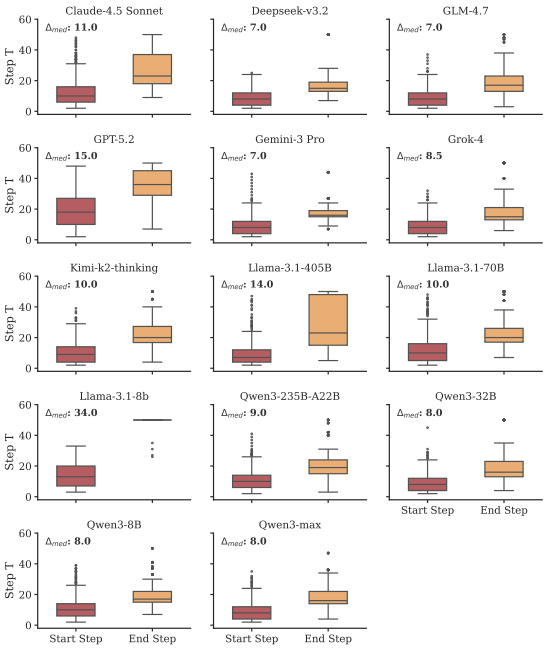


Figure 4: **Paragraph Lifecycle Analysis.** Distribution of *Start Step* vs. *End Step*. GPT-5.2 exhibits **sequential drafting**; Gemini and DeepSeek adopt a **"batch" approach**.

Dynamics. The **Agentic Dynamics** metrics distinguish refinement *frequency* from *self-assessed gains*:

***A. Refinement Frequency and Gain*: Gemini-3 Pro** has low $\rho_{\text{ref}} = 2.4\%$ but high $\delta_{\text{gain}} = 27.96$, whereas **Claude-4.5** and **Qwen3-32B** exhibit $\rho_{\text{ref}} > 100\%$ with moderate self-assessed gains. Frequent refinement therefore need not imply larger review-score gains, and neither metric alone establishes external quality improvement.

***B. Planning Horizon Length*: GPT-5.2** and **Claude-4.5** generate long outlines ($|\mathcal{O}| > 10$), increasing the probability of exceeding T_{max} and lowering $\mathcal{S}$ (64.3% for GPT-5.2). **Gemini-3 Pro** adopts concise planning, achieving superior robustness ($\mathcal{S} = 95.1\%$). This suggests "shorter" planning is more token-efficient and less prone to agentic drift.

Strategic Differences in Synthesis Trajectories. We analyze the lifecycle of paragraph generation across 14 LLM trajectories. As illustrated in Figure 4, we identify two paradigms. ***Interleaved***

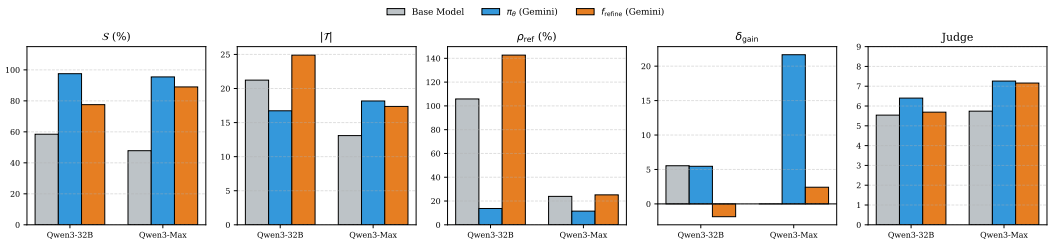


Figure 5: Sensitivity analysis of reasoning vs. generation capabilities within **RAVEL**.

Sequential Synthesis (ISS): Exemplified by **GPT-5.2** and **Claude-4.5**, these models exhibit a "plan-write-review-refine" loop per section, delaying subsequent sections until the current one satisfies τ. ***Parallelized Batch Synthesis (PBS):*** **Gemini-3 Pro** and **DeepSeek-v3.2** adopt a "write-all-then-refine" strategy with $T_{\text{start}} \approx 0$ for all paragraphs. While both achieve comparable quality, **PBS** generally demonstrates superior efficiency by avoiding the "refinement loops" that trap **ISS** strategies.

Linking C3EBENCH and RAVEL. Across 13 models with paired results, static task scores correlate positively with **RAVEL**'s final End2End Judge score (Table 3). Trajectory-metric associations were not significant; this small sample does not establish independence. These results link the evaluations on output quality, while **RAVEL** additionally characterizes behavior.

Table 3: Task–Judge Spearman correlations ($n = 13$ models).

Cloze	Expand	Edit	End2End
0.76	0.66	0.73	0.69

5.4 REASON & GENERATE: WHO TAKES THE DECISIVE ROLE?

To investigate the "review-refine deadlock," we use **Qwen3-32B** and **Qwen3-Max** as backbones and replace either the policy (π_θ) or refiner (f_{refine}) with Gemini-3 Pro. These controlled substitutions distinguish the contributions of action selection and local revision.

The results reveal a critical insight: **reasoning dominates generation in complex text synthesis.** Integrating a stronger reasoner (π_θ=Gemini3-Pro) significantly increases synthesis success rates ($\mathcal{S} > 95\%$) across both backbones. Notably, for Qwen3-32B, the superior reasoner drastically suppresses the refinement frequency (ρ_{ref}), suggesting that optimal reasoning over next-step actions effectively obviates the need for iterative review-refine loops. Counter-intuitively, simply enhancing the generator (f_{refine}=Gemini3-Pro) does not guarantee quality improvements, though it does help optimize task success rate. For instance, applying the Gemini refiner to Qwen3-32B resulted in a negative review-score gain ($\delta_{\text{gain}} = -1.86$), whereas the Gemini-driven reasoner yielded a positive gain for Qwen3-Max ($\delta_{\text{gain}} = 21.66$). This confirms that high-quality synthesis stems more from structural reasoning than from local generation. An extended cross-model validation with additional base models and augmentors is provided in Appendix K.

Self-scoring and protocol sensitivity. Self-review scores exceed external judgments, while varying τ produces small quality changes on the tested subset (Appendices L.1–L.2). Protocol ablations preserve model ranking but raise Qwen3-Max's Judge from 5.92 to 7.08 under a fixed schedule (Appendix L.3). Thus, protocol choice affects absolute quality as well as completion; behavioral metrics should be interpreted alongside external quality.

5.5 META EVALUATION OF LLM-AS-A-JUDGE

We validate the judge against human preferences using 200 randomly selected instructions (50 per task, with even genre coverage) and nine model outputs per instruction from Table 2. Thirty-six writing experts rated outputs on a 1–5 scale after guideline training (Appendices I.1, I.5); agreement statistics appear in Appendix I.2. We pair human-rated outputs and measure consistency between human and judge preferences. Ablating rubrics, traits, and references from the protocol in Section 4.1 isolates their contributions (Table 4).

Table 4: Ablation on LLM-as-a-judge setting.

SETTING	CLOZE	EXPAND	EDIT	END2END
CURRENT	**88.0**	**88.0**	**90.0**	**92.0**
– RUBRIC	76.0	78.0	74.0	72.0
– TRAITS	78.0	82.0	78.0	76.0
– REFERENCE	72.0	64.0	64.0	56.0

Results in Table 4 reveal that the existence of high quality reference is the most crucial part in the evaluation performance (+36% in END2END). The rubric and trait design are also important designs, improving performance from (+11% to +13.5%). Conclusively, the current LLM judge configuration validates the alignment of our evaluation framework. A comprehensive comparison against alternative evaluation methods (overlap metrics, self-consistency, self-planned criteria) is provided in Appendix I.3.

6 RELATED WORK

6.1 BENCHMARKING LLM WRITING

Prior research on evaluating LLM-generated writing has primarily focused on creative story generation, emphasizing fluency and coherence through datasets like RocStories (Mostafazadeh et al., 2016) and metrics like OpenMEVA (Guan et al., 2021). While these works highlight narrative quality, their scope is constrained to specific genres (e.g., fiction) and narrow evaluation dimensions. Recent benchmarks for general text generation emphasize instruction-following (Zheng et al., 2023; Liu et al., 2024), lexical quality and coherence (Zhang et al., 2024a;b), as well as domain expertise (Liang et al., 2023). However, they inadequately address the open-ended nature of writing tasks. Very recent work Wu et al. (2025) seeks an instruction-following way for writing evaluation. For instance, reference-based metrics (Deutsch et al., 2022) prioritize structural conformity over creative divergence, while existing LLM-as-a-judge methods face challenges in capturing genre-specific stylistic subtleties. C3EBENCH advances these researches by expanding evaluation to **12** diverse genres beyond story generation, and explicitly addressing the complex actual scenarios rather than only leaving the text synthesis as a black-box process.

6.2 LLM-BASED EVALUATION

Recent advances in LLM-based evaluation utilize proprietary models for automated scoring through prompt engineering (Zheng et al., 2023; Liu et al., 2023) or supervised training on human annotations (Wang et al., 2024b; Ke et al., 2024). These methods surpass traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) in efficiency and alignment with human correlation, particularly for constrained tasks such as summarization. However, their reliability weakens in the context of open-ended writing evaluation: verbosity bias (Zheng et al., 2023), positional bias (Wang et al., 2024a), and rubric dependency (Ke et al., 2024; Kim et al., 2024a) hinder their generalizability. In contrast, attempts (Wu et al., 2025) that involve LLMs autonomously generating evaluation criteria and rubrics emerged, but their robustness remain unexamined.

7 CONCLUSION

This work introduces **RAVEL** and C3EBENCH to shift text synthesis evaluation from static outputs to dynamic agentic processes. Our empirical analysis demonstrates that synthesis success stems primarily from reasoning efficacy (planning and critiquing) rather than raw generative capacity (editing or drafting). We also identify a significant bottleneck in autonomous planning, where models excel at local drafting but struggle with high-level structural coherence without explicit scaffolding. These findings suggest that the path toward true intent-driven composition lies not in scaling generation but in developing robust autonomous reasoning capabilities.

ETHICS STATEMENT

Human participation and ethics review. The study was reviewed and approved by a university Institutional Review Board. Annotators were recruited primarily from freelance writers and graduate students in the humanities, social sciences, and journalism. They participated voluntarily after being informed that the annotation tasks supported research on generative AI, and were compensated at USD 10 per hour. Further details are provided in Appendix I.1.

Data use. Use of the released benchmark is restricted to research on the evaluation of generative AI. Data sources and the filtering and review procedures are described in Appendix G.1 and Section 4.2, respectively.

Potential harms and safeguards. AI-generated text may contaminate web data if circulated without clear attribution, potentially affecting subsequent model training and evaluation. To help identify such content, **RAVEL** appends an explicit AI-generated notice to its outputs after generation. This notice makes their origin visible but does not prevent removal of the notice or subsequent redistribution. Further risks and limitations are discussed in Appendices B and A.

AI USE STATEMENT

During data production and evaluation, generative AI was primarily used to assist with the preliminary filtering and classification of raw data, and as an automated judge to provide feedback for a subset of our automated evaluation metrics.

During manuscript and artifact preparation, generative AI was primarily used for typesetting, formatting checks, and organizing and releasing the accompanying code repository.

We have reviewed and verified the AI-assisted work described above and take responsibility for the final content of the paper and accompanying materials.

REFERENCES

Anthropic. Claude 4.5 system card: Reasoning, coding, and agentic capabilities. Technical report, Anthropic PBC, 2025. URL https://www.anthropic.com/claude-sonnet-4-5-system-card.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context llms, 2024. URL https://arxiv.org/abs/2408.07055.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang

Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Daniel Deutsch, Rotem Dror, and Dan Roth. On the limitations of reference-free evaluations of generated text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10960–10977, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.753. URL https://aclanthology.org/2022.emnlp-main.753/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents, 2025. URL https://arxiv.org/abs/2506.11763.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023. URL https://arxiv.org/abs/2302.04166.

Gemini Team, Google. Gemini 3 pro system card. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

GLM-Team. Glm-4.7: Advancing the coding capability. Technical report, ZhipuAI, 2025. URL https://z.ai/blog/glm-4.7.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,

Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle

Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 6394–6407, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.500. URL https://aclanthology.org/2021.acl-long.500/.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13034–13054, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.704. URL https://aclanthology.org/2024.acl-long.704/.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=8euJaTveKw.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models, 2024b. URL https://arxiv.org/abs/2405.01535.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li,

Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. Criticbench: Benchmarking llms for critique-correct reasoning, 2024. URL https://arxiv.org/abs/2402.14809.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. AlignBench: Benchmarking Chinese alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11621–11640, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.624. URL https://aclanthology.org/2024.acl-long.624/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. Plotmachines: Outline-conditioned generation with dynamic plot state tracking, 2020. URL https://arxiv.org/abs/2004.14967.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model, 2022. URL https://arxiv.org/abs/2208.11663.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua

Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Together Computer. RedPajama: An open source recipe to reproduce LLaMA training dataset, April 2023. URL https://github.com/togethercomputer/RedPajama-Data.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=5Nn2BLV7SB.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Cunxiang Wang, Zhen Wu, Xinyu Dai, Yue Zhang, Wei Ye, and Shikun Zhang. Trustjudge: Inconsistencies of llm-as-a-judge and how to alleviate them, 2025. URL https://arxiv.org/abs/2509.21117.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing, 2025. URL https://arxiv.org/abs/2503.05244.

xAI. Grok 4.1 model card: Scaling reinforcement learning for reasoning agents. Technical report, xAI, 2025. URL https://x.ai/blog/grok-4.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision, 2022. URL https://arxiv.org/abs/2210.06774.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Xuanming Zhang, Zixun Chen, and Zhou Yu. ProLex: A benchmark for language proficiency-oriented lexical substitution. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8475–8493, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.502. URL https://aclanthology.org/2024.findings-acl.502/.

Xuanming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, and Zhou Yu. DECOR: Improving coherence in L2 English writing with a novel benchmark for incoherence detection, reasoning, and rewriting. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11436–11458, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.639. URL https://aclanthology.org/2024.emnlp-main.639/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/

91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

APPENDIX CONTENTS

Section			Page
A	**Limitations**		**20**
B	**Broader Impacts**		**20**
C	**Data Construction Details**		**21**
	C.1	END2END Instruction Prompts	21
	C.2	CONDITION Instruction Prompts	21
	C.3	CLOZE Instruction Prompts	22
	C.4	EDIT Instruction Prompts	23
	C.5	Instruction Templates for Human Annotators	23
D	**Human-in-the-Loop Guideline**		**23**
	D.1	Task Description	23
	D.2	Annotation Fields	24
	D.3	Evaluation Criteria	24
E	**Contamination Analysis**		**25**
F	**Data Examples**		**27**
G	**Data Quality**		**31**
	G.1	Crawling Sources	31
	G.2	Source Text Genre Distribution	32
	G.3	Prompts for Source Text Filtering	32
H	**Evaluation Prompts**		**34**
	H.1	END2END Evaluation Prompts	34
	H.2	CLOZE Evaluation Prompts	35
	H.3	EXPAND Evaluation Prompts	35
	H.4	EDIT Evaluation Prompts	36
I	**Meta Evaluation Details**		**36**
	I.1	Human Annotator Details	36
	I.2	Additional Annotation Details	36
	I.3	Extended Meta-Evaluation: Method Selection	37
	I.4	Absolute Score Reliability	38
	I.5	Completion Annotation Guidance	39
J	**Prompts for RAVEL Implementation**		**41**
K	**Extended Analysis: Reasoning vs. Generation**		**43**
L	**Self-Scoring and Protocol Sensitivity**		**44**
	L.1	Self-Review versus External Judgment	44
	L.2	Sensitivity to the Stopping Threshold	45
	L.3	Action-Protocol Ablation	45
M	**Reproducibility Statement**		**46**

A LIMITATIONS

Our study presents two primary limitations regarding benchmark language coverage and generalizability.

First, C3EBENCH is currently constructed exclusively for English and Chinese. While these languages represent distinct linguistic typologies, the absence of low-resource languages limits our ability to generalize findings across broader linguistic families.

Second, the **RAVEL** action space $\mathcal{A}$ is deliberately restricted to "close-book" synthesis operations (planning, drafting, critiquing, reviewing), excluding external tool use such as web search. This design choice was made to isolate the LLM's internal reasoning and compositional coherence from confounding variables like retriever performance or factual hallucination. Consequently, our evaluation characterizes "closed-book" synthesis capabilities; extending this framework to "open-book" settings involving retrieval-augmented generation or multi-modal integration features is another distinct challenge for future research.

Finally, we employ a standard single-pass LLM-as-a-judge rather than complex multi-agent judging frameworks. This decision is pragmatic: our single-pass judge demonstrates high human alignment ($\approx 90\%$), providing sufficient discriminatory power while avoiding the prohibitive computational costs of iterative judging. This efficiency is critical for scaling evaluation to the super-long text synthesis scenarios.

B BROADER IMPACTS

Positive Impacts. **RAVEL** and C3EBENCH contribute to a more transparent and fine-grained understanding of LLM text synthesis capabilities, which benefits both the research community and downstream practitioners. By decomposing monolithic generation into observable agentic operations (outlining, drafting, reviewing, refining), our framework enables developers to diagnose specific capability bottlenecks rather than relying on opaque aggregate scores. This can guide more targeted and efficient model improvement, reducing wasteful scaling efforts. Furthermore, the reverse-construction methodology and human-in-the-loop quality assurance pipeline provide a replicable paradigm for building high-quality, reference-grounded benchmarks, raising the standard for evaluation rigor in the community. Our finding that reasoning capability dominates raw generation capacity offers actionable insight for LLM production: investing in planning and self-critique mechanisms may yield greater returns than scaling generation alone.

Potential Negative Impacts. As an evaluation framework for text synthesis, this work could indirectly accelerate the development of more capable text generators. While our benchmark is designed for diagnostic evaluation, the insights it reveals (e.g., effective refinement strategies, optimal planning horizons) could be leveraged to improve automated content generation systems, which may be misused for producing disinformation, academic dishonesty, or other deceptive content at scale. Additionally, the benchmark data is sourced from crawled professional writings; despite our safety filtering and PII removal procedures (Section 4.2), residual risks of reproducing copyrighted material or culturally sensitive content may persist. We encourage users of C3EBENCH and **RAVEL** to adhere to responsible use guidelines and to deploy the insights from this work toward improving content authenticity detection alongside generation quality.

Ethics approval. The study was reviewed and approved by a university Institutional Review Board. Further details on human participation and safeguards are provided in the Ethics Statement.

C Data Construction Details

C.1 Prompts for constructing END2END task instruction

END2END task instruction: Reverse Instruction Engineering

"You are an expert Literary Editor. Your task is to identify and extract the most stylistically brilliant and contextually significant continuous segment from a given text."

Role: You are an expert Prompt Engineer and Creative Writing Specialist. Your task is to perform "Reverse Instruction Engineering" on a piece of high-quality human text.

Input: A text snippet from a JSONL file (field: `content`).
Genres: The text will belong to one of these categories: Essay (academic/analytical), Story (short story), Fiction (novel chapter), or Speech (political/historical).

Task: Analyze the provided `content` and design a writing instruction (the "query") that would guide an LLM to reproduce the original text as closely as possible in terms of theme, tone, and organization.

Output Requirements:
You must return a JSON object with the following five fields:

- `"genre"`: Identify the specific writing genre.
- `"brief"`: A concise summary of the central theme or main idea.
- `"audience"`: Define the target audience and the required tone/register.
- `"word"`: An approximate word count requirement based on the original text.
- `"query"`: The final, comprehensive writing instruction. **This field must not exceed 100 words.**

JSON Schema:

```
{
  "genre": "...",
  "brief": "...",
  "audience": "...",
  "word": "...",
  "query": "..."
}
```

Constraint: Do not include any introductory or concluding remarks. Return ONLY the raw JSON object.

Input Content to Analyze:
`[TEXT_CONTENT]`

C.2 Prompts for constructing CONDITION task instruction

CONDITION task instruction: Reverse Instruction Engineering

"You are an expert Literary Editor. Your task is to identify and extract the most stylistically brilliant and contextually significant continuous segment from a given text."

Role: You are an expert Prompt Engineer and Creative Writing Specialist. Your task is to perform "Reverse Instruction Engineering" on a piece of high-quality human text.

Input: A text snippet from a JSONL file (field: `content`).
Genres: The text will belong to one of these categories: Essay (academic/analytical), Story (short story), Fiction (novel chapter), or Speech (political/historical).

Task: Analyze the provided `content` and design a writing instruction (the "query") that would guide an LLM to reproduce the original text as closely as possible in terms of theme, tone, and organization.

Output Requirements:
You must return a JSON object with the following six fields:

1. `"genre"`: Identify the specific writing genre.
2. `"brief"`: A concise summary of the central theme or main idea.
3. `"structure"`: Specific requirements for the flow, outline, or narrative arc.
4. `"audience"`: Define the target audience and the required tone/register.
5. `"word"`: An approximate word count requirement based on the original text.

6. `"query"`: The final, comprehensive writing instruction. **This field must not exceed 300 words.** It should integrate the genre, brief, structure, and audience.

JSON Schema:

```
{
  "genre": "...",
  "brief": "...",
  "structure": "...",
  "audience": "...",
  "word": "...",
  "query": "..."
}
```

Constraint: Do not include any introductory or concluding remarks. Return ONLY the raw JSON object.

Input Content to Analyze:
`[TEXT_CONTENT]`

C.3 Prompts for constructing CLOZE task instruction

CLOZE task instruction: Segment Extraction

"You are an expert Literary Editor. Your task is to identify and extract the most stylistically brilliant and contextually significant continuous segment from a given text."

Role: You are an expert Literary Editor and Writing Coach with a deep understanding of rhetoric, narrative structure, and advanced linguistics.

Task: Analyze the provided text (essay, story, fiction chapter, or speech). Your goal is to identify and extract a "Golden Segment"—a continuous block of text that represents the pinnacle of writing quality within the piece.

Selection Criteria:

1. **Length:** The extracted segment must constitute between 20% and 50% of the total length of the original text.
2. **Quality:** Features sophisticated vocabulary, evocative imagery, unique rhetorical devices, or unexpected narrative turns.
3. **Contextual Significance:** Must be integral to the flow, showing how the author builds an argument or advances a plot.
4. **Continuity:** Must be one unbroken, continuous passage.

Constraints:

- Do **NOT** paraphrase. The extracted text and sentences must be verbatim.
- Ensure the `"start_sentence"` and `"end_sentence"` are complete, recognizable sentences.

Output Format (JSON):

```
{
  "start_sentence": "...",
  "end_sentence": "...",
  "selected_text": "..."
}
```

Input Content:
`[TEXT_CONTENT]`

C.4 Prompts for constructing EDIT task instruction

EDIT task instruction: Comparative Draft Review & Revision

Task: I have a draft snippet (AI-generated) here, and I would like you to perform an in-depth review. To help you provide more precise improvement suggestions, I will provide an "Ideal Reference" (original human writing) as the goal. Please compare the two, identify the issues in the draft, and provide revision instructions.

Note: Your final output must be revision suggestions addressed to the author. You are **strictly prohibited** from mentioning terms like "human original," "reference sample," "comparison," or "original author" in your output. You must act as if you are reviewing the draft directly and guiding the revisions toward the desired outcome by identifying its flaws.

Input Data:

- Draft Content (to be reviewed): `{ai_content}`
- Ideal Reference (for internal comparison only): `{human_middle}`

Writing Requirements:

1. **Problem Diagnosis:** Identify specific issues in the draft regarding lexical expressiveness, logical tension, emotional insight, etc.
2. **Revision Advice:** Provide specific directions for rewriting. You may draw upon the elements of the "Ideal Reference" to guide the transformation of the draft.
3. **Tone:** Professional, sharp, and inspiring.
4. **Length:** The combined length of the Problem Diagnosis and Revision Advice should not exceed 300 words.

Output Format:
Please begin your response directly starting with: **Problem Diagnosis & Revision Suggestions**

C.5 Templates of instruction construction for human

Human annotators are instructed to construct the instruction based on the following template. They are not instructed to hard copy but rather containing the necessary information in the template.

Instruction Template for Cloze

Input: Genre
Please fill in the blanks in the following {genre}, marked with [fill in the blank] signs. You should comprehensively consider the context and ensure the completion quality.

Instruction Template in End2End/Expand

Inputs: Genre,Topic,Summary,Word counts
Please write a {genre} about {Topic}. {summary}. You should write in approximately {word counts}.

D Human-in-the-Loop Guideline

D.1 Task Description

Your task is to evaluate and compare four different writings based on a provided writing instruction. Each writing is a response to the same instruction, and your goal is to pick the one that fits the instruction with the highest quality. Use the evaluation criteria provided below to make your judgment. The selected writing should be the one that most effectively fulfills the writing instruction and demonstrates the highest level of quality across both content and format.

D.2 Annotation Fields

D.2.1 Visible Inputs

- **Writing Instruction** : A clear description of the requirements or objectives for the writing task (e.g., structure, tone, purpose, or audience).

- **Guiding Information** : If applicable, specific details that the writings are expected to follow (e.g., key points, required examples, or constraints). For tasks requiring "guide generation," ensure the writings strictly adhere to these details.

- **Writing 1/2/3/4** : The individual LLM writings submitted for judging.

D.2.2 Your Observations

- Write down notes on how each writing satisfies the instruction and aligns with the evaluation criteria.

- Highlight specific strengths and weaknesses of each writing that influenced your judgment.

D.2.3 Annotation Process

Step 1: Read Each Writing Thoroughly

- Carefully read each writing submission. - Pay attention to how well the author has addressed the writing instruction and incorporated the guiding information provided. - Consider the quality of the arguments, organization, and style of each piece. Make sure to read thoroughly before forming a judgment.

Step 2: Apply the Quality Criteria

- Systematically assess each writing response against the evaluation criteria outlined below. - Use both content and format criteria to conduct your evaluation and determine the strengths and weaknesses of each submission. - You may apply a pointwise scoring system (e.g., rating each category from 1 to 5) to help you compare the writings more quantitatively. These scores should support — but not replace — your final judgment.

Step 3: Select the Best Writing

- Based on your evaluation in Step 2, determine which writing best fulfills the writing instruction and meets the specified quality criteria. - Document your reasoning for selecting the chosen writing. Highlight why the selected piece was superior and what weaknesses were present in the others.

D.3 Evaluation Criteria

Your evaluation should be based on two main areas: Content and Format . Each area contains specific criteria to guide your assessment:

D.3.1 Content

1. Theme/Argument/Topic Fit :

- How well does the writing address the objective of the instructions?

- Are the arguments or ideas relevant and clearly aligned with the given topic?

- Does the writing stay focused, or does it go off-topic?

2. Tone and Language :

- Is the tone appropriate for the audience and purpose outlined in the writing instruction?

- Does the writing use clear, engaging, and professional language where required?

- Is the tone consistent throughout the piece?

3. Attractiveness of Opening and Profound Ending :

- Does the writing start with a strong and engaging opening that catches the reader's attention?

- Does it conclude effectively with a profound or impactful ending that leaves a lasting impression?

4. Rhetoric, Logic, and Examples :

- Does the writing employ effective rhetoric (e.g., persuasive techniques, vivid imagery, or strong analogies)?

- Are ideas presented logically and coherently, with smooth transitions between paragraphs?

- Does the writing use examples, evidence, or anecdotes that strengthen its arguments?

D.3.2 FORMAT

1. Basic Format Requirements of the Genre

- Does the writing follow the structural conventions of the specified genre (e.g., essay, article, guide, etc.)?

- Are any mandatory elements of the format (e.g., headings, bullet points, or lists) included and used appropriately?

- Avoiding Abrupt Bullets or Unordered Lists :

- Does the writing avoid disorganized or improperly formatted lists or bullet points that disrupt the flow of the content?

- Are lists used sparingly and only when they enhance clarity?

2. Adequate Titling and Subtitle Structures

- Does the writing include an appropriate, engaging, and informative title?

- If subtitles are required or used, are they logical, helpful, and aligned with the overall structure of the piece?

D.3.3 ADDITIONAL CONSIDERATIONS

- **Consistency with Instruction and Guiding Information**

Always double-check whether the writing adheres to the writing instruction and any specific guiding information provided. A failure to follow core requirements should result in a lower ranking.

- **Avoid Personal Bias**

Focus on the objective quality of the writing, not on personal preferences or subjective interpretations that are unrelated to the task.

- **Use a Systematic Approach**

Ensure that you assess each writing fairly and systematically using the outlined evaluation criteria. If you're unsure between two submissions, revisit the instruction and criteria to resolve ambiguity.

E CONTAMINATION ANALYSIS

A key concern for any new benchmark is whether its data has been seen during LLM pre-training or post-training, which would inflate evaluation scores and undermine the validity of results. To empirically verify that C3EBENCH is free from such contamination, we evaluate sampled pairs from C3EBENCH against known corpora—RedPajama-1T (Together Computer, 2023), GSM8K (Cobbe et al., 2021), and MBPP (Austin et al., 2021)—across both open-source and proprietary LLMs using three complementary metrics:

- **Perplexity (PPL):** Lower values indicate higher exposure during pre/post-training. A high PPL on C3EBENCH would suggest minimal prior exposure.
- **Min-k% Prob:** Averages the log-probabilities of the k% least predictable tokens. High probabilities on these rare tokens explicitly expose pre-training leakage.

- **Greedy Recall (GR):** Measures ROUGE-L at zero temperature. Higher scores indicate direct memorization of the reference text by the LLM.

Table 5: Contamination analysis of C3EBENCH compared against known corpora. PPL = Perplexity (↑ = less exposed), Min-kp = Min-k% Prob (↓ = less leaked), GR = Greedy Recall (↓ = less memorized). "–" indicates metrics unavailable for proprietary models.

Corpus	Metric	Qwen3-1.7B	Qwen3-8B	Qwen3-32B	GLM4-9B	Llama-3.1-8B	GPT-5.2	Gemini-3.1
RedPajama (N=4000)	PPL	17.1	12.1	11.6	12.5	9.2	–	–
	Min-kp	−7.6	−6.9	−6.6	−6.9	−6.0	–	–
	GR	11.1	10.9	11.0	10.4	11.0	11.6	12.3
GSM8K (N=500)	PPL	4.4	2.7	2.4	3.0	2.1	–	–
	Min-kp	−6.5	−4.7	−4.4	−5.1	−3.2	–	–
	GR	28.4	23.5	25.6	21.2	30.8	41.3	35.0
MBPP (N=500)	PPL	17.0	12.8	9.9	9.2	12.0	–	–
	Min-kp	−8.8	−8.5	−8.2	−7.2	−8.0	–	–
	GR	9.0	8.3	8.1	15.8	13.1	19.8	7.4
C3EBENCH (N=500)	PPL	**26.7**	**19.6**	**17.9**	**27.6**	**12.4**	–	–
	Min-kp	−8.5	−7.9	−7.7	−8.4	−6.7	–	–
	GR	**6.9**	**6.2**	**6.1**	**6.6**	**6.8**	**9.7**	**8.7**

Results. Compared to heavily exposed corpora, C3EBENCH exhibits clear signs of non-contamination across all three metrics:

- **High PPL:** C3EBENCH yields the highest perplexity across all open-source models (e.g., 26.7 for Qwen3-1.7B vs. 4.4 on GSM8K), indicating minimal pre-training exposure.
- **Low Min-k% Prob:** The predictability of rare tokens in C3EBENCH is comparable to raw code corpora (MBPP), confirming no pre-training leakage.
- **Lowest Greedy Recall:** C3EBENCH achieves the lowest GR across all models (e.g., 6.1–9.7), far below known-contaminated benchmarks like GSM8K (21.2–41.3), confirming zero direct memorization.

These results collectively confirm that C3EBENCH scores genuinely reflect LLM reasoning and generation capabilities rather than memorized training data.

F DATA EXAMPLE

Table 6: Examples of the four subtasks in C3EBENCH. **For brevity and clarity, the data examples and references are simplified**.

Task	Instruction & Input	Reference
End2End	**Instruction:** Write a formal academic essay comparing Autism and Williams Syndrome as neurodevelopmental disorders. Use separate labeled sections for each condition, covering: definition, causes, physical and behavioral characteristics, diagnostic methods, and treatment/intervention strategies. Target about 1300 words. **Input:** (None)	Concepts of Autism and Williams Syndrome Autism is primarily a physiological problem that results due to problems that occur during the neurodevelopmental process. The disorder manifests itself in the early years of a child's life, with long-lasting effects that are not curable but controllable and easy to deal with on condition that, the concerned parties take the necessary intervention actions after suspecting that a child is suffering from the disorder. The disorder interferes with the growth of important human interactions, for example .....
Expand	**Instruction:** Write a short, rhymed love poem of about 120–180 words, arranged in four quatrains. Address a long-term partner in the first person. **Input:** *Genre:* Story (lyrical poem) *Brief:* A reflective poem where a speaker wonders if they would change the past, only to find comfort in their current love. *Condition:* Use four quatrains. Shift from past questions to an intimate present scene. End with a reassuring quote from the partner.	Ah, love, could we find but one / Of all the dreams we lost / Would we pick it up again / Regardless of the cost? ... Almost asleep on my arm / At the ending of the day / You touch the question carefully / And smile, as if to say: "It doesn't matter how rough the route / When you've finally found some peace / Sometimes the shelter matters most / And the passage matters least."
Cloze	**Instruction:** Please fill in the blanks marked with [fill in the blank] in the following article based on the context. **Input:** *Prefix:* "Alone Together" – Technology and Human Interaction Essay. The author examines the impacts of ICT on individuals in relation to the desire for companionship and time wastage. *Suffix:* Technology is like a revolution. Turkle is of the opinion that advancements in this field are inevitable. The same is evident in the numerous hardware and software designed to suit the 21st century.	Human beings are social creatures. As such, socializing is a common phenomenon. Turkle argues that ICT can be used to meet social demands. Many people associate technology with utopia. Those who lack time to go to social places may use sites like Facebook to maintain a professional lifestyle. Companionship is complicated; technology exists to make life easy, and dating sites help people meet soul mates.

Task	Instruction & Input	Reference
Edit	**Instruction:** Rewrite the content based on the background material while incorporating professional critique and suggestions. Improve expression, logic, and emotional tension. **Input:** *Content:* The health care crisis in California is shaped by social and economic factors. Many residents are uninsured. Budget shortfalls have led to cuts in safety-net services. *Critique:* Lack of specificity and vivid detail. Needs specific figures (e.g., deficit numbers). The causal chain is weak and needs clearer signposting.	The situation in California is among the worst in the nation. One primary reason is the massive budget deficit. As of 2003, the deficit was $35 billion, and by 2010, it reached $41 billion. This leads to deteriorating services for the poor. Another factor is the continuous premium increase, which is a major concern for employees. Additionally, illegal immigrants constitute roughly 20% of the state's 6 million uninsured residents, highlighting a national crisis in medical costs.

The following is an example of CLOZE task from Chinese, and below to it is the corresponding translation.

Example for CLOZE in Chinese (Argumentative Essay Domain)

Instruction

请根据上下文补全以下文章中用**[fill in the blank]** 特殊符号标记出的内容。

Input

Prefix: 最近，有媒体盘点出了中国超级工程里的"世界之最"，在网络上引发了一大波热议：
白鹤滩水电站是目前世界在建规模最大、技术难度最高的水电工程；港珠澳大桥是世界上总体跨度最长的跨海大桥；新疆和若铁路开通运营，让世界首条环沙漠铁路线完成"最后一块拼图"……
有网友称，"中国制造就是中国骄傲"。
而如果我们往深了扒一扒，超级工程的背后，实际上凝聚了大量自主研发的新科技，科技自强自立的背后，最终则是创新的驱动。
习近平总书记在党的二十大报告中22次提到创新，并深刻指出：坚持创新在我国现代化建设全局中的核心地位。报告中还有一处提到：创新是第一动力。
我们来理一理，"核心地位+第一动力"，创新的分量为何这么重?

一

从人类历史来看，社会生产力的每一次发展、科学技术的每一次进步，无不是通过创新实现的。
欧美几个发达国家就是抓住了科技和产业革命的创新机会而一跃跨入现代化行列，实现大国崛起和民族振兴，并引领时代的走向和世界的发展。
有创新就会有发展，谋创新就能谋未来。涅槃于一穷二白旧社会的中国式现代化，也经历了无数次以创新求发展的浴火重生。
特别是新时代以来，在创新驱动发展战略的指引下，我国的"创新型国家"的建设稳步加快。从2012年到2021年，全社会研发投入从1.03万亿元增长到2.79万亿元，全球创新指数排名从第34位上升到第12位。
科技创新在企业壮大、产业升级、区域发展、重大工程建设等方面发挥了重要作用，有力支撑了高质量发展，带动一些关键核心技术相继实现突破，取得了载人航天、探月探火、深海深地探测、超级计算机等重大成果。
九天之外传来的"感觉良好"，深潜海底万米的"妙不可言"，乘坐"复兴号"飞驰万里，睁开"天眼"仰观浩渺宇宙.....这些，都成了网民心中中国科技创新的"名场面"，成了我们心中升腾起的自信和自豪。

二

[fill in the blank]
Suffix: 这样的故事还有不少。这些年，我们在科技"从模仿到创新"的转型过程中遭遇了"追赶的极限"，关键领域核心技术被"卡脖子"的问题愈发突出。
特别是中美贸易摩擦中，我国"缺芯少核"的科技短板暴露了出来。美西方国家利用技术优势地位一方面禁止关键技术流入中国，推动高科技产业链的"对华脱钩"；另一方面阻碍我国核心技术研发，企图将我国彻底压制在产业链中低端。
在激烈的国际竞争中，惟创新者进，惟创新者强，惟创新者胜。正是因为我国科技实力和世界领先水平的差距在不断缩小，一些领域实现了从"跟跑"到"并跑"甚至"领跑"，才引发了美西方国家的战略焦虑，并招致不惜成本的封锁和打压。
然而，我们的目标绝不是跟着西方国家亦步亦趋。我们要开拓出中国式现代化路径，这是一条从未有人走过的路。为人类实现现代化提供新选择，科技创新在其中的核心作用无疑更加凸显。

三

东部沿海省份浙江，为创新之路探了路。
早在2006年，习近平同志在浙江工作时就为浙江定下了用15年时间进入创新型省份行列，基本建成科技强省的目标。当年的"全省科学技术大会"这个会议名称，被习近平同志修改为"全省自主创新大会"。几字之变，意图更加清晰，导向更加明确。

一路走来，“自主创新”这面旗帜始终在之江大地上高高飘扬。今天的浙江，已经拥有良好的科创环境和氛围，三大科创高地加速打造。很多人一提到科创大走廊、之江实验室、西湖大学就想到浙江，这些高能级的平台不仅是浙江的“标签”，也正成为创新的沃土。
有活力就有人才，浙江也越来越成为顶尖人才的向往之地。截至今年8月，全省研发人员总量已达77.58万人，这就意味着大概每1000个浙江人中就有12个科研人员。
而这些科研平台、科研技术、创新力量，则前所未有地融入到百姓的日常当中。在全国率先启动数字化改革一年多来，浙江打造出一批实用、管用的重大应用。“海外智慧物流”“浙农服”“健康码”“政采云”……一个个有着鲜明浙江烙印的数字化应用，便企惠民，香飘墙外、飞向万家。
每个时代，都有打开创新之门的钥匙。比如第一次工业革命是蒸汽机，第二次工业革命是电气化。今天，浙江则以“数”谋“新”，做第一个吃螃蟹的人。

四

今天的世界瞬息万变。大变局之下，唯一的“不变之道”就是以变应变、以新应变。创新，该怎么创？如何新？
“必须坚持科技是第一生产力、人才是第一资源、创新是第一动力，深入实施科教兴国战略、人才强国战略、创新驱动发展战略，开辟发展新领域新赛道，不断塑造发展新动能新优势。”党的二十大报告中的这段话，为创新之路擘画了清晰的领域和路径。
此外，笔者认为，以创新驱动发展还要坚持好以下几个关键点。
创新靠不得别人，还得靠自己。创新能力，关乎一个国家在世界格局中的地位，甚至关乎着国家安全。在世界竞技赛中，跟着别人跑随时可能会被绊倒，只有把创新的自主权、技术的所有权、发展的主动权紧紧攥在自己手中，才能跑出速度、跑到前列。
创新的重要目的之一，是整合资源，打通链条、畅通循环。中国已经是全球第二大经济体，依靠传统的土地、资源和低成本人力来驱动发展已经没有竞争力，也不会有出路。只有用好新型举国体制优势，发挥创新的核心作用，打通不受制于人的产业链、供应链，才能在经济发展中涌现出无数“风口”，在国际竞争中站稳脚跟。
真正的创新，最终要落脚于民。近年来，我国科技创新能力不断提升，越来越多的创新成果广泛应用于民生领域。高铁网络、电子商务、移动支付、互联网+、共享经济……正在深刻改变着人们的衣食住行。不过，实现“人的现代化”也还有很多空白领域，如何围绕老百姓的切身需要，填补这些空白，是需要瞄准的“靶子”。
赶考路上，需要创新来“澎湃”。坚持创新在我国现代化建设全局中的核心地位，坚持创新是第一动力，不仅要让1不断地递增出N，也要探索如何让更多的0实现1的突破。

Reference

科技创新是大国竞争的核心领域。一个国家科技创新能力的高低，决定了其在国际竞争中的水平。
一个经典的故事是，1960年前后，一套重量为3公斤的精密光学坐标镗床主轴轴承，外商对我们的要价竟相当于和轴承同等重量的黄金或6吨重的对虾。直到我们通过自主创新成功攻关，才不再需要依赖进口。
这至少告诉我们两个道理：第一，关键核心技术要不来、买不来、讨不来。只有把它牢牢攥在自己手中，才能从根本上保障国家总体安全。第二，在现代世界体系中，不同国家有着不同的分工。位于“中心地区”的发达国家享有先进技术和高附加值产业，而位于“边缘地区”的欠发达国家只能提供原材料、自然资源和廉价劳动力。这一格局让资本和价值源源不断地向“中心地区”聚集并导致严重的两极分化。

Chinese CLOZE example translated

Instruction

Complete the contents whose position is marked with `[fill in the blank]` according to contexts.

Inputs

***Preifx*:** Recently, the media compiled a list of “world’s best” super projects in China, sparking lively discussions online:
Baihetan Hydropower Station is currently the largest under-construction hydropower project in the world, with the highest technical difficulty; the Hong Kong–Zhuhai–Macau Bridge is the longest cross-sea bridge in the world; the opening and operation of the Xinjiang Hotan-Ruoqiang Railway has completed the “last piece of the puzzle” for the world’s first desert-circling railway line...
Some netizens remarked, “Made in China is China’s pride.”
However, when we dig deeper, we find that behind these super projects lie significant new technologies developed independently, backed by the drive for technological self-reliance and self-strengthening, which in turn is fueled by innovation.
In the report to the 20th National Congress of the Communist Party of China, General Secretary Xi Jinping mentioned innovation 22 times and profoundly emphasized: Innovation must occupy the core position in China’s overall modernization strategy. The report also stated: Innovation is the primary driving force.
Let’s unpack this—“core position + primary driving force.” Why does innovation weigh so heavily?

I

In human history, every advancement in social productivity and every progress in science and technology has always been achieved through innovation.
Several developed Western countries, such as those in Europe and North America, managed to seize the opportunities brought by technological and industrial revolutions, propelling themselves into the ranks of modernized nations, achieving national rejuvenation and rise to prominence, and leading the trajectory of their times and global progress.
Where there is innovation, there is development; where there is a plan for innovation, there is a plan for the future. China’s modernization, which rose from a once-impoverished and backward society, has also undergone countless “phoenix-like rebirths” driven by innovation to seek development. Especially since the advent of the new era, under the guidance of the innovation-driven development strategy, China has been steadily accelerating its progress as an “innovative nation.” From 2012 to 2021, nationwide R&D expenditures increased from 1.03 trillion yuan to 2.79 trillion yuan, and the global innovation index ranking rose from 34th to 12th.
Technological innovation has played a vital role in driving business growth, industrial upgrades, regional development, and the construction of major projects. It strongly supports high-quality development, enabling breakthroughs across critical core technologies in areas such as manned spaceflight, lunar and Mars exploration, deep-sea and deep-earth exploration, and supercomputers.

The “feeling good” phrase transmitted from outer space, the “beyond words” achievement of deep-sea dives exceeding 10,000 meters, the miles sped through on the “Fuxing” bullet train, and the vast universe explored using the “Sky Eye”... All these iconic moments of China’s technological innovation have captured netizens’ imaginations, igniting pride and confidence in all our hearts.

II

`[fill in the blank]`

***Suffix*:**

There are many more stories like this. In recent years, during China’s transformative journey from “imitation” to “innovation,” we have encountered the “limits of catching up,” with challenges in critical core technologies increasingly coming to the forefront.

Particularly during the U.S.-China trade friction, the technological shortcomings labeled as China’s “chip deficiency” and “lack of core technologies” were laid bare. Western countries, leveraging their technical dominance, simultaneously imposed bans on transferring critical technologies to China and tried to “decouple” high-tech industrial chains from China. They also sought to obstruct China’s R&D of core technologies in an attempt to suppress China to the lower ends of the industrial chain.

In the fierce international competition, only those who innovate advance, only those who innovate become stronger, and only those who innovate win. It is precisely because the gap between China’s technological strength and world-leading levels is narrowing, with some fields accomplishing shifts from “running behind” to “running alongside” or even “leading,” that strategic anxiety has arisen among Western countries, prompting them to resort to cost-no-object blockades and suppression.

However, our goal is not to follow in the footsteps of Western nations. Our aim is to pioneer a uniquely Chinese path to modernization—a road never before taken. Offering humanity an alternative modernization model makes the core role of technological innovation even more prominent.

III

The eastern coastal province of Zhejiang has been a trailblazer in the journey of innovation. Back in 2006, while working in Zhejiang, Comrade Xi Jinping set the goal of making Zhejiang an innovation-oriented province within 15 years and essentially building it into a province strong in science and technology. The conference, originally named the “Provincial Science and Technology Conference,” was renamed by Xi Jinping as the “Provincial Independent Innovation Conference.” This subtle change in wording carried a clearer intent and a more focused objective.

Over time, the flag of “independent innovation” has flown high across the land of Zhejiang. Today, Zhejiang boasts an excellent environment and atmosphere for scientific and technological innovation, with three major innovation centers being rapidly developed. Mentioning the Innovation Corridor, the Zhijiang Laboratory, or Westlake University immediately brings Zhejiang to mind. These high-caliber platforms are not only among Zhejiang’s prominent “labels” but are also becoming fertile ground for innovation.

Where there is vitality, there is talent. Zhejiang has increasingly become a magnet for top-tier talent. As of this August, the total number of R&D personnel in the province had reached 775,800, meaning that approximately 12 out of every 1,000 people in Zhejiang work in research.

These research platforms, technologies, and innovation resources have also been unprecedentedly integrated into the daily lives of ordinary people. Thanks to Zhejiang’s bold steps in launching its digitization reform efforts, practical and user-friendly applications have emerged, such as “Overseas Smart Logistics,” “Zhejiang Agricultural Services,” “Health Code,” and “Government Procurement Cloud.” These digital services, bearing the unmistakable imprint of Zhejiang, have benefited businesses and citizens alike, extending their influence far beyond the region.

Every era has its own key that unlocks the door to innovation. For instance, the steam engine in the First Industrial Revolution and electrification in the Second Industrial Revolution. Today, Zhejiang is creating the new with “data,” becoming the first to “try new things.”

IV

The world today is undergoing rapid changes. In an age of great transformations, the only “constant” is to adapt to change with change, and to respond to the new with the new. How should we innovate? What constitutes “new”?

“We must uphold the principle that science and technology are the primary productive forces, talent is the primary resource, and innovation is the primary driving force. We must intensively implement the strategies of rejuvenating the country through science and education, strengthening the nation through talent, and driving development through innovation. We must continuously open new fields and tracks for development and create new momentum and new advantages for growth.” This excerpt from the 20th National Congress report outlines a clear roadmap for the path of innovation.

Moreover, the author believes that driving development through innovation requires adhering to the following key principles: Innovation cannot rely on others; it must depend on ourselves. Innovation capacity determines a nation’s standing in the global landscape and even its national security. In the global competition arena, following others always carries the risk of being tripped. Only by firmly grasping the autonomy of innovation, ownership of core technologies, and initiative in development can we achieve speed and move to the forefront.

One of the primary objectives of innovation is to integrate resources, streamline the chain, and ensure smooth circulation. As the world’s second-largest economy, China can no longer depend on traditional drivers such as land, resources, and low-cost labor for competitiveness or growth. By utilizing the advantages of the new nationwide system and emphasizing the role of innovation, China can build an autonomous and robust industrial and supply chain, generate numerous “opportunities” for economic growth, and secure its position in international competition.

True innovation must ultimately focus on people. In recent years, China’s technological innovation prowess has steadily improved, leading to the widespread application of many innovative achievements in the realm of public welfare. High-speed rail networks, e-commerce, mobile payments, Internet+, the shared economy... all these have profoundly transformed people’s livelihoods. However, there remain many gaps in achieving “human modernization.” Addressing these gaps and meeting the genuine needs of ordinary people becomes the target to aim for.

On this challenging journey, innovation serves as the driving force that powers us forward. By maintaining innovation as the core position in China’s modernization strategy and upholding it as the primary driving force, we must not only ensure the continuous transformation of 1 into N but also explore how to turn more zeros into breakthroughs of 1.

Reference

Technological innovation is the core arena of competition among major powers. The level of a country’s capacity for technological innovation determines its standing in international competition.

A classic story goes that, around 1960, a set of precision optical coordinate boring machine spindle bearings weighing 3 kilograms was offered to us by foreign sellers at a price equivalent to either the same weight in gold or 6 tons of shrimp. It was not until we achieved a breakthrough through independent innovation that we no longer needed to rely on imports.

At least two lessons can be drawn from this: First, key and core technologies cannot be obtained by asking, buying, or begging. Only by firmly holding them in our own hands can we fundamentally ensure the overall security of the nation. Second, in the

modern world system, different countries have different roles. Developed countries in the "core regions" enjoy advanced technology and high value-added industries, while less developed countries in the "peripheral regions" can only supply raw materials, natural resources, and cheap labor. This structure causes capital and value to continuously flow toward the "core regions," resulting in severe polarization.

G Data Quality

G.1 Crawling Sources

For Chinese part, we crawled data from the following high quality and reputable sources:

1. *Chinese Writer Website* (**CN Writer**, 中国作家网) [1] : this cite collects all publishable fictions, proses, poets from professional writers from China, powered by Chinese Association of Writer. The writings are all professionally written. The total number of raw data is approximately 5k.
2. *The pivot website for example essays* (**PW4ES**, 第一范文网) [2] : this cites collects numerous functional writing sources, such as contracts, plans, conclusions, thoughts, speeches and deliveries etc. The writings are of high quality and they serve as examples for learners. The total number of raw data is approximately 30k.
3. *September for example essays* (**SeptES**, 九月范文网) [3]: this cites complements to the above cites, with additional functional writings. The writings are of high quality and they serve as examples for learners. The total number of raw data is approximately 30k.
4. *Zhejiang Publicity* (**ZJPub, 浙江宣传**) [4] : this cites collects numerous argumentatives, critics targeting at social/historical/cultural affairs. These articles are targeting electronic self-media readers, and are written by professional newspaper writers. The total number of raw data is approximately 10k.
5. *Cite for Officials* (**Officials**, 公文网) [5]: this cites collects examples for official articles writings, including propaganda, deliveries, announcements, etc. We purchased the articles from the cite instead of crawling for its commercial use. The articles are written by expert civil servants from the government, and is of high quality. The total number of raw data is approximately 20k.

For English part, we crawled data from the following high quality and reputable sources:

1. *American Rhetoric*[6] : This website records famous speeches in American history, including historical speeches as well as parliamentary speeches and questions.
2. *Obook*[7]: This website records numerous English published books with a wild range of genres, including fiction, prose, poet, novel across 16 century to contemporary.
3. *IvyPanda*[8]. This website serves top level example essays across 32 topics, including art, business, culture, environment, history, music and so on. We use huggingface dataset `qwedsacf/ivypanda-essays` [9] from the same source and the number is approximately 100K.

[1] https://www.chinawriter.com.cn/
[2] https://www.diyifanwen.com/
[3] https://www.chinesejy.com/
[4] https://zjnews.zjol.com.cn/zjxc/
[5] https://www.gongwen.com.cn/
[6] https://www.americanrhetoric.com/top100speechesall.html
[7] https://www.obooko.com/
[8] https://ivypanda.com/
[9] https://huggingface.co/datasets/qwedsacf/ivypanda-essays



Figure 6: Writing genre distribution and Taxonomy.

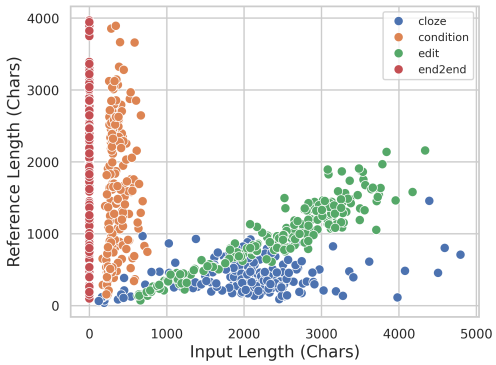


Figure 7: Correlations between the data input-instruction and reference.

G.2 The Source Professional Writing Genre Distribution

G.3 Prompts for Source Text Filtering

Scoring Filter Prompt

Input: content

Please act as a professional fiction reviewer to evaluate the following novel and rate it based on the specified dimensions. For each dimension, assign a score between 1 and 5 and provide a brief explanation. Finally, give the fiction a total score (between 1 and 5).

[Fiction Start]

{content}

[Fiction End]

[Criteria Start]

1. Plot and Structure

- Compactness of the Plot : Are the plotlines smooth and tight? Do they hold enough allure to sustain the reader's interest?
- Structural Layout : Is the novel's structure reasonable? Does it avoid excessive drag or hollow portions in the story? For medium to long-form novels, are there clear stages of exposition, rising action, climax, resolution, and reversals?
- Sense of Rhythm : Is the progression of the story balanced? Does the unfolding of events carry tension and momentum, especially in medium to long-form novels, where pacing is critical?

2. Character Development

- Depth of Characters : Are the characters well-rounded and multi-dimensional? Do they exhibit unique personalities and undergo meaningful changes?
- Character Growth : Does the novel reasonably portray the growth, transformation, or conflicts of its characters? Are there evident internal struggles or character arcs?
- Character Relationships : Are the interactions between characters natural? Do they contribute meaningfully to the advancement of the plot?

3. Themes and Ideas

- Depth of Theme : Does the novel have a clear and compelling theme? Is the theme substantial and thought-provoking?
- Expression of Ideas : Does the novel convey profound thoughts or ideas through its characters, plot, or symbolic elements? Does it inspire reflection in its readers?
- Social and Cultural Context : Does the novel provide deep insight into a particular era, society, or culture through its story and characters?

4. Language and Writing Style

- Language Style : Is the author's language vivid and elegant? Can it effectively convey the emotions and thoughts of the characters?
- Adaptability of Language : Does the language align with the story's atmosphere and context? Does it enhance the emotional intensity of the novel?
- Detail Description : Are the descriptive details fitting and appropriate? Do they aid in character-building, setting the mood, or driving the story forward?

5. Emotional Resonance

- Emotional Depth : Does the novel evoke emotional resonance in readers? Can it make readers empathize and emotionally invest in the story?
- Authenticity of Emotions : Are the emotions in the novel realistic and believable? Do they have the power to move the reader?

6. Innovation and Uniqueness

- Innovative Elements : Does the novel showcase originality in some areas? Does it challenge traditional narrative conventions or stylistic norms?
- Unique Perspective : Does it approach a topic or tell its story from a distinctive angle? Does it reflect a strong, memorable voice or personality?

[Criteria End]

Table 7: Filter score from the coarse rubric scoring system implemented with Claude-4-5-sonnet.

	CN Writer	PW4ES	SeptES	ZJPub	Officials
1	0	153	20	0	2
2	12	351	134	3	15
3	1137	13188	10261	272	8705
4	4468	72908	3216	722	6957
5	19	861	73	86	204

Begin your evaluation by assigning a score between 1 and 10 for each dimension, along with a brief explanation. Conclude with the novel's overall score (1 to 5). A score of 1–2 indicates the dimension performed poorly, 3-4 means it was average, and 5 means it excelled in the dimension. Please use the following example output format:

"Plot and Structure": 2
"Character Development": 3
"Themes and Ideas": 4
"Language and Writing Style": 3
"Emotional Resonance": 3
"Innovation and Uniqueness": 2
"Overall Rating": 3

Table 7 lists the score distribution from the filter.

H Evaluation Prompts

H.1 Evaluation Prompts for END2END

Eval Prompt: End-to-End Creative Evaluation

 # Role
You are a versatile literary editor and critic specializing in instruction adherence and creative quality across essays, poetry, and fiction.

 # Evaluation Criteria
Instruction Adherence (Completion of core requirements), **Structure & Logic** (Genre-appropriate framework), **Creativity & Diversity** (Idiomatic vocabulary), **Form & Format** (Visual layout).

 # Scoring Rubric (Baseline: 6)

- **8–10:** Exceptional creativity or depth; infectious language.
- **6–7:** Fully satisfies instructions with smooth prose.
- **4–5:** Follows instructions superficially; lacks depth.
- **1–3:** Significant language barriers or wrong format.

 # Output Format (JSON)

```
{
  "score": [1-10],
  "analysis": {
    "instruction_and_logic": "...",
    "language_and_creativity": "...",
    "format_check": "..."
  },
  "verdict": "..."
}
```

H.2 Evaluation Prompts for CLOZE

Eval Prompt: Cloze Task Evaluation

 # Role
You are an expert professor specializing in textual logic and semantic analysis. Your task is to evaluate the quality gap between a generated text and a standard Reference in a Cloze (fill-in-the-blank) task.

 # Task & Data Fields
Compare the following fields: 1. [Context], 2. [Reference Answer] (Baseline: 6 points), 3. [Candidate Answer].

 # Evaluation Criteria

- **Semantic Fit:** Accuracy of meaning within the context.
- **Cohesion:** Naturalness of transitions between sentences.
- **Grammar & Expression:** Correctness of word collocations and syntax.
- **Comparison Gap:** Whether Candidate optimizes expression or introduces errors.

 # Scoring Rubric (1–10)
Use Reference as **6-point** baseline:

- **8–10:** Superior to Reference. More idiomatic or sophisticated.
- **6–7:** Equal or slightly better. Completely fluent.
- **4–5:** Sub-par. Semantically correct but awkward.
- **1–3:** Significant errors. Logical breaks or inappropriate context.

 # Output Format (JSON)

```
{
  "score": [1-10],
  "critique": {
    "pros_cons": "...",
    "key_reason": "..."
  },
  "verdict": "..."
}
```

H.3 Evaluation Prompts for EXPAND

Eval Prompt: Outline-Constrained Exapnding Evaluation

 # Role
You are a chief judge proficient in multi-genre literary creation and editing, specializing in assessing works based on outline fidelity and artistic creativity.

 # Evaluation Criteria (Traits)

- **Outline Fidelity:** Coverage of key nodes (years, names, philosophies).
- **Genre Suitability:** Matches characteristics of target genre.
- **Logical Flow:** Organic integration rather than mechanical listing.
- **Creativity:** Vividness and rhetorical richness compared to Reference.

 # Scoring Rubric (1–10)

- **9–10:** Perfectly covers outline; infectious prose.
- **7–8:** Accurately follows outline; clear logic.
- **6:** Baseline (Equal to Sample).
- **4–5:** Stiff logic or misses 1–2 minor details.
- **1–3:** Severely deviates or wrong genre.

 # Output Format (JSON)

```
{
```

```
  "score": [1-10],
  "analysis": {
    "outline_completeness": "...",
    "logic_and_style": "...",
    "vs_reference": "..."
  },
  "verdict": "..."
}
```

H.4 Evaluation Prompts for EDIT

Eval Prompt: Editing Evaluation

 # Role
Senior literary editor specializing in evaluating a writer's revision ability by comparing Draft, Critique, and Reference.

 # Evaluation Criteria (Traits)

- **Instruction Adherence:** Replacing "telling" with "showing."
- **Literary Tension:** Conveying raw emotion rather than "safe" observations.
- **Imagery Precision:** Details carrying the weight of the theme.
- **Flow & Pace:** Unexpected perspectives or psychological shifts.

 # Output Format (JSON)

```
{
  "score": [1-10],
  "critique_fulfillment": {
    "imagery_change": "...",
    "emotional_depth": "...",
    "ending_treatment": "..."
  },
  "vs_reference": "...",
  "verdict": "..."
}
```

I Meta Evaluation Details

I.1 Human Annotator Details

We hired 36 experts in writing with at least bachelor's degree and 23 of them are pursuing master degree or PhD degree in university. 29 of the experts major in literature, history, philosophy, journalism and communication, sociology, phychology and pedagogy. 7 of them are from engineering majors such as environment/engergy/computer science. Annotators were recruited primarily from freelance writers and graduate students in the humanities, social sciences, and journalism. They volunteered to participate after being informed that the annotation tasks supported research on generative AI. Annotators were compensated at USD 10 per hour; each annotation task comprised assessments of nine LLM-generated texts.

I.2 Additional Details on Annotation

Sample Construction. We selected 200 instructions from C3EBench (50 per task: Cloze, Edit, Expand, and End2End), ensuring random and even coverage of all genres. For each instruction, 9 LLM-generated writings are collected from: GPT-5.2, Gemini-3, Claude, Grok-4, GLM-4.7, Qwen3-Max, Kimi-K2, DeepSeek-V3.2, and LLaMA-3.1-405B. Each writing is evaluated on a 1–5 scale along dimensions of content quality, format adherence, and structural coherence.

Annotation Process. To ensure consistency, each set of nine writings corresponding to the same instruction was evaluated by a single annotator, and each individual LLM-generated writing was cross-scored by two annotators for validation. The overall inter-annotator agreement is 0.71 using Cohen's Kappa and 0.87 using Pearson correlation, indicating high consistency among human raters. We averaged the two scores to obtain a single final score for each writing, thereby preserving the diversity of human judgments. Cases with significant disagreement between the two annotators were reviewed by a third expert to resolve edge cases. For all data to be open-sourced, we engaged the five experts who achieved the greatest agreement with other annotators throughout the process and asked them to re-check all annotations.

I.3 Extended Meta-Evaluation: Evaluation Method Selection

Beyond the component ablation presented in the main text (Table 4), we conducted a comprehensive comparison of alternative evaluation methods to justify our final LLM-as-a-judge protocol. The base model for LLM-as-a-judge is GPT-5.2-1211. The evaluation prompt design is detailed in Appendix H, which primarily includes judging criteria, scoring rubrics, dimensional traits, and references. During development, we tested the following alternatives:

- **Overlap Metrics:** BLEU-1, BLEU-rt (brevity-penalty-removed BLEU), and ROUGE-L as reference-based baselines.
- **LLM-judge w/o components:** LLM-judge prompts without rubrics, traits, or human references (corresponding to the ablation in Table 4).
- **Self-Consistency (SC):** Repeat the evaluation N times and average the scores across trials as the final score.
- **Self-Plan:** Given each query, text-to-judge, and reference, the LLM first explicitly plans the evaluation criteria, then uses that self-generated prompt to judge.
- **Self-Plan + SC:** Combining self-planned criteria with self-consistency averaging.

We report system-level Pearson (ρ), Kendall (τ), and Spearman (σ) correlations with human judgments for all four C3EBench tasks, alongside the evaluation cost.

Table 8: System-level correlation with human judgments across evaluation methods. ρ = Pearson, τ = Kendall, σ = Spearman. **Bold** indicates the best per column. Cost is the total API cost in USD for evaluating the full meta-evaluation set.

Method	Cost ($)	Cloze			Expand			End2End			Edit		
		ρ	τ	σ	ρ	τ	σ	ρ	τ	σ	ρ	τ	σ
BLEU-1	–	0.85	0.67	0.80	0.65	0.54	0.69	0.70	0.50	0.62	0.62	0.43	0.58
BLEU-rt	–	0.19	0.06	0.15	−0.25	−0.20	−0.19	−0.45	−0.22	−0.27	0.19	−0.23	−0.08
ROUGE-L	–	0.87	0.67	0.75	0.06	0.14	0.20	0.46	0.22	0.32	0.46	0.35	0.46
LLM-judge (ours)	34.2	**0.87**	**0.78**	**0.78**	**0.85**	**0.76**	**0.89**	**0.89**	**0.78**	**0.88**	**0.88**	**0.93**	**0.93**
w/o rubrics	31.3	0.67	0.62	0.59	0.62	0.63	0.72	0.63	0.52	0.72	0.75	0.77	0.86
w/o traits	31.6	0.69	0.58	0.67	0.65	0.63	0.62	0.58	0.62	0.76	0.75	0.62	0.78
w/o references	28.7	0.45	0.32	0.23	0.52	0.58	0.63	0.17	0.23	0.09	0.34	0.26	0.41
SC (N=5)	171	0.87	0.78	0.78	0.85	0.76	0.89	0.89	0.78	0.88	0.88	0.93	0.93
SC (N=10)	342	0.87	0.78	0.78	0.85	0.82	0.89	0.89	0.78	0.88	0.88	0.93	0.93
Self-Plan	45.3	0.62	0.69	0.72	0.85	0.76	0.79	0.67	0.62	0.67	0.82	0.77	0.88
Self-Plan + SC (N=5)	452	0.67	0.78	0.78	0.85	0.76	0.79	0.67	0.62	0.72	0.82	0.77	0.88

Analysis and Justification. We selected the LLM-judge with rubrics, traits, and human references as our final evaluation protocol for the following reasons:

1. **Self-Consistency offers no gain.** SC (N=5 and N=10) yields virtually identical correlations to the single-pass LLM-judge while incurring 5–10× the cost ($171–$342 vs. $34.2). Unlike reasoning tasks where SC benefits from diverse solution trajectories, evaluation tasks produce highly deterministic scoring with minimal variance across trials, making SC cost-inefficient.
2. **Self-Planning is suboptimal.** LLM-generated evaluation criteria inherently diverge from human judgment preferences, particularly on subjective writing dimensions. Self-Plan achieves lower correlation than the human-designed rubric on Cloze (ρ: 0.62 vs. 0.87), End2End (ρ: 0.67 vs.

0.89), and Edit (ρ: 0.82 vs. 0.88). Combining Self-Plan with SC (N=5) at \$452 still underperforms the single-pass LLM-judge at \$34.2.

3. **Overlap metrics are unreliable across tasks.** While BLEU-1 and ROUGE-L show reasonable correlation on Cloze (where references are short, context-bounded spans), they fail on Expand (ρ: 0.06–0.65) and End2End (ρ: -0.45–0.70), where valid outputs may diverge substantially from the reference in wording while matching in quality. BLEU-rt with removed brevity penalty performs worst, often showing negative correlations.

I.4 Absolute Score Reliability

The system-level correlations reported in Table 8 verify that the LLM judge produces reliable *relative* rankings among models. To further assess *absolute* scoring reliability, we compute the correlation between human judgments (1–5 scale) and LLM scores (1–10 scale, linearly rescaled) at two granularity levels:

- **System-level:** We correlate the averaged scores per model on each task subset. As shown in Table 8, system-level correlations are consistently high (Pearson $\rho \geq 0.85$ across all tasks), confirming strong alignment in overall LLM rankings.
- **Text-level:** For each instruction, we compute the correlation between the human and LLM score vectors across the 9 generated texts (i.e., comparing two 9-dimensional vectors), then average across all 200 instructions. Results are reported in Table 9.

Table 9: Text-level correlation with human judgments. r = Pearson, τ = Kendall, ρ = Spearman. Each cell reports the average correlation across the 9 LLM-generated texts per instruction.

Method	Cloze			Expand			End2End			Edit		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
LLM-judge (ours)	**.64**	**.58**	**.61**	**.62**	**.57**	**.61**	**.56**	**.55**	**.54**	**.67**	**.66**	**.64**
w/o rubrics	.48	.42	.45	.47	.41	.43	.39	.35	.35	.47	.43	.43
w/o traits	.44	.46	.43	.41	.39	.41	.35	.37	.38	.43	.44	.41
w/o references	.17	.14	.14	.25	.21	.23	.19	.17	.19	.31	.31	.33

Analysis. Text-level correlations are consistently lower than their system-level counterparts (e.g., Pearson $r = 0.56$–0.67 at text-level vs. $\rho = 0.85$–0.89 at system-level for the full LLM-judge). This gap reveals two complementary insights:

1. **Individual-text absolute scores are not perfectly reliable.** For a portion of instructions, the LLM judge's pointwise scores diverge from human ratings when viewed in isolation. This is expected given the inherent subjectivity of open-ended writing evaluation.
2. **Aggregate absolute scores yield reliable partial orderings.** When viewing a task subset as a whole, per-model averaged scores produce a stable and accurate ranking among LLMs, validating the use of absolute scoring for benchmark-level conclusions.

The component ablation at text-level further reinforces the findings from system-level analysis: removing references causes the most severe degradation (r drops to 0.17–0.31), followed by removing traits and rubrics.

On Pointwise vs. Pairwise Judgment Consistency. Our main-text meta-evaluation (Table 4) adopts a pairwise (relative) protocol, while the absolute score analysis above employs pointwise scoring. Prior work (Wang et al., 2025) notes that inconsistencies between these two paradigms primarily arise when (1) ambiguous "tie" judgments are prevalent in pairwise settings, or (2) vague rubrics cause information loss in pointwise scoring. Our evaluation design mitigates both failure modes: the pairwise setup excludes tie-pairs when organizing comparison samples, and our evaluation prompts utilize clear rubrics that explicitly define the standard for each scoring interval (Appendix H). The strong alignment observed at both system-level and text-level confirms that our protocol is robust to the choice of evaluation paradigm.

I.5 Completion Annotation Guidance

Completion Writing Scoring Criteria

I. Task Objectives, Fields & Techniques

A. Task Objectives

Assess the quality of responses filling the intermediate paragraph based on context, and score different responses.

- Responses A, B, and C are the model's completions for the text at "fill in the blank" position.
- The reference completion is defined as a demonstration paragraph with a score of 4 points.
- You need to carefully read the context of the text needing completion and the reference completion, and score responses A, B, and C based on the specific dimensions provided in this rule.

B. Field Description

Fixed Fields (No annotation needed)

- **Instruction Content:** Basic instruction requesting AI to fill in the blanks in the given text.
- **Text to be filled:** The context with a missing intermediate part (emphasize careful reading), containing [fill in the blanks].
- **Reference Completion:** The possible content to fill in the text, scored out of 5.
- **Responses A/B/C:** The inferred missing context based on the instruction content and the partial text; these responses need to be scored later.

Note: Replies may contain conversational content, which can be ignored, and only the fill-in content should be evaluated. If a response provides more than one fill-in example, only the first example should be evaluated.

Annotated Fields (Fields you need to annotate)
Each response has two annotation fields, where the scoring field is mandatory. Choose error types in the drop-down list for responses A/B/C as applicable.

- **Annotation Field 1: score A/B/C**
  Score the content format of response A/B/C based on the relevant rules in this document (e.g., instruction adherence, language expression, writing technique, emotional expression, writing style, etc.).
- **Annotation Field 2: Errors in Responses A/B/C (drop-down menu)**
  [Note:] This field is required if the score is below 3. Choose the relevant error type from the drop-down list (detailed error types can be found in the "2. Penalty Items - Error Types" section below).

C. Techniques / Points to Note

- Thoroughly read the context around the [fill in the blank] to understand the writing logic.
- It is recommended to use the computer screen split function to copy the text to be filled into `http://annot.xhanz.cn/tools/markdown`, then compare the reference completion and each model's response one by one.
- Fact-check if there is factual content.
- Accelerate the judgment process by referencing the "III. Scoring Basis (0) Scoring Logic" section.

II. Scoring Basis

Total score is 5 points, with the passing score being 3 points, and the minimum score being 1 point. The reference completion quality corresponds to a 4-point standard.

- **High-Quality Response:** 4–5 points
- **Passing Response:** 3 points
- **Low-Quality Response:** 1–2 points

- **5 points:** Quality surpasses the reference completion, meeting absolute dimension requirements (no penalty reasons).
- **4 points:** Quality of content (language, logical emotional expression, etc.) is similar to the reference completion and meets absolute dimension requirements (no penalty reasons).
- **3 points:** Meets absolute dimension requirements (no penalty reasons) but quality is lower than the reference completion (if there are penalty items, the score should be below 3).
- **2 points:** 1–2 absolute dimensions are not met (requires penalty reasons).
- **1 point:** (requires penalty reasons)
  - More than 2 absolute dimensions are not met;
  - Or, the response performs well in other dimensions (can be scored 3–5 points), but there is a severe security issue, or the [filling instruction] is not followed. In such cases, directly score 1 point.

Scoring Logic

1. **Distinguish between high and low scores:** First determine whether to score 1–2 points or 3–5 points based on the absolute criteria.
2. **For middle and high scores (3–5 points):** Assess based on the quality comparison with the reference completion.
3. **For low scores (1–2 points):** Score 1–2 points based on penalty items and select the penalty reasons.
4. **Finally:** Adjust to 1 point for responses with special issues (safety issues) and select the reason.

III. Detailed Standards

4–5 points Standard

Should be considered high-quality, from the following aspects:

- **Language Expression:** Is the language more accurate and clear? Is the vocabulary more varied? Is the sentence structure more flexible?
- **Content Richness:** Does it appropriately cite speech, poetry, or allusions?
- **Writing Techniques:** Are rhetorical devices used more aptly and skillfully?
- **Emotional Expression:** Is the expression more natural and forceful?

(A) Absolute Criteria (For a baseline score of 3)

- **Consistency:** The completion should align with the context in perspective, narrator, logic, and style.
- **Fact Consistency:** Facts should align with the context and have no errors in external knowledge.
- **Fluency:** No language errors, logical contradictions, or mixed language issues.

(B) PENALTY ITEMS - ERROR TYPES

If the following errors are present, the score **must** be below 3.

- **A. Consistency Issues:** Format inconsistency (e.g., response is a single sentence while context expects a paragraph), content inconsistency, or repeated content.
- **B. Accuracy Issues:** Factual errors in quotes, published knowledge, dates, or common sense (e.g., "the sun rising from the west").
- **C. Fluency Issues:**
  - Unmeaningful repetition.
  - Mixed Language: Mixing languages (e.g., "I say this is not okay" mixed with other scripts) deducts points.
  - Special Character Issues: Unfit characters or odd symbols like ^, &.

(C) SPECIAL CASES: SAFETY ISSUES

Directly score 1 point for:

- Violent, obscene, or abusive content.
- Inducing self-harm or illegal activities.
- Defamation or incorrect representation of national leaders/governments.

J PROMPTS FOR **RAVEL** IMPLEMENTATION

System Prompt: Policy

Role: You are a "Writing Project Manager" proficient in the entire content creation lifecycle. You are responsible for building high-quality content from scratch. By observing the current state, you autonomously decide which tool to call next.

WritingState Field Analysis:
You will receive a real-time `WritingState` JSON object with the following fields:
- `meta`: Contains the article's `topic` and `style_guide`.
- `outline`: A list structure storing the article skeleton.
  - `id`: Unique identifier for the section.
  - `section_title`: Title of the section.
  - `points`: Core points that must be covered in this section.
  - `status`: Lifecycle status of the section (`pending`, `drafted`, `revision_needed`, `completed`).
- `manuscript`: A key-value map (Key is `section_id`) storing specific content, summary, score, and feedback.

Tools & Documentation (Action Space):
1. `plan_outline(topic: str, style_guide: str)`: Generate a full-text outline based on topic and style.
2. `write_paragraph(topic, style_guide, section_id, ...)`: Write a specific section based on core points and previous summary.
3. `review_content(section_id, style_guide, points, content)`: Evaluate the quality of a drafted section.
4. `revise_paragraph(section_id, style_guide, points, content, feedback)`: Revise a section using feedback.
5. `finish()`: Announce the end of the task if all sections are `completed`.

Output Format:
You must strictly return the following JSON format:

```
{
  "thought": "Deep thinking based on current state...",
  "action": "Tool Name",
  "params": {"parameter_name": "value"}
}
```

Action as Function: Outlining

Role: Senior Content Planning Editor
Task: Generate a logically rigorous and clearly structured article outline based on the topic and style guide.

Constraints:
1. Must output in JSON format.
2. The outline should contain several core sections designed to fit the genre specified in the `Style_guide`.
3. Each section must provide specific writing points (`points`).

Output Format Example:

```
{
  "title": "Article Title",
  "outline": [
    {"section_title": "Introduction", "points": "Background, core
        thesis"},
    {"section_title": "...", "points": "..." }
  ]
}
```

Action as Function: Drafting

Role: You are a senior copywriter, skilled at constructing penetrating body content based on rigorous logical outlines.

Task: Write a high-quality paragraph based on the provided global information and current section elements.

Input Data:
- **Topic:** `topic`
- **Style Guide:** `style_guide`
- **Section Title:** `section_title`
- **Core Points:** `points`
- **Previous Summary:** `prev_summary` (If "None" or empty, start the article directly).

Writing Requirements:
1. **Style Adherence:** Strictly follow the `Style Guide`.
2. **Logical Flow:** Ensure the first sentence connects naturally with `Prev Summary`. Do not repeat facts; build upon them.
3. **Content Richness:** Enrich `Points` with facts, data, logical deduction, or case studies.
4. **Output Restrictions:** Output only the body content; do not include section titles.

Format Example:

```
{
  "content": "Drafted body content here...",
  "summary": "Summarize the core content and ending state..."
}
```

Action as Function: Review

Role: Senior Content Reviewer
Task: Perform a multi-dimensional evaluation of the provided paragraph and provide suggestions for improvement.

Audit Criteria:
- **Fulfillment:** Is all core information in `Points` reflected?
- **Style Adherence:** Is the `Style Guide` strictly followed?
- **Logic & Cohesion:** Is internal logic consistent? Natural transition?
- **Depth:** Is the content superficial?

Scoring Rules:
- **9-10:** Perfectly meets requirements, inspiring language.

- **8-8.9:** Meets requirements; minor flaws (Passed).
- **Below 8:** Clearly point out missing points or sub-standard style.

Input Data: `{{style_guide}}`, `{{points}}`, `{{content}}`, `{{prev_summary}}`

Format Example:

```
{
  "score": 8.5,
  "feedback": "Logic is clear but details are insufficient..."
}
```

Action as Function: Refine

Role: Senior Polishing Editor

Tasks:
1. **Precise Repair:** Target and optimize all deficiencies in `Feedback`.
2. **Maintain Integrity:** Ensure full coverage of `Points`.
3. **Style Alignment:** Cross-reference the `Style Guide`.
4. **Seamless Transition:** Ensure flow from `Prev Summary`.

Input Data:
- **Original Content:** `{{content}}`
- **Reviewer Feedback:** `{{feedback}}`
- **Points:** `{{points}}`
- **Style Guide:** `{{style_guide}}`

Format Example:

```
{
  "revised_content": "High-quality revised body text...",
  "change_log": "Briefly describe the main adjustments made..."
}
```

K Extended Analysis: Reasoning vs. Generation

The analysis in Section 5.4 demonstrates that reasoning dominates generation in agentic text synthesis, based on swapping Gemini-3 Pro into two Qwen3 backbones. A natural concern is whether this conclusion generalizes beyond the specific base–augmentor pairing. To validate our findings, we expand the study across additional **base models** (Llama-3.1-70B, GLM-4.7) and **augmentors** (GPT-5.2, Claude-4.5), yielding a broader cross-model analysis.

Findings. The extended results in Table 10 are consistent with the claims in Section 5.4:

- **Upgrading the reasoner consistently improves synthesis.** Across all four base models and three distinct augmentors, replacing the reasoner (π_θ) with a stronger model yields the highest success rates ($\mathcal{S}$) and best final judge scores. Notably, the reasoner upgrade positively correlates with δ_{gain} and final scores, confirming that high-quality action planning is the primary driver of synthesis quality.
- **Upgrading the generator yields marginal or negative gains.** A stronger refiner (f_{refine}) provides smaller success rate improvements and does not guarantee better final scores. High refinement rates (ρ_{ref}) under the refiner-upgrade setting do not translate to quality improvements, reinforcing that iterative local editing cannot compensate for suboptimal structural reasoning.
- **Augmentor–base compatibility matters.** We observe that Gemini-3 Pro, despite being a SOTA model, struggles to assist weaker base models when acting solely as a refiner (e.g., $\delta_{\text{gain}} = -12.16$ for Llama-3.1-70B). In contrast, GPT-5.2 and Claude-4.5 as refiners avoid negative gains but still underperform their reasoner counterparts. This suggests that refiner effectiveness is bottlenecked by the base model's ability to produce structurally sound drafts for the refiner to improve upon.

Table 10: Extended sensitivity analysis of reasoning vs. generation capabilities within **RAVEL**. For each base model, we compare the default configuration (*Base*) against upgrading only the reasoner (π_θ) or only the refiner (f_{refine}) to a stronger external model.

Base Model	**Configuration**	$\mathcal{S}$ (%)	T	ρ_{ref} (%)	δ_{gain}	**Judge**
Llama-3.1-70B	Base	91.8	24.1	18.7	11.39	4.58
	π_θ=Gemini-3 Pro	**98.9**	18.8	0	0	**5.62**
	f_{refine}=Gemini-3 Pro	92.9	24.6	25.4	−12.16	4.27
GLM-4.7	Base	73.3	18.0	38.5	0	6.31
	π_θ=Gemini-3 Pro	**94.6**	19.9	50.4	0.29	**6.67**
	f_{refine}=Gemini-3 Pro	90.8	20.6	52.3	0	6.29
Qwen3-32B	Base	58.5	21.2	105.8	5.54	5.54
	π_θ=GPT-5.2	**82.3**	31.4	12.3	14.5	**6.05**
	f_{refine}=GPT-5.2	77.3	21.0	78.4	0	5.73
Qwen3-Max	Base	47.8	13.1	23.9	0	5.74
	π_θ=Claude-4.5	**93.2**	23.4	54.3	23.1	**6.57**
	f_{refine}=Claude-4.5	89.4	19.8	0	0	5.96

L Self-Scoring and Protocol Sensitivity

We examine how internal review scores, the stopping threshold, and the action protocol affect the interpretation of **RAVEL**'s metrics. External Judge scores assess final-text quality; completion and trajectory statistics describe behavior under the specified execution protocol.

L.1 Self-Review versus External Judgment

The main experiments use a shared threshold $\tau = 8.0$, without per-model calibration. Table 11 compares mean internal review scores with external Judge scores for six models from the self-scoring analysis. Internal scores are consistently higher, with larger gaps for weaker models. These are descriptive differences between section-level self-reviews and final-article judgments, which use different evaluation contexts; they are not paired calibration errors on identical texts. The Judge values below are those reported for this diagnostic analysis and are kept separate from Table 2.

Table 11: Self-scoring diagnostic. Self-review averages include all reviews; pass rate is the fraction meeting $\tau = 8.0$. Gap is mean self-review minus external Judge, rounded to one decimal.

Model	Self-review	Pass rate (%)	Judge	Gap
Gemini-3 Pro	9.44	99	7.05	+2.4
GPT-5.2	9.04	99	6.74	+2.3
Claude-4.5 Sonnet	8.89	95	6.71	+2.2
Qwen3-32B	8.64	95	5.56	+3.1
Hermes-3-Llama-70B	8.94	100	4.56	+4.4
Llama-3.1-8B	8.25	96	1.66	+6.6

High review pass rates do not guarantee episode completion: $\mathcal{S}$ also depends on planning, action selection, and the step budget. For example, GPT-5.2's mean outline has 15.5 nodes and its completion rate is 64.3% in Table 2, whereas Gemini-3 Pro uses 7.1 nodes and completes 95.1% of episodes. This is consistent with planning and convergence contributing to completion, but does not rule out self-scoring calibration as a confound.

We retain the common threshold as a documented protocol setting. We report $\mathcal{S}$, η_{traj}, and ρ_{ref} separately from external quality rather than combining them into a weighted score; efficiency comparisons should consider matched quality. Calibration on an independent set remains a possible extension.

L.2 Sensitivity to the Stopping Threshold

We re-ran autonomous **RAVEL** on a fixed seed-0 subset of 50 End2End items for GPT-5.2, Qwen3-Max, and Qwen3-32B at $\tau \in \{2, 5, 8\}$. To expose τ as an experimental control, the policy prompt's hard-coded 8.0 rule was removed; the review rubric was unchanged. Every article was scored by the same external GPT-5.2-1211 Judge.

Table 12: Threshold sensitivity on 50 End2End items per model. Each cell is external Judge / completion rate $\mathcal{S}$ (%).

Model	$\tau = 2$	$\tau = 5$	$\tau = 8$
GPT-5.2	7.42 / 96	7.44 / 98	7.46 / 96
Qwen3-Max	5.80 / 22	5.66 / 20	5.70 / 20
Qwen3-32B	5.54 / 100	5.56 / 100	5.59 / 100

Table 12 shows within-model Judge ranges of 0.04, 0.14, and 0.05 points, respectively, with unchanged model ranking. Completion varies little: Qwen3-Max remains at 20–22% even under the lowest threshold. These results suggest limited threshold sensitivity within this subset, but neither establish invariance outside the tested conditions nor eliminate calibration concerns. Because the sweep modifies the policy prompt, it is a separate experiment from the main results.

L.3 Action-Protocol Ablation

We evaluate the same three models on a fixed seed-0 subset of 50 End2End items under four protocols ($3 \times 50 \times 4 = 600$ runs): autonomous action selection; a forced outline→draft→review→refine schedule; no-review, which skips both review and the review-score gate; and no-refine, which retains review but disables revision. All outputs are scored by the same external GPT-5.2-1211 Judge. Table 13 reports quality and completion for each protocol.

Table 13: Action-protocol ablation on 50 End2End items per model. Each cell is external Judge / completion rate $\mathcal{S}$ (%). Completion is measured under each protocol's own rules.

Model	Autonomous	Fixed	No-review	No-refine
GPT-5.2	7.40 / 96	7.40 / 96	7.34 / 100	7.50 / 96
Qwen3-Max	5.92 / 24	7.08 / 100	7.10 / 100	7.20 / 100
Qwen3-32B	5.50 / 100	5.92 / 100	5.86 / 100	5.80 / 100

The ranking GPT-5.2 > Qwen3-Max > Qwen3-32B holds under all four protocols, but absolute quality depends on the protocol. A fixed schedule raises Qwen3-Max's Judge score by 1.16 points and its completion rate by 76 percentage points. Qwen3-32B gains 0.42 Judge points with unchanged completion, while GPT-5.2's mean Judge score is unchanged. Thus, protocol choice can affect quality even when completion is stable.

The no-review and no-refine variants do not consistently lower quality. These results suggest that the benefit of autonomous review and revision depends on the model and workflow; they do not establish an isolated causal contribution of either sub-ability. In particular, removing review also removes its score gate, so higher completion under no-review cannot be interpreted as stronger self-assessment. The results concern three models and one 50-item subset, without reported uncertainty estimates, and should not be generalized to all models or prompts.

M REPRODUCIBILITY STATEMENT

Draft template – pending author confirmation; not submission-ready.

The anonymous repository at https://anonymous.4open.science/r/ICLR2027-RAVEL-Reasoning-Agents-Text-Eval/ provides **[TO COMPLETE: currently available code, data or permitted data subsets, evaluation scripts, configurations, dependency specifications, and outputs/logs]**. The release corresponding to this paper is **[TO COMPLETE: version or commit identifier]**. Data construction, evaluation prompts, and agent implementation prompts are documented in Appendices C, H, and J, respectively; human evaluation is described in Appendix I. The repository documents **[TO COMPLETE: commands and configuration paths for the main results, ablations, and supplementary experiments; identify any experiments not covered]**. Model versions and access requirements, generation settings, random seeds, repetitions, and resource requirements are specified in **[TO COMPLETE: exact locations and any unavailable information]**. The LLM judge used in the main experiments and extended meta-evaluation is GPT-5.2-1211. Reproduction has been checked by **[TO COMPLETE: who ran which experiments, in what environment, and with what outcome; state if a clean reproduction has not been performed]**. Remaining constraints include **[TO COMPLETE: proprietary API access or version changes, stochastic variation, cost, and data-redistribution restrictions, as applicable]**.