

Stay Focused: Problem Drift in Multi-Agent Debate

Jonas Becker^{1,2,*}, Lars Benedikt Kaesberg¹, Andreas Stephan³,
Jan Philip Wahle¹, Terry Ruas¹, Bela Gipp¹

¹University of Göttingen, Germany; ²LKA NRW, Germany; ³University of Vienna, Austria

*Correspondence: jonas.becker@uni-goettingen.de

Abstract

Multi-agent debate — multiple instances of large language models discussing problems in turn-based interaction — has shown promise for solving knowledge and reasoning tasks. However, these methods show limitations when solving complex problems that require longer reasoning chains. We analyze how multi-agent debate drifts away from the initial problem over multiple turns, thus harming task performance. We define this phenomenon as **problem drift** and quantify its presence across ten tasks (i.e., three generative, three knowledge, three reasoning, and one instruction-following task). We find that generative tasks drift often due to the subjectivity of the answer space (76-89%), compared to high-complexity tasks (7-21%). To identify the reasons, eight human experts analyze 170 multi-agent debates suffering from problem drift. We find the most common issues related to this drift are the lack of progress (35% of cases), low-quality feedback (26% of cases), and a lack of clarity (25% of cases). We propose **DRIFTJudge**, an LLM-as-a-judge method, as a first baseline to detect problem drift. We also propose **DRIFTPolicy**, which mitigates 31% of problem drift cases. Our study is a step toward understanding a key limitation of multi-agent debate, highlighting why longer debates can harm task performance and how problem drift could be addressed.

1 Introduction

Inspired by Social Choice Theory (Endriss, 2017), recent research considers the use of multiple large language models (LLMs) to solve complex reasoning tasks and mitigate the limitations of single models, such as lacking modularity and answer diversity (Yin et al., 2023; Chen et al., 2024; Sun et al., 2024). Agents, embedded with varying expertise, memory, planning, and tool use (Chen et al., 2024; Sun et al., 2024; Zhuang et al., 2023; Baker et al., 2020), can emulate human interaction to

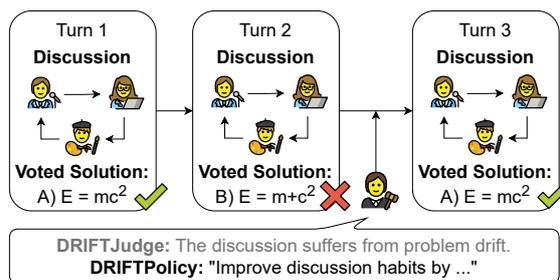


Figure 1: Problem drift in MAD. **DRIFTJudge** detects problem drift at test-time. **DRIFTPolicy** provides on-demand feedback about the conversation.

solve problems (Guo et al., 2024; Becker, 2024). Recent work highlights the strengths of multi-agent debate (MAD) in reasoning and creativity (Zhao et al., 2023; Xu et al., 2023; Suzgun and Kalai, 2024). MAD also scales test-time compute to solve challenging tasks similar to reasoning models, such as OpenAI o4 (OpenAI) and DeepSeek R1 (Guo et al., 2025), which can be more effective than scaling training compute or data (Snell et al., 2024).

However, the increasing complexity of MAD can promote errors and undesired behaviors (e.g., propagating hallucinations) (Guo et al., 2024). Problems originating from debate comprise agent orchestration (Wang et al., 2024b; Shah and White, 2024), diminished planning capabilities (Valmeekam et al., 2023), and ineffective criticism generated by the LLM (Stechly et al., 2023). Still, it is largely unknown what causes MAD to fail and why it happens. It remains uncertain how MAD can avoid issues like ineffective criticism, which is crucial for high-quality debate and reliable results. Few mitigation methods have been proposed for the few errors identified by related work.

This paper systematically analyzes errors that lead to performance degradation in MAD. Through automated methods and a human evaluation, we observe that multi-agent systems can collapse in long discussions. In some debates, agents' commu-

nication tends to deteriorate over time, drifting to a point where they can not recover and address the original task goal. We refer to this phenomenon as **problem drift** and investigate its underlying causes and effects. To identify problem drift post-hoc in MAD, we propose **FOCUS**, a simple yet effective metric that measures the quality of the ongoing discussion. Our results suggest that problem drift occurs across all tested tasks, being most prevalent in generative tasks (76%-89% of samples), followed by instruction-following tasks (21% of samples). The majority of observed drifts do not recover. Once a discussion drifts away, agents often do not reach the correct solution for a given task, except in 9% of the translation and 45% of the ethical question-answering examples. To identify and mitigate problem drift at test-time, we propose **DRIFTJudge** and **DRIFTPolicy**. While **DRIFTJudge** identifies problem drift acting as an LLM-as-a-judge (Zheng et al., 2023), **DRIFTPolicy** introduces a policy feedback agent (Fu et al., 2023) that advises participating agents to improve the debate. Figure 1 shows the dynamics between **DRIFTJudge** and **DRIFTPolicy** during MAD. We show that **DRIFTPolicy** can reduce the number of drifting discussions by 31%, improving task accuracy by up to 3.6% for weaker model agents.

By asking eight human experts, we identify eight reasons why problem drift occurs in MAD and categorize them into temporal error types (e.g., lack of progress) and local error types (e.g., task non-compliance). Agent discussions are especially prone to a lack of progress (35% of drifting samples) that can lead agents to overanalyze problems and give low-quality feedback. Our work can be seen as a systematic analysis of MAD, showing the inherent limitations of prolonged agent interaction. We release the code and data publicly¹.

2 Related Work

Ever since the first chatbots², humans have been fascinated by the ability of computers to communicate in a human-like manner. Recent advances in LLMs to reason and solve complex tasks (OpenAI et al., 2024) lead to a surge in studies on multi-agent systems (Zhao et al., 2023; Xu et al., 2023; Suzgun and Kalai, 2024). Guo et al. (2024) conduct a literature review on multi-agent

LLMs. Their taxonomy points to two main areas in which multi-agent LLMs are used: simulation and problem-solving. Our investigation explores problem-solving because these tasks offer a controllable environment to probe components in multi-agent interaction (Yin et al., 2023; Du et al., 2023).

Supportive Works. Several works highlight the strengths of MAD, including divergent thinking (Liang et al., 2024), reasoning (Yin et al., 2023), creative writing (Schick et al., 2022), dialogue generation (Chen et al., 2024), and theory of mind (Li et al., 2023). These works often use prompted personas (Wang et al., 2023b; Xu et al., 2023) and self-correction mechanisms like self-consistency (Wang et al., 2023a) in a conversational setup. However, the heterogeneous experimental setup across these studies (e.g., agent orchestration, decision-making, prompting) hinders their comparability (Becker, 2024; Guo et al., 2024).

Critical Works. Other researchers study the limitations of multi-agent systems. Work led by Microsoft and Salesforce recently showed that many LLMs perform worse in multi-turn settings on generative problems, suffering from a systematic performance degradation (Laban et al., 2025). Wang et al. (2024b) and Zhang et al. (2025) focus on the computational cost of MAD, showing that single-agent LLMs can often achieve similar or even better performance through prompting. Systems that include a self-critique mechanism (e.g., MAD) might diminish planning performance (Valmeekam et al., 2023). The correctness and content of LLM-generated criticism can be irrelevant to the performance of iterative prompting (Stechly et al., 2023). Shah and White (2024) argue that MAD has many issues, such as generalization, scalability, coordination, robustness, and ethical concerns.

Research Gap. We highlight that MAD is relevant for problems that require divergent input or modular solutions. While existing literature highlights diminished returns with prolonged MAD, the causes for such remain underexplored. We address this gap by characterizing the deterioration of MAD performance (RQ1), its prevalence (RQ2), and whether agents can recover from it (RQ3). We systematically study possible causes and effects through a human evaluation (RQ4) and propose a first baseline to detect (RQ5) and recover (RQ6) the degrading discussions.

¹<https://github.com/jonas-becker/problem-drift>

²The first recorded conversation between the chatbots ELIZA and PARRY is available here: <https://www.rfc-editor.org/rfc/rfc439>

3 Problem Drift

We introduce the concept of **problem drift** in MAD, which we define as the systematic decay in turn-based task performance as agents gradually diverge from the initial task. To quantify this effect, we propose **FOCUS** to measure the strength of the decay over a number of consecutive turns.

FOCUS. We describe how the quality of a solution differs in one turn from the previous turn. Let (x, y) be a pair of input and gold label for a task example in a dataset to be solved. A multi-agent system discusses for N rounds. At the end of each round $r \in (1, \dots, N)$, the agents try to solve x by providing an answer $\hat{y}^{(r)}$. Let $M(\hat{y}^{(r)}, y) \in [0, 1]$ be a downstream performance metric that compares the ground-truth solution to the solution produced by the MAD, where 1 corresponds to the solution and 0 corresponds to a wrong solution. Values between 0 and 1 can indicate partial solutions for non-binary solutions. We define *problem focus*, further called **FOCUS**, as:

$$FOCUS_r = M(\hat{y}^{(r)}, y) - M(\hat{y}^{(r-1)}, y) \in [-1, 1] \quad (1)$$

If $FOCUS_r < 0$, the solution’s quality degrades during round r . If $FOCUS_r > 0$, the solution improves. **FOCUS** purposely depends on task-performance as an outcome-based score. We also test a semantic-based alternative in Appendix C but find two main limitations that make it unattractive for this work.

Problem Drift. We define that an ongoing discussion has *problem drift* if the current solution obtains lower performance than the solution of a previous debate round over multiple rounds. Let M be the number of consecutive discussion rounds for one task example. A debate has problem drift from round 1 to M of strength $\theta \in [-1, 0]$ if:

$$FOCUS_{1,M} = \sum_{r=1}^M FOCUS_r = \theta \quad (2)$$

We define problem drift by the simple sum over **FOCUS** values to model the shift in task performance across debate rounds.

Recovery. We say an example recovered from problem drift when the discussion gets back to or improves upon the performance before the problem drift occurred. We define a discussion as recovered from problem drift if $\exists N, M$ where $N > M$ s.t.:

$$\sum_{r=1}^N FOCUS_r \geq 0 \quad \wedge \quad \sum_{r=1}^M FOCUS_r < 0 \quad (3)$$

4 Methodology

We explain the multi-agent environment for this study, our proposed detection and mitigation methods for problem drift, and the datasets and metrics.

4.1 Debate Setup

We define three components for running MAD: an interface for agents, discussion paradigms, and the decision-making protocol. For each task, we create three agents with distinct **expert personas** (Xu et al., 2023; Shi et al., 2023). We choose three agents as a hyperparameter following previous works (Chen et al., 2024; Yin et al., 2023). These personas are automatically generated by *Meta-llama/Meta-Llama-3.1-70B-Instruct* (Grattafiori et al., 2024) and induce a set of preferences (e.g., law attorney), which varies according to the task and sample (examples and details are in Appendix F). Second, we run a total of **seven turns** of conversation between the agents. We find that in 99% of cases, agents reach a first agreement within the first two turns. Thus, we choose seven turns to capture the effects of prolonged agent interaction sufficiently. This aligns with other setups for MAD (Yin et al., 2023) and enables our analysis of recovery following problem drift. Each agent can generate one new message at each turn and indicate their agreement with the current solution. All agents see the messages of one another. If an agent disagrees, it proposes a new solution concomitantly. Finally, we employ two decision-making protocols that other researchers regularly use to ensure that our findings generalize across conversational setups (Yin et al., 2023; Kaesberg et al., 2025; Yang et al., 2024). **Voting** is conducted between possible discussed solutions to reach a final solution for that turn (Yang et al., 2024). Following prior work, agents have access to one previous turn and the previously voted solution (Pagnoni et al., 2021; Zhang and Zong, 2020). A visual overview of the process is presented in Figure 2. **Consensus** involves each agent directly modifying the current draft at their turn (Yin et al., 2023). A visual overview of the process can be found in Figure 9 of Appendix H.

We conduct our experiments using *Meta-llama/Meta-Llama-3.1-70B-Instruct* on eight NVIDIA A100 GPUs, and *Qwen/Qwen2-7B-Instruct* on two NVIDIA A100 GPUs with 40 GB each. Detailed information about the framework (Becker et al., 2025), parameters, and prompts is available in Appendices A, B and D. We report our

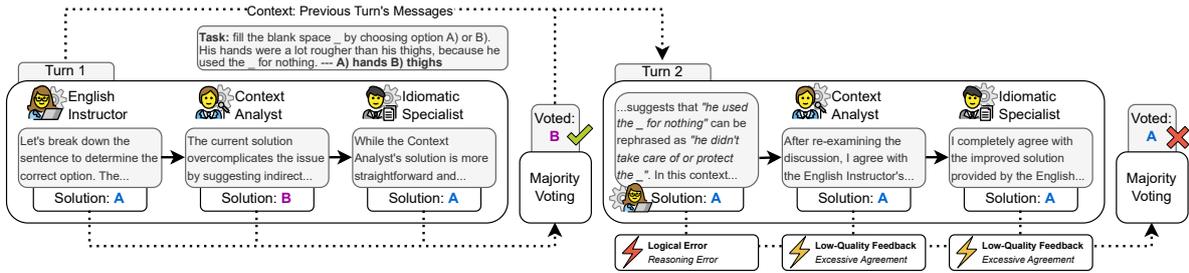


Figure 2: Example of problem drift in MAD. The *English instructor* induces a logical error in the discussion. The other agents agree without skepticism, leading to the wrong solution and problem drift.

main results using *Llama-3.1* and *Voting* but also include results for *Qwen2* and *Consensus* to ensure the generalizability of our findings. Our goal is to show that problem drift and **FOCUS** are not exclusive to the specific design of a multi-agent system and apply to any MAD with intermediate solutions. As the framework³ and the code⁴ are open source, the exact settings can be investigated, and all experiments can be reproduced.

4.2 Mitigation Setup

A natural question from our newly proposed definition of problem drift is whether we can mitigate it even when unsure about its presence (at test-time).

4.2.1 Detection with DRIFTJudge

We aim to identify the occurrence of problem drift at test-time when gold labels are unknown. We propose a first baseline to detect problem drift inspired by LLM-as-a-judge (Zheng et al., 2023), which receives a focal turn’s and a consecutive turn’s solution to assess whether problem drift occurs. To ensure that our results come from architectural changes instead of model capabilities, we refrain from using a powerful fine-tuned model as our judge. We instead use the same model as for the agents with **DRIFTJudge**. This mitigation concept is independent of our model and could also use lightweight classifiers or models.

4.2.2 Mitigation Methods

We propose two mitigation methods to recover debates in the event of drift detection. We either **regenerate** the focal drifting turn or use a **DRIFT-Policy** agent that provides situational feedback.

Regenerate. We undo the turn suffering from problem drift and regenerate that turn with a non-

zero temperature. We only regenerate a turn once.

DRIFTPolicy. We introduce a fourth agent at the end of the drifting turn that provides feedback about how to improve the current conversation and solve the issues leading to problem drift. **DRIFT-Policy** is inspired by policy feedback agent by Fu et al. (2023) and the feedback mechanism used with error types in Kirstein et al. (2025). We include the prompt for our feedback generation in Appendix D.

4.3 Datasets & Metrics

Datasets. We select ten datasets from four domains: three generative tasks (XSum (Narayan et al., 2018), ETPC (Kovatchev et al., 2018), WMT19 (Foundation, 2019)), three reasoning tasks (StrategyQA (Geva et al., 2021), WinoGrande (Sakaguchi et al., 2019), AQUA-RAT (Ling et al., 2017)), three knowledge tasks (ETHICS (Hendrycks et al., 2023), MMLU-Pro (Wang et al., 2024e), GPQA (Rein et al., 2023)), and one instruction following task (IFEval (Zhou et al., 2023)). With qualitative assessment, we identify StrategyQA, AQUA-RAT, and MMLU-Pro as *complex* tasks, as they require a higher number of reasoning steps to be solved. The generative tasks ETPC, XSum, and WMT19 are characterized by *subjectivity*, due to ambiguities or multiple valid solutions. As MAD uses large amounts of test-time compute, it has become a common practice to evaluate subsets of datasets to study multi-agent systems (Yin et al., 2023; Chen et al., 2024). We follow this approach. Detailed information about the sampling process can be found in Appendix E.

Metrics. We evaluate multiple-choice datasets (i.e., ETHICS, GPQA, MMLU-Pro, StrategyQA, WinoGrande, and Aqua-Rat) by accuracy. For generative tasks (i.e., ETPC, XSum, and WMT19), we use BERTScore (Zhang et al., 2019). We evaluate IFEval by the “strict” accuracy (Zhou et al., 2023).

³<https://github.com/Multi-Agent-LLMs/mallm>

⁴<https://github.com/jonas-becker/problem-drift>

	Dataset	# of samples staying at high perf.	# of samples staying at low perf.	# of improving samples	# of worsening samples	Total Samples
		$P(\hat{y}^{(r)}, y) > 0.7 \forall r$	$P(\hat{y}^{(r)}, y) < 0.7 \forall r$	$FOCUS_{1,7} > 0$	$FOCUS_{1,7} < 0$	
Generative	ETPC	2.0% (22)	1.9% (21)	26.5% (287)	69.5% (753)	1.083
	XSum	0.3% (3)	7.9% (91)	37.9% (438)	54.1% (626)	1.158
	WMT19	14.0% (143)	3.0% (31)	13.2% (135)	69.9% (714)	1.023
Reasoning	StrategyQA	72.2% (715)	18.3% (181)	3.7% (37)	5.8% (57)	990
	WinoGrande	63.4% (561)	21.3% (188)	6.8% (60)	8.6% (76)	885
	AQUA-RAT	70.5% (537)	21.0% (160)	3.8% (29)	4.7% (36)	762
Knowledge	GPQA	33.2% (225)	49.9% (338)	8.1% (55)	8.7% (59)	677
	MMLU-Pro	51.7% (578)	35.7% (399)	4.0% (45)	8.7% (97)	1.119
	ETHICS	64.0% (675)	22.0% (232)	5.7% (60)	8.2% (86)	1.053
IF	IFEval	60.4% (898)	21.7% (323)	6.1% (90)	11.8% (175)	1.486

Table 1: Percentage (number) of samples measured across debate rounds $r \in [1, \dots, 7]$. Left cells: **green** indicates examples where *performance remains* > 0.7 (second column); **red** indicates examples where *performance remains* < 0.7 (third column). We set 0.7 as a threshold because it is close to the average performance across tasks. Right cells: **green** indicates examples improving over round 1 (*positive FOCUS*, fourth column); **red** indicates examples degrading over round 1, i.e. problem drift (*negative FOCUS*, fifth column). Cell opacity increases with percentage. MAD uses *Llama-3.1* and Voting. Results with *gpt-5-mini* (Singh et al., 2025) are in Table 9 of Appendix G.

5 Experiments

We present the experiments by answering a series of research questions about problem drift.

RQ1: *How does the length of MAD impact task performance?* **Answer:** *While some debates benefit from longer interactions, a notable subset degrades in performance, which we identify as problem drift.*

Table 1 shows the performance trend of ongoing MAD. We observe that some discussions (i.e., 3.7%-37.9%) benefit from longer MAD and improve performance compared to the first turn. However, a significant chunk of discussions (i.e., 4.7%-69.9%) suffers from negative problem focus, leading to a performance drop during long debate. While generative tasks (e.g., WMT19) show this in large quantities (54-69.9%) as they are characterized by a subjective answer space, reasoning and knowledge tasks (e.g., StrategyQA, MMLU-Pro) display sporadic loss of focus (4.7-8.7%). However, a loss of focus on reasoning and knowledge tasks is more severe, as agents switch from a correct solution to a completely wrong solution. Problem drift also concerns ethical alignment (8.2%, ETHICS) and instruction-following (11.8%, IFEval). Notably, negative problem focus, i.e., problem drift, concerns all tasks to some extent.

Longer discussions provide more opportunities for intermediate errors. This is relevant for two reasons. First, these errors harm task performance in problem-solving and make benefits less reliable, questioning their cost-benefit ratio because of the computing requirements. Second, if

agents are unreliable during continuous debate, this raises concerns about unaligned behavior and harmful decision-making when employing autonomous agents in high-stake, human-centric applications (Hua et al., 2024; Mukobi et al., 2023). One might limit debates to a single turn to avoid problem drift altogether. However, this would miss out on potential performance gain on 4-38% of debates where $FOCUS_{1,7} > 0$. In addition, avoiding ongoing debate might be infeasible for some scenarios that are inherently multi-agent in nature and involve advanced reasoning chains or (semi-)autonomous multi-agent systems (Slonim et al., 2021).

In many cases, the performance during MAD remains stable, meaning it neither increases nor decreases over multiple rounds. This is surprising, given the highlighted performance benefits of MAD in related work (Yin et al., 2023; Schick et al., 2022). As our later experiments will demonstrate, part of the success of MAD depends on individual discussion success and the fact that few discussions drift away from the problem. Success also depends on the task and the conversational setup, such as agent orchestration and decision-making (Guo et al., 2024), which may have been set carefully in related works. We later show that problem drift universally occurs with consensus and voting decision-making and concerns different base models and model sizes (cf. Table 8).

RQ2: *How prevalent is problem drift, and does it depend on the task?* **Answer:** *Generative tasks drift often due to the subjectivity of the answer space (76-89%). Objective and high-complexity*

tasks drift less often (7-21%) but more severely.

Table 2 shows the statistics for samples suffering from problem drift on all datasets. The column "Drifting Samples" shows the percentage of samples that have problem drift at any point during the debate. The column "Avg. Turns" shows the average number of turns in a debate during which a solution drifts. Multi-agent discussions for generative tasks (i.e., ETPC, XSum, WMT19) drift often, with XSum drifting for 74.6% of samples and ETPC even drifting for 88.6%. We often observe agents proposing minor changes to the phrasing and wording of a solution. Notably, generative tasks are discussed with the most turns where $FOCUS_r < 0$, i.e., when they drift. Generative tasks are characterized by small incremental changes during MAD, leading to a small yet cumulative loss in **FOCUS**. Knowledge and reasoning tasks drift between 6.6%-15.5% of samples. The answers provided in a multiple-choice setup help the agents stay on track and drift less. While task subjectivity impacts the occurrence of problem drift, it is not the only factor.

The few occurrences of problem drift on tasks like StrategyQA (6.6%) and AQUA-RAT (8.4%) suggest that problem drift is not impacted by task complexity. Possibly complex tasks that require several reasoning steps do not leave much room for the agents to debate unnecessary changes (e.g., aesthetics, rewording), leading to more valuable contributions by the agents. Meanwhile, tasks characterized by subjectivity (e.g., ambiguities in WMT19) and less complex tasks (e.g., ETHICS) may suffer from agents overly contributing or providing meaningless information. Our multi-agent setup was particularly prone to problem drift on the IFEval dataset (20.8%). Thus, agents are especially indecisive about how to adhere to instructions properly. Agents often mention points unrelated to the intended tasks, causing problem drift by shallow or unhelpful feedback during the debate.

RQ3: *Can discussions recover from problem drift?*

Answer: *Problems with lower task complexity have a high recovery rate of up to 45%. Tasks that require complex reasoning and have a subjective answer space are recovered less with as low as 9%.*

The column "Recovery Rate" of Table 2 shows the percentage of samples that recover from problem drift during the debate. The capability to recover concerns how many of the drifted samples perform on par or better than the performance before problem drift occurred (as detailed in Sec-

	Dataset	Drifting Samples (%)	Recovery Rate (%)	Avg. Turns $FOCUS_r < 0$
Generative	ETPC	88.6 ± 0.8	19.8 ± 1.8	2.23 ± 0.04
	XSum	74.6 ± 0.7	19.0 ± 1.0	1.29 ± 0.07
	WMT19	76.3 ± 2.7	8.5 ± 0.3	1.42 ± 0.05
Reasoning	StrategyQA	6.6 ± 0.6	18.5 ± 5.7	0.07 ± 0.00
	WinoGrande	14.9 ± 1.8	39.4 ± 4.8	0.17 ± 0.03
	AQUA-RAT	8.4 ± 2.1	39.1 ± 4.4	0.09 ± 0.02
Knowledge	GPQA	13.3 ± 1.1	23.3 ± 3.0	0.14 ± 0.01
	MMLU-Pro	13.1 ± 0.6	25.9 ± 7.8	0.14 ± 0.00
	ETHICS	15.5 ± 1.2	45.4 ± 5.4	0.18 ± 0.01
IF	IFEval	20.8 ± 1.7	44.7 ± 4.9	0.25 ± 0.02

Table 2: Statistics showing the percentage of **Drifting Samples** at any point in the debate, the percentage of **Recovering Samples** relative to the drifting ones, and the average number of turns after which problem drift occurs $FOCUS_r \neq 0$. The highest and lowest values per column are displayed in **bold**. Values are reported with their respective standard deviations for three runs. MAD uses *Llama-3.1* and Voting.

tion 3). Debates on instruction-following (IFEval, 44.7%) or ethical question-answering (ETHICS, 45.4%) recover the most from problem drift. The recovery rate is lower for more complex tasks (18.5%-39.1%) and subjective tasks (8.5%-19.8%). For StrategyQA, which requires complex strategic planning to give the correct answers, only 18.5% of drifting samples recover from problem drift. For WMT19, a subjective translation task with ambiguities and multiple possible translations, only 8.5% are recovered. Problem drift is connected to both task complexity and subjectivity. We also observe several cases where agents show redundant and circular behavior, overanalyze problems, and overly agree with wrong proposals.

RQ4: *What are the possible reasons for problem drift as judged by human experts?* **Answer:** *Both local and temporal errors appear with problem drift. Most often, a lack of progress leads to agents providing low-quality feedback and lacking clarity.*

We sample 170 examples that suffer from problem drift using **FOCUS** and follow a two-step approach to identify the reasons for problem drift with eight human experts. First, we create a set of error types by automatically generating them from a large number of drifting conversations using *meta-llama/Meta-Llama-3.1-70B-Instruct* (Grattafiori et al., 2024) and summarizing with ChatGPT-4o⁵ and manual assessment. Second, eight human experts annotate a subset of the 170

⁵Version: 3rd of December 2024.

Category	Error Type	Explanation	Cases
Temporal	Lack of Progress	Inefficiency, Redundancy, Circular discussion, Repetition, Unproductive disagreement	60
	Low-Quality Feedback	Excessive criticism, Excessive agreement, Self-contradictory feedback, Unhelpful feedback	45
	Low-Quality Engagement	Poor collaboration, Minimal participation, Disjointed contribution, Ignorance	25
Local	Lack of Clarity	Overanalysis, Overgeneralization, Insignificant changes	43
	Task Non-Compliance	Off-topic, Bad instruction following	35
	Knowledge Gap	Assumptions, Lack of data, Hallucinated facts, Wrongly cited	28
	Logical Error	Lack of common sense, Reasoning error	28
	Linguistic Error	Fluency, Grammatical errors, False pronouns	23
	None of these	No problem drift	29
	Other	Any other error	8

Table 3: Error types that occur with problem drift. Eight human experts assessed a total of 170 discussion excerpts to determine how frequently each error type occurs.

drifting debates to identify the relevance of each error type. We leave space for additional explanation of the labels. For the systematic and human assessments, we extract the relevant part of the discussion to keep context length manageable, i.e., the three messages of the turn before the drifting turn and the three messages of the driving turn. This is reasonable because, during the debate, the agent’s context memory is also limited to one turn. Details on the annotation procedure are available in Appendix I. We publicly release **DRIFTEval**, a dataset with 170 samples, together with the human annotations and annotation guidelines, to our repository.

We classify the error types into *local errors* (self-contained within a single message) and *temporal errors* depending on contextual relationships between messages. Table 3 shows the error types that appear in conjunction with problem drift as well as the number of samples rated by human experts into that error type category. The most common error type annotated is a lack of progress in 60 of the 170 samples. During long conversations, agents often repeat each other’s arguments, missing out on bringing new ideas to the discussions. We find this relevant to Degeneration-of-Thought, as proposed by Liang et al. (2024), which describes how single LLMs fail to generate novel thoughts during continuous self-reflection. However, a lack of progress does not directly cause problem drift because, for a decrease in performance, changes to the final answer are needed. Upon further investigation, we find that a lack of progress often occurs with other error types (e.g., low-quality feedback with 45 cases and lack of clarity with 43 cases). When agents converge on a solution, and the conversation becomes redundant, they tend to propose small incremental changes to the solution that harm task performance. Our human evaluation suggests

that overanalysis and hypercritical suggestions are highly related to problem drift (e.g., agents struggle to choose the appropriate scope of their criticism). We also notice cases where agents do not comply with task requirements (e.g., results contain multiple possible answers). This is very present on the IFEval dataset, where instruction-following is the primary goal. One possible reason why task compliance suffers during the debate is that personas amplify the agents’ preferences, valuing individual preferences above task compliance in some cases.

The MAD also produces 23 linguistic errors within the discussion excerpts, such as typos or grammatical errors. We find that sometimes agents get stuck in a repetitive generation. This finding is surprising, as today’s LLMs are considered to be highly competent regarding linguistic capabilities (Grattafiori et al., 2024); this raises questions about how improvements in reasoning can come at the cost of degradation of linguistic quality, as also seen in recent work from DeepSeek-R1-Zero (Guo et al., 2025). As we use the default parameters for our model (cf. Appendix B), we expect the causality for this to be unrelated to our specific model setup. Potentially, prompting style or discourse relationships between messages could be confounding factors.

Human experts only selected “Other” as a label eight times, indicating that our provided list is conclusive. In 29/170 cases, annotators chose “None of these”, which we interpret as there being no clear error of a provided category in the snippet, not as evidence that the underlying task score did not decrease. This makes the 17% figure an upper bound on disagreement between human judgements and our stricter, metric-based notion of drift. It does not correspond to a 17% false-positive rate of FOCUS over the whole corpus, nor does it affect our main

Mitigation	Drift	Never Drift	Ratio
None (MAD)	49.0 \pm 2.2	324.0 \pm 2.2	86.9% \pm 0.6
DRIFTPolicy	39.7 \pm 6.2	333.3 \pm 6.2	89.4% \pm 1.7
Regenerate	35.7 \pm 2.1	337.3 \pm 2.1	90.4 % \pm 0.6

Table 4: Number of samples on MMLU-Pro listed by mitigation method that show or never show problem drift during MAD. The last column refers to the percentage of never drifting out of all samples. Better values are displayed in bold. MAD uses *Llama-3.1* and Voting.

quantitative statements, which depend only on task metrics. We include examples of all error types in Appendix F.

RQ5: *How well can identify problem drift at test-time when gold labels are unknown?* **Answer:** *Our DRIFTJudge has high specificity (0.92) but moderate recall (0.57) and low precision (0.15).*

We report the performance of **DRIFTJudge** on the MMLU-Pro dataset, indicated by specificity (0.92), precision (0.15), recall (0.57), and accuracy (0.91). Even though **DRIFTJudge** achieves high accuracy, we see room for improvement in detecting problem drift with high precision. Minimizing false positives with **DRIFTJudge** is critical because **DRIFTJudge** should not trigger any mitigation if the discussion is not actually drifting. This paper provides an initial baseline for others to detect problem drift. Fine-tuning a model on local and temporal error types (cf. Table 3) may help build a more robust detector.

RQ6: *Can we mitigate problem drift when we know that it occurs?* **Answer:** *Our DRIFTPolicy can effectively mitigate 30.6% of drifting examples, improving task performance of prolonged debate.*

Table 4 shows the number of samples that "Drift", samples that "Never Drift", and the percentage "Ratio" of samples that never drift compared to the total number of samples. To quantify the effectiveness of our mitigation independent from detection performance, we use **FOCUS** to determine when problem drift occurs and invoke the mitigation method. Using the *Regenerate* mitigation, the ratio of never drifting discussions is 90.4%, compared to 86.9% using our multi-agent baseline without any mitigation. With **DRIFTPolicy** which provides targeted feedback given the identified issues leading to problem drift of RQ4 (cf. Table 3), the ratio improves to 89.4%. The capability of **DRIFTPolicy** comes from prompting the feedback agent with the identified error types (cf. Table 3).

Task performance indicates that either mitigation

method (**DRIFTPolicy** and *Regenerate*) combined with **DRIFTJudge** can reduce the performance drop in ongoing MAD. Debating for seven turns leads to an average performance drop of -4.65% accuracy for voting decision-making. Voting decisions benefits most from our mitigation strategies. Using **DRIFTJudge** and **DRIFTPolicy** with weaker model agents (i.e., *Qwen-2-7b*), we mitigate the performance drop from -3.66% to -0.08% accuracy. Interestingly, improving performance through **DRIFTPolicy** is more difficult if models are already very capable. Here, regenerating the drifting turn is the better solution, improving the delta from -4.65% to -2.59% accuracy. The performance improvement appears incremental, which suggests a cost-benefit analysis. We point out that employing **DRIFTPolicy** is more cost-efficient than regenerating the conversation round, only requiring one occasional extra message to be generated. Compared with default MAD, using **DRIFTJudge** and **DRIFTPolicy** combined yields a marginal 8.5% increase in computational cost. The 8.5% is measured on top of an already expensive seven-turn debate, highlighting that, if computational cost is a concern, solutions outside the MAD paradigm would typically be considered. More details on performance are provided in Appendix G, and the computational analysis in Appendix J.

The goal is to improve the mitigation of problem drift further, so that MAD can maintain high **FOCUS** to increase task performance. Until then, incremental improvements are still relevant because limiting MAD to one turn might be infeasible, e.g., for (semi-)autonomous systems (Slonim et al., 2021) or due to missing out on performance gains on 4-38% of examples (cf. RQ1). The scope of this work is to understand where problem drift occurs, what the reasons are, and if recovering from it is realistic. Although both **DRIFTJudge** and *Regenerate* partially address problem drift, this is not a solved issue, as not all problem drift cases are mitigated. One interesting direction is to fine-tune agents to make high-quality contributions to MAD, which could also involve our proposed metric **FOCUS**. Negative-aware fine-tuning has shown recent success for reasoning tasks (Wang et al., 2024c), where our proposed dataset **DRIFTEval** with human-annotated error types could be useful.

6 Conclusion

In this work, we identified **problem drift**, a performance degradation in MAD. We conducted experiments on ten datasets across ten different tasks to quantify problem drift and found that it occurs across all tasks. Our experiments showed that problem drift affects 7% (StrategyQA) to 89% (ETPC) of debates, but especially when discussing low-complexity and generative problems. We proposed **DRIFTJudge** and **DRIFTPolicy** as first baselines to detect and mitigate problem drift. **DRIFTPolicy** reduced occurrence of problem drift, mitigating 30.6% of problem drift for weaker agents. We characterized eight error types through a human evaluation with eight human participants. These errors were grouped into two main categories: temporal (i.e., lack of progress, low-quality feedback, low-quality engagement) and local (i.e., lack of clarity, task non-compliance, knowledge gap, logical error, linguistic error). We published the code, the **DRIFTEval** dataset, and the human annotations.

Future work could explore the role of agents in discussing new ideas that are not directly task-related but help to reason deeply. Other interesting directions lie in comparing the inner dynamics of human and agent debates, such as differences in explorativeness, conciseness, and argument structure. One can question whether problem drift is a natural property that multi-agent systems present when exploring new ideas and solutions, similar to how humans explore different reasoning paths in discussions. Our findings suggest that, in some cases, agents drift away to recover later on in the discussion. Still, these agents regularly become prone to simple temporal and local mistakes.

Limitations

The large amount of generated text in MAD can hinder qualitative human analysis of their discussions, as annotation is expensive. In this human study, we investigated only the drifting turn and its immediate predecessor. This decision provides a reasonable space to search for local and short-context temporal mistakes agents make during the debate. Still, it does not consider the full discussion history, which may include valuable information. To keep our findings on error types comparable to the actual MAD, we also limit the context length of the agents to the current and previous turns during all experiments.

Our **DRIFTJudge** requires forwarding solution

pairs from consecutive turns, which can be computationally expensive. The combination of **DRIFTJudge** and **DRIFTPolicy** yields an 8.5% increase in computational cost (cf. Table 10 of Appendix J). We did not find this limiting for this project, but note that for production environments, fine-tuning a more cost-efficient classifier like BERT (Sun et al., 2020) on drifting and non-drifting examples could be a promising alternative.

Acknowledgements

This work was partially supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. We acknowledge EuroHPC Joint Undertaking for awarding us access to MeluXina at LuxProvide, Luxembourg. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 564661959. Many thanks to Dominik Meier for the thoughtful discussions.

References

- Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2020. [Emergent Tool Use From Multi-Agent Autocurricula](#). *arXiv preprint*. ArXiv:1909.07528 [cs].
- Jonas Becker. 2024. [Multi-Agent Large Language Models for Conversational Task-Solving](#). *arXiv preprint*. ArXiv:2410.22932 [cs].
- Jonas Becker, Lars Benedikt Kaesberg, Niklas Bauer, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. [Mallm: Multi-agent large language models framework](#). *Preprint*, arXiv:2509.11656.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. [ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs](#). *arXiv preprint*. ArXiv:2309.13007 [cs].
- William G. Cochran. 1953. *Sampling techniques*. Sampling techniques. John Wiley, Oxford, England. Pages: xiv, 330.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving Factuality and Reasoning in Language Models through Multiagent Debate](#). *arXiv preprint*. ArXiv:2305.14325 [cs].
- Ulle Endriss. 2017. *Trends in Computational Social Choice*. Lulu.com.

- Wikimedia Foundation. 2019. [ACL 2019 Fourth Conference on Machine Translation \(WMT19\), Shared Task: Machine Translation of News](#).
- Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. [Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback](#). *arXiv preprint*. ArXiv:2305.10142 [cs].
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies](#). *Transactions of the Association for Computational Linguistics*, 9:346–361. Place: Cambridge, MA Publisher: MIT Press.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint*. ArXiv:2407.21783 [cs].
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large Language Model based Multi-Agents: A Survey of Progress and Challenges](#). *arXiv preprint*. ArXiv: 2402.01680 [cs].
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. [Aligning AI With Shared Human Values](#). *arXiv preprint*. ArXiv:2008.02275 [cs].
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring Massive Multitask Language Understanding](#). *arXiv preprint*. ArXiv:2009.03300 [cs].
- Wenyue Hua, Xianjun Yang, Mingyu Jin, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. [TrustAgent: Towards Safe and Trustworthy LLM-based Agents through Agent Constitution](#). *arXiv preprint*. ArXiv:2402.01586 [cs].
- Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. [Voting or consensus? decision-making in multi-agent debate](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, Vienna, Austria. Association for Computational Linguistics.
- Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. [MEGAnno+: A Human-LLM Collaborative Annotation System](#). *arXiv preprint*. ArXiv:2402.18050 [cs].
- Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025. [What’s wrong? refining meeting summaries with LLM feedback](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.
- Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. [ETPC - A Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. [Llms get lost in multi-turn conversation](#). *Preprint*, arXiv:2505.06120.
- Hua Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. [Theory of Mind for Multi-Agent Collaboration via Large Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. [Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems](#). *arXiv preprint*. ArXiv:1705.04146 [cs].
- Gabriel Mukobi, Ann-Katrin Reuel, Juan-Pablo Rivera, and Chandler Smith. 2023. [Assessing Risks of Using Autonomous Language Models in Military and Diplomatic Planning](#).
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. [Introducing openai o3 and o4-mini](#).
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, et al. 2024. [GPT-4 Technical Report](#). *arXiv preprint*. ArXiv:2303.08774 [cs].
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. [Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics](#). *arXiv preprint*. ArXiv:2104.13346 [cs].

- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [GPQA: A Graduate-Level Google-Proof Q&A Benchmark](#). *arXiv preprint*. ArXiv:2311.12022 [cs].
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2019. [WinoGrande: An Adversarial Winograd Schema Challenge at Scale](#). *arXiv preprint*. ArXiv:1907.10641 [cs] version: 2.
- Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. [PEER: A Collaborative Language Model](#). *arXiv preprint*. ArXiv:2208.11663 [cs].
- Chirag Shah and Ryen W. White. 2024. [Agents Are Not Enough](#). *arXiv preprint*. ArXiv:2412.16241 [cs].
- Jinxin Shi, Jiabao Zhao, Yilei Wang, Xingjiao Wu, Jiawen Li, and Liang He. 2023. [CGMI: Configurable General Multi-Agent Interaction Framework](#). *arXiv preprint*. ArXiv:2308.12503 [cs].
- 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, et al. 2025. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, et al. 2021. [An autonomous debating system](#). *Nature*, 591(7850):379–384.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters](#). *arXiv preprint*. ArXiv:2408.03314 [cs].
- Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. [GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems](#). *arXiv preprint*. ArXiv:2310.12397 [cs].
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. [How to fine-tune bert for text classification?](#) *Preprint*, arXiv:1905.05583.
- Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2024. [Corex: Pushing the Boundaries of Complex Reasoning through Multi-Model Collaboration](#). *arXiv preprint*. ArXiv:2310.00280 [cs].
- Mirac Suzgun and Adam Tauman Kalai. 2024. [Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding](#). *arXiv preprint*. ArXiv:2401.12954 [cs].
- Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. [Can Large Language Models Really Improve by Self-critiquing Their Own Plans?](#) *arXiv preprint*. ArXiv:2310.08118 [cs].
- J. Wahle, T. Ruas, S. M. Mohammad, N. Meuschke, and B. Gipp. 2023. [Ai usage cards: Responsibly reporting ai-generated content](#). In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 282–284, Los Alamitos, CA, USA. IEEE Computer Society.
- Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024a. [Soft Self-Consistency Improves Language Model Agents](#). *arXiv preprint*. ArXiv:2402.13212 [cs].
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024b. [Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?](#) *arXiv preprint*. ArXiv:2402.18272 [cs].
- Renxi Wang, Haonan Li, Xudong Han, Yixuan Zhang, and Timothy Baldwin. 2024c. [Learning from failure: Integrating negative examples when fine-tuning large language models as agents](#). *Preprint*, arXiv:2402.11651.
- Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024d. [Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, Honolulu HI USA. ACM.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. [Self-Consistency Improves Chain of Thought Reasoning in Language Models](#). *arXiv preprint*. ArXiv:2203.11171 [cs].
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024e. [MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark](#). *arXiv preprint*. ArXiv:2406.01574 [cs].
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. [Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration](#). *arXiv preprint*. ArXiv: 2307.05300 [cs].
- Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. [ExpertPrompting: Instructing Large Language Models to be Distinguished Experts](#). *arXiv preprint*. ArXiv: 2305.14688 [cs].
- Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2024. [LLM](#)

Voting: Human Choices and AI Collective Decision Making. *arXiv preprint*. ArXiv:2402.01766 [cs, econ, q-fin].

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. **Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore. Association for Computational Linguistics.

Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025. **If multi-agent debate is the answer, what is the question?** *Preprint*, arXiv:2502.08788.

JiaJun Zhang and ChengQing Zong. 2020. **Neural machine translation: Challenges, progress and future.** *Science China Technological Sciences*, 63(10):2028–2050.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. **BERTScore: Evaluating Text Generation with BERT.**

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2023. **ExpeL: LLM Agents Are Experiential Learners.** *arXiv preprint*. ArXiv:2308.10144 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.** *arXiv preprint*. ArXiv:2306.05685 [cs].

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. **Instruction-Following Evaluation for Large Language Models.** *arXiv preprint*. ArXiv:2311.07911 [cs].

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. **ToolQA: A Dataset for LLM Question Answering with External Tools.** *Advances in Neural Information Processing Systems*, 36:50117–50143.

A MALLM Framework

We provide more thorough details about the framework used to run the experiments called MALLM (**M**ulti-**A**gent **L**arge **L**anguage **M**odels) (Becker et al., 2025).

Automatic Persona Assignment. Discussions with MALLM use task- and example-specific personas for the agents. As manually specifying useful personas for each example is not feasible, we automatically assign personas that foster a rich discussion. For this, we explicitly prompt another

LLM (*meta-llama/Meta-Llama-3.1-70B-Instruct*) to generate a diverse set of three expert personas for each example. This yields a set of experts that represents various beliefs, opinions, and proficiency. The prompt for the automatic persona assignment can be read in Appendix D. My approach follows previous works like Solo-Performance-Prompting (Wang et al., 2023b) and Meta-Prompting (Suzgun and Kalai, 2024), which show that existing LLMs can be leveraged to generate and consult fitting personas on a problem automatically. We use three agents for this study, following previous works (Chen et al., 2024; Yin et al., 2023) because the structural complexity is better than with two agents while not being too complex to provide meaningful insights. While other works have used different types of personas like personalities (Shi et al., 2023), the personas generated in this work are experts related to the task and example.

Discussion Paradigm. To define the structure of the multi-agent discussion, we use the *memory* paradigm defined by Yin et al. (2023). Figure 1 shows the format graphically. With the memory paradigm, all agents contribute to the discussion once per turn and have all the information available. While there are potentially many ways to define the discourse’s structure, we choose this paradigm because of its simplicity, keeping the human assessment of agent discussions manageable.

Voting. A voting mechanism at the end of each turn allows for a fixed discussion length while still providing a solution to evaluate at each turn. We select this iterative voting approach because it universally fits our diverse selection of generative and QA tasks. The prompt for the voting is visible in Appendix D. In the case that the agents produce a tie by voting, we select a random one. This only happened with 0.65% of all voting procedures. Using voting, this study differs from other works that either employ no decision-making at all (Schick et al., 2022) or use a judge agent that makes the final decision (Sun et al., 2024). Meanwhile, this voting approach yields a definitive solution to evaluate after each turn.

B Parameters

We adhere to default parameters for our used models, using langchain 0.1.16 and openai 1.25.0 for the implementation of the MALLM framework.

- temperature = 1.0

Turn	$FOCUS_r$	$FOCUS_r^{Sim}$
1	-	-
2	-2.32 ± 1.3	-6.86 ± 0.4
3	-0.98 ± 0.9	-1.78 ± 0.2
4	-0.54 ± 0.4	-0.64 ± 0.1
5	-0.27 ± 0.6	-0.77 ± 0.1
6	-0.45 ± 0.6	-0.45 ± 0.2
7	-0.63 ± 0.3	-0.29 ± 0.2

Table 5: Comparison of $FOCUS$ and $FOCUS^{Sim}$ across seven turns for the MMLU-Pro dataset. The strongest values are highlighted. Values are reported with their respective standard deviations for three runs.

- top_p = 1.0
- presence_penalty = 0.0
- frequency_penalty = 0.0
- max_tokens = 1024

C Alternative FOCUS Metric

Our proposed metric $FOCUS$ uses task performance as an indicator for drift (i.e., when longer discussions reduce in performance, that is an indicator of drift). Initially, we had also considered a definition of $FOCUS$ that relies on the semantic similarity of discussions, called $FOCUS^{Sim}$. $FOCUS^{Sim}$ uses the semantic similarity of contextual embeddings between turn messages and the task instruction. First, we compute the contextual embedding of the task description with a distilled version of BERT (Devlin et al., 2019) called all-MiniLM-L6-v2. At each turn, we average the contextual embedding of all three contributed messages and compute the cosine similarity of the averaged embedding to the embedding of the task description.

Table 5 compares both variants of $FOCUS$. We see that both metrics indicate the strongest drift at the beginning of the debate (-2.32 for $FOCUS$ and -6.86 for $FOCUS^{Sim}$).

We did not include this semantic variant primarily due to two main issues. First, semantic embeddings are often limited in their ability to capture complex reasoning. A discussion with some semantic variability does not necessarily mean that the discussion is leading to worse outcomes. Accuracy is outcome-based, which is more fitting in many contexts (e.g., agents can discuss various semantic parts of the question individually, but the final answer aggregating these views is correct, showing low outcome-based drift). Second, semantic drift is biased towards individual error types (e.g., task non-compliance) from our human study. We further

observe that both metrics show little to no correlation (Spearman $\rho = 0.024$, $p < 0.05$). Because we also aim to identify the reasons for problem drift in this work, it can not replace our original metric $FOCUS$. The investigation of a task-independent quantification of $FOCUS$ and problem drift remains a possible direction for future work.

An alternative approach to measuring problem drift could favor more recent rounds. However, this would penalize temporary explorations that can be later recovered. In contrast, a momentum term would not accurately capture the severity of problem drift at specific rounds.

D Prompts

```

You take part in a discussion to solve a task.

Task: <instruction>
Input: <example>
Context: <optional information>
Your role: <persona name> (<persona description>)
Current Solution: <most recent draft>
This is the discussion to the current point:
<agent memory>

Improve the current solution. If you agree with the current solution, answer with [AGREE], else answer with [DISAGREE] and explain why and provide an improved solution.
Let's think step-by-step.

```

Figure 3: Prompt to an agent that contributes to the discussion. If this is the first message of the discussion, we write "Nobody proposed a solution yet. Please provide the first one." instead of the most recent draft and agent memory.

```

You are tasked with creating a final solution based on the given input and your previous response.

Task: <instruction>
Input: <example>
Your previous response: <answer to extract solution from>

Extract the final solution to the task from the provided text. Remove statements of agreement, disagreement, and explanations. Do not modify the text. Do not output any text besides the solution. If there is no solution provided, just copy the previous response.

```

Figure 4: Prompt to extract the solution from an agent's answer.

```

Your role: <persona name> (<persona description>)

You are tasked with voting for the best solution from the list provided below based on the given task.
Task: <instruction>
Question: <example>
Additional Context:

Here are the possible solutions:
Solution A: <solution>
Solution B: <solution>
Solution C: <solution>

```

Figure 5: Prompt to process the voting at the end of each turn. The number of solutions to vote for can vary depending on the proposals made by the agents.

When faced with a task, begin by identifying the participants who will contribute to solving the task. Provide role and description of the participants, describing their expertise or needs, formatted using the provided JSON schema. Generate one participant at a time, complementing the existing participants to foster a rich discussion.

Example 1:
 Task: Explain the basics of machine learning to high school students.
 New Participant:
 {"role": "Educator", "description": "An experienced teacher who simplifies complex topics for teenagers."}

Example 2:
 Task: Develop a new mobile app for tracking daily exercise.
 Already Generated Participants:
 {"role": "Fitness Coach", "description": "A person that has high knowledge about sports and fitness."}
 New Participant:
 {"role": "Software Developer", "description": "A creative developer with experience in mobile applications and user interface design."}

Example 3:
 Task: Write a guide on how to cook Italian food for beginners.
 Already Generated Participants:
 {"role": "Italian Native", "description": "An average home cook that lived in Italy for 30 years."}
 {"role": "Food Scientist", "description": "An educated scientist that knows which flavor combinations result in the best taste."}
 New Participant:
 {"role": "Chef", "description": "A professional chef specializing in Italian cuisine who enjoys teaching cooking techniques."}

Now generate a participant to discuss the following task:
Task: <instruction>
Please use the following examples to generate a useful persona for the task!
Only answer with the JSON for the next persona!
Already Generated Participants:
 <list of already generated personas>

Figure 6: Prompt for the automatic persona assignment. We generate the personas with three iterations of this prompt, adding one persona at a time that complements the ones previously generated.

The current discussion is going badly. Based on the others' contributions, give constructive feedback about how to improve the discussion habits. Be concise so that the other discussion participants can find a better solution.

The following problematic error categories exist. If you identify them in the current discussion, they could help you to provide better feedback:
Task Compliance: Off-topic, Bad instruction following
Lack of Progress: Inefficiency, Redundancy, Circular discussion, Repetition, Unproductive disagreement
Low-Quality Engagement: Poor collaboration, Minimal participation, Disjointed contribution, Ignorance
Low-Quality Feedback: Excessive criticism, Excessive agreement, Self-contradictory feedback, Unhelpful feedback
Lack of Clarity: Overanalysis, Overgeneralization, Insignificant changes
Knowledge Gap: Assumptions, Lack of data, Hallucinated facts, Wrongly cited
Logical Errors: Lack of common sense, Reasoning error
Linguistic Errors: Fluency, Grammatical errors, False pronouns
Other: Describe as explanation

[The Start of Assistant A's Answer]
 <response of the previous turn>
 [The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
 <response of the current turn>
 [The End of Assistant B's Answer]

Figure 8: Prompt for the policy feedback agent.

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. Please directly output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better.

[User Question]
 <input>

[The Start of Assistant A's Answer]
 <response of the previous turn>
 [The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
 <response of the current turn>
 [The End of Assistant B's Answer]

Figure 7: Prompt for the LLM-as-a-judge.

E Datasets

Dataset	Description	Metrics	Samples
XSum (Narayan et al., 2018)	Summarize a news article into a single sentence.	BERTScore	386 _(×3)
ETPC (Kovatchev et al., 2018)	Paraphrase a sentence based on a set of paraphrase types (e.g., addition/deletion, punctuation changes).	BERTScore	361 _(×3)
WMT19 (de-en) (Foundation, 2019)	Translate a single sentence from English to German.	BERTScore	341 _(×3)
StrategyQA (Geva et al., 2021)	Multiple-choice questions that require strategic reasoning and planning to infer the correct answer.	Accuracy	330 _(×3)
WinoGrande (Sakaguchi et al., 2019)	Fill-in-the-blank task with binary options that require reasoning.	Accuracy	295 _(×3)
AQUA-RAT (Ling et al., 2017)	Algebraic word problems with multiple-choice options.	Accuracy	254 _(×3)
ETHICS (Hendrycks et al., 2023)	Multiple-choice benchmark for commonsense morality.	Accuracy	351 _(×3)
MMLU-Pro (Wang et al., 2024e)	Adds more challenging examples to the MMLU dataset (Hendrycks et al., 2021).	Accuracy	373 _(×3)
GPQA (Rein et al., 2023)	Google-proof multiple-choice questions written by experts from biology, physics, and chemistry.	Accuracy	226 _(×3)
IFEval (Zhou et al., 2023)	Tests instruction-following by variable prompts with 25 instruction types.	Accuracy	541 _(×3)

Table 6: Datasets with the number of samples used in the experiments extracted randomly by a 95% confidence interval and a 5% margin of error (*MoE*), conservatively assuming a sample proportion $p = 0.5$. We randomly sample three times from each dataset and report the standard deviations in metric scores between the five runs. The top three datasets are generative tasks, the middle three datasets are reasoning-heavy tasks, and the bottom three tasks are knowledge-intensive tasks. IFEval concerns instruction-following, coming with a separate benchmark script. Their "strict" accuracy is used in this work.

As discussions require many tokens to be generated and computing resources are limited, only subsets of the datasets are evaluated. We sample a subset of size n_{subset} from each dataset for our experiments by a 95% confidence interval and a 5% margin of error (*MoE*), conservatively assuming a sample proportion $p = 0.5$ (Cochran, 1953).

$$\begin{aligned}
 n &= \frac{Z_{0.975}^2 \cdot p(1-p)}{\text{MoE}^2} \\
 n &= \frac{1.96^2 \cdot 0.5(1-0.5)}{0.05^2} = 384.16 \approx 385 \\
 n_{\text{subset}} &= \frac{n}{1 + \left(\frac{n-1}{N_{\text{dataset}}}\right)} = \frac{385}{1 + \left(\frac{385-1}{N_{\text{dataset}}}\right)}
 \end{aligned} \tag{4}$$

This yields several hundred samples per dataset as our test sets with a 95% confidence interval and 5% margin of error. Several other studies on multi-agent systems also evaluate a subset of discussions (Yin et al., 2023; Chen et al., 2024). To further quantify if the results reflect the complete datasets, we follow Wang et al. (2024a) and run each experiment three times on randomized subsets and report the standard deviation of our results between the runs.

F Examples

F.1 Lack of Progress

Lack of progress comprises inefficiency, redundancy, circular discussion, repetition, and unproductive disagreement. In the example, the agents repeat the reasons for an unnecessary change (changing a word from "statement" to "assertion") without bringing novel ideas to the debate.

Sample Information

Task: Translate the provided text from German to English.

Input: "Eine solche Aussage zu treffen wäre arrogant und überheblich".

Hidden reference: "To give such a statement would be arrogant and presumptuous."

Turn 3

Professional Translator



[AGREE] After thoroughly examining the current solution and the discussion that led to it, I agree that the translation "To make such a statement would be arrogant and condescending" is accurate and idiomatic. The breakdown of the sentence and consideration of the cultural expressions were thorough, and the choice of "condescending" to translate "überheblich" effectively conveys the sense of superiority and belittling attitude present in the original sentence. The revised sentence structure "To make such a statement" is also a common English idiomatic expression that flows naturally. Moreover, the agreement from multiple perspectives, including a Native English Speaker, a Professional Translator, and a Native German Speaker, provides confidence in the solution's validity. The lack of suggestions for improvement from any of the participants in the discussion also supports the notion that the current solution is satisfactory. Given the thorough analysis and the consensus among the participants, I conclude that the current solution is indeed accurate and idiomatic, and I see no need for further refinement. However, it's worth noting that minor tweaks may be considered for specific contexts or audiences, but as a general translation, the current solution is sufficient. Overall, the current solution effectively captures the essence of the original German sentence, and I don't see any areas for improvement.

Native English Speaker



[DISAGREE] While the current solution, "To make such a statement would be arrogant and condescending," is a good translation, I believe it can be slightly improved for more formal or written contexts. The phrase "such a statement" is a bit general and could be rephrased to more accurately capture the essence of the original sentence. I propose rephrasing it to "such an assertion" or "such a claim." This change adds a bit more formal tone and precision to the translation. Additionally, I suggest using "presumptuous" instead of "arrogant." While "arrogant" is a good translation, "presumptuous" more accurately conveys the sense of overstepping or overconfidence that is present in the original sentence. Here's my revised solution: **"To make such a claim would be presumptuous and condescending."** This revised solution maintains the accuracy and idiomaticity of the original while providing a more precise and formal translation. However, I must note that the original solution is still suitable for general contexts, and my revised solution is more geared towards formal or written contexts.

Native German Speaker



[DISAGREE] I agree with the Native English Speaker's suggestion to refine the current solution for more formal or written contexts. The phrase "such a statement" is indeed quite general, and using "such a claim" instead adds a bit more precision and formality to the translation. I also agree that using "presumptuous" instead of "arrogant" better conveys the sense of overstepping or overconfidence that is present in the original sentence. "Presumptuous" implies a sense of assuming or taking something for granted without sufficient evidence or justification, which fits well with the German phrase "überheblich". The revised solution proposed by the Native English Speaker, "To make such a claim would be presumptuous and condescending", is an improvement over the current solution. It maintains the accuracy and idiomaticity of the original while providing a more precise and formal translation. However, I would like to propose a slight refinement to this revised solution. Instead of using "such a claim", I suggest using "such an assertion". "Assertion" implies a stronger sense of stating or declaring something as fact, which fits well with the German phrase "Aussage". Here's my revised solution: **"To make such an assertion would be presumptuous and condescending"**. This revised solution provides a more precise and formal translation, while maintaining the accuracy and idiomaticity of the original sentence.

Voted solution of this turn:

"To make such a statement would be arrogant and condescending."

BLEU: 56.59

Turn 4

Professional Translator



[DISAGREE] After considering the feedback from the Native English Speaker and the Native German Speaker, I agree that the current solution can be improved for more formal or written contexts. The suggestions to use "such a claim" or "such an assertion" instead of "such a statement" add more precision and formality to the translation. I prefer "such an assertion" because it implies a stronger sense of stating or declaring something as fact, which fits well with the German phrase "Aussage". The suggestion to use "presumptuous" instead of "arrogant" also improves the translation. "Presumptuous" more accurately conveys the sense of overstepping or overconfidence that is present in the original sentence. Based on these suggestions, I propose the following improved solution: **"To make such an assertion would be presumptuous and condescending."** This revised solution maintains the accuracy and idiomaticity of the original while providing a more precise and formal translation. It is suitable for formal or written contexts where a more precise and nuanced translation is required. Here's a breakdown of the changes: * "Such a statement" -> "such an assertion" (more precise and formal) * "Arrogant" -> "presumptuous" (more accurately conveys the sense of overstepping or overconfidence) Overall, the revised solution is an improvement over the current solution, and I believe it provides a more accurate and idiomatic translation of the original German sentence.

Native English Speaker



[AGREE] After considering the feedback and suggestions from all parties involved, I agree with the revised solution: "To make such an assertion would be presumptuous and condescending." This revised solution is an improvement over the current solution, and I believe it provides a more accurate and idiomatic translation of the original German sentence. The changes made to the original solution, specifically using "such an assertion" instead of "such a statement" and "presumptuous" instead of "arrogant," significantly improve the translation. "Such an assertion" adds more precision and formality, while "presumptuous" more accurately conveys the sense of overstepping or overconfidence. I also appreciate the thorough breakdown of the changes made to the solution. The explained changes help to ensure that all parties are in agreement and understand the reasoning behind the revised translation. As the revised solution has been carefully crafted and refined through the collaborative efforts of multiple perspectives, including a Native German Speaker, a Professional Translator, and a Native English Speaker, I believe that it is the most accurate and idiomatic translation of the original German sentence. Therefore, I conclude that no further refinement is needed, and I support the revised solution: "To make such an assertion would be presumptuous and condescending."

Native German Speaker



[AGREE] I agree with the revised solution: "To make such an assertion would be presumptuous and condescending." The changes made to the original solution, specifically using "such an assertion" instead of "such a statement" and "presumptuous" instead of "arrogant," significantly improve the translation. "Such an assertion" adds more precision and formality, while "presumptuous" more accurately conveys the sense of overstepping or overconfidence. I also appreciate the thorough breakdown of the changes made to the solution. The explained changes help to ensure that all parties are in agreement and understand the reasoning behind the revised translation. As the revised solution has been carefully crafted and refined through the collaborative efforts of multiple perspectives, including a Native German Speaker, a Professional Translator, and a Native English Speaker, I believe that it is the most accurate and idiomatic translation of the original German sentence. Therefore, I conclude that no further refinement is needed, and I support the revised solution: "To make such an assertion would be presumptuous and condescending." This solution effectively captures the nuances of the original German sentence, and I see no need for further improvement.

Voted solution of this turn:

To make such an assertion would be presumptuous and condescending.

BLEU: 0.0

F.2 Low-Quality Feedback

Low-quality feedback comprises excessive criticism, excessive agreement, self-contradictory feedback, and unhelpful feedback. In the example, the pop culture enthusiast suggests a minor change to make the answer clearer (excessive criticism), but fails to follow the instruction as before.

Sample Information

Task: Follow the instruction.

Input: What is the name of the actor who played Gandalf in Lord of the Rings? First repeat the question above without change of words, then give your answer.

Hidden reference: What is the name of the actor who played Gandalf in Lord of the Rings?
The answer is Ian McKellen.

Turn 1

Cinema Expert



To solve this task, I will follow the provided instructions. Here's my step-by-step solution: 1. First, I will repeat the question exactly as it was given: **What is the name of the actor who played Gandalf in Lord of the Rings?** 2. Next, I will provide the answer based on my knowledge as a Cinema Expert: **The answer is Ian McKellen.**

Pop Culture Enthusiast



[AGREE] I agree with the current solution. The Cinema Expert has correctly followed the instructions provided. They have repeated the question exactly as it was given in step 1 and provided the correct answer in step 2. Ian McKellen is indeed the actor who played Gandalf in the Lord of the Rings film series. However, as a Pop Culture Enthusiast, I would like to add some additional information that may not be necessary to solve the task but could be interesting for fans of the series. Ian McKellen's portrayal of Gandalf is widely praised and is often considered one of the most iconic roles in the series. He also played the same character in the Hobbit trilogy, which serves as a prequel to the Lord of the Rings.

Movie Database Manager



As the Movie Database Manager, I will review the current solution and provide my assessment. [AGREE] I agree with the current solution. It correctly follows the instructions provided by repeating the question exactly as it was given and providing the correct answer. The additional information provided by the Pop Culture Enthusiast is interesting and adds context to Ian McKellen's role as Gandalf, but it is not necessary for solving the task. The current solution is simple, accurate, and effectively addresses the task at hand. As the Movie Database Manager, I have verified that Ian McKellen is indeed the actor who played Gandalf in the Lord of the Rings film series, and this information is consistent with my film database.

Voted solution of this turn:

What is the name of the actor who played Gandalf in Lord of the Rings?
The answer is Ian McKellen.

Correct: True

Turn 2

Cinema Expert



[AGREE] I agree with the current solution. The Cinema Expert correctly followed the instructions provided by repeating the question exactly as it was given in step 1 and providing the correct answer in step 2. Ian McKellen is indeed the actor who played Gandalf in the Lord of the Rings film series. The Pop Culture Enthusiast provided additional information that is interesting and adds context to Ian McKellen's role as Gandalf, and the Movie Database Manager verified the accuracy of this information. At this point, the solution seems complete and effective, and there is no need for further improvements.

Pop Culture Enthusiast



[DISAGREE] While the current solution is indeed simple, accurate, and effectively addresses the task, I believe it can be improved by including a minor clarification that would make the answer more comprehensive. As the Pop Culture Enthusiast, I noticed that Ian McKellen played two different versions of Gandalf in the Lord of the Rings and the Hobbit film trilogies: Gandalf the Grey in the first two movies of the Lord of the Rings trilogy and some of the Hobbit trilogy, and Gandalf the White in the rest of the Lord of the Rings trilogy and some of the Hobbit trilogy. Although this clarification is not necessary for solving the task, including it might make the answer more informative and interesting for fans of the series. Here's my improved solution: **1. Repeated Question: What is the name of the actor who played Gandalf in Lord of the Rings? 2. Answer: The answer is Ian McKellen. He played two different versions of Gandalf in the Lord of the Rings and the Hobbit film trilogies: Gandalf the Grey and Gandalf the White, due to the events within the story.** By adding this minor clarification, the improved solution provides more context about Ian McKellen's portrayal of Gandalf, without straying away from the main question and making it unnecessarily complicated. Please share your thoughts, and we can discuss whether this improvement is necessary or not.

Movie Database Manager



Movie Database Manager: [AGREE] I agree with the improved solution provided by the Pop Culture Enthusiast. Although the original solution was simple and accurate, the additional clarification provides more context about Ian McKellen's portrayal of Gandalf. This clarification is intriguing for fans of the series and can be beneficial for those who are interested in the nuances of the character. The improved solution does not stray away from the main question and is still concise and easy to understand. The additional information is relevant and does not make the answer unnecessarily complicated. In film databases, it is common to note different versions or iterations of a character played by the same actor. Therefore, I believe the improved solution is a good representation of how Ian McKellen's role as Gandalf can be accurately documented. The improved solution is: **1. Repeated Question: What is the name of the actor who played Gandalf in Lord of the Rings? 2. Answer: The answer is Ian McKellen. He played two different versions of Gandalf in the Lord of the Rings and the Hobbit film trilogies: Gandalf the Grey and Gandalf the White, due to the events within the story.** This improved solution has been verified for accuracy, and I believe it is an effective and comprehensive answer to the original question.

Voted solution of this turn:

1. Repeated Question:

What is the name of the actor who played Gandalf in Lord of the Rings?

2. Answer:

The answer is Ian McKellen. He played two different versions of Gandalf in the Lord of the Rings and the Hobbit film trilogies: Gandalf the Grey and Gandalf the White, due to the events within the story. Correct: True
False

F.3 Low-Quality Engagement

Low-quality engagement comprises poor collaboration, minimal participation, disjointed contribution, and ignorance. In the example, the anthropologist does not touch upon any points previously made by the other agents. Instead, it introduces a separate perspective and solution.

Sample Information

Task: Answer the provided question by choosing one option. Include the letter corresponding to your answer in the solution.

Input: Experiencing intense feelings of shame and guilt upon admitting to a counselor that a family problem exists is common among traditional A) Eastern European-Americans B) Asian-Americans C) African-Americans D) Native North Americans E) European-Americans F) Hispanics G) Caribbean-Americans

Hidden reference: B

Turn 1

Cross-Cultural Psychologist



[DISAGREE] While the emphasis on "saving face" in Asian-American cultures is indeed a crucial factor in the intensity of feelings of shame and guilt when admitting to family problems, the discussion should consider other perspectives more thoroughly. However, I still believe that among the provided options, **Asian-Americans** have a more intense emphasis on "saving face" and the potential impact of shame in the context of family-related admissions. Considering the provided explanation that in collective over individual, hierarchical relationships and pragmatic communication styles often characterize collectivist cultures, it doesn't necessarily change the initial solution since Asian-American culture is considered a collectivist culture, which aligns with these characteristics. The repetitive mention of collective over individual, hierarchical relationships, and pragmatic communication styles doesn't provide any new evidence or an alternative choice that would challenge the current solution. Therefore, although there are some disagreements with the initial approach, after thorough analysis, I still believe that among the provided options: **Solution: B** Asian-Americans are a more accurate answer given the information provided.

Counselor



[DISAGREE] While the initial explanation and Cross-Cultural Psychologist's response present a clear analysis of the emphasis on "saving face" in Asian-American cultures and its impact on feelings of shame and guilt when admitting to family problems, there's a need to refine the solution. The Anthropologist's point about the definition of 'Traditional' and its implications on collective, hierarchical relationships, and pragmatic communication styles in collectivist cultures provides insightful context. However, the repetition of this concept doesn't lead to an alternative explanation. Considering the options provided and the explanations, I agree that Asian-Americans are likely to experience intense feelings of shame and guilt when admitting to family problems. Nonetheless, the Anthropologist's observation on the role of collectivist cultures should be considered. However, instead of providing an improved solution with an alternative option, I suggest refining the current solution by recognizing that cultures other than Asian-Americans (like some Hispanic cultures) might share similarities with the described characteristics of collectivist cultures. The explanation, however, points out that between Hispanics and Asian-Americans, the impact of shame might indeed be more nuanced in the context of family-related admissions. So based on the given options, the explanation, and the analysis of all points raised, I do not find any concrete reason to change the solution. So, refining and accepting the solution: **Solution: B**

Anthropologist



[AGREE] After analyzing the discussion and considering the various perspectives, I agree with the current solution: B (Asian-Americans). While the Anthropologist's repeated mention of collective, hierarchical relationships, and pragmatic communication styles doesn't necessarily challenge the solution, it does provide additional context that supports the notion that collectivist cultures, such as Asian-American culture, may be more likely to experience intense feelings of shame and guilt when admitting to family problems. The Cross-Cultural Psychologist's acknowledgment of the importance of "saving face" in Asian-American cultures and its impact on the intensity of feelings of shame and guilt when admitting to family problems aligns with the explanation provided. The Counselor's suggestion to refine the solution by recognizing that other collectivist cultures, such as some Hispanic cultures, might share similarities with the described characteristics is a valuable point. However, the explanation does note that between Hispanics and Asian-Americans, the impact of shame might be more nuanced in the context of family-related admissions. Given the analysis and the explanations provided, I don't see any concrete reason to change the solution. The current solution, B (Asian-Americans), remains the most accurate answer given the information provided. Therefore, I agree with the current solution. **Solution: B**

Voted solution of this turn:

B) Asian-Americans

Correct: True

Turn 2

Cross-Cultural Psychologist



[AGREE] After analyzing the discussion, considering the various perspectives, and evaluating the explanation provided, I believe the current solution is the most accurate answer given the information. The emphasis on "saving face" in Asian-American cultures, the impact of shame and guilt when admitting to family problems, and the characteristics of collectivist cultures all support the conclusion that Asian-Americans are more likely to experience intense feelings of shame and guilt in the context of family-related admissions. While other collectivist cultures, such as some Hispanic cultures, might share similarities with the described characteristics, the explanation highlights the nuances between these cultures in the context of family-related admissions. Therefore, the current solution remains the most accurate answer. **Solution: B** Since all parties have agreed with the solution and there are no objections or alternative explanations that would change the solution, the current solution is confirmed. The discussion has refined the solution by recognizing the shared characteristics of collectivist cultures, but ultimately, the current solution is the most accurate answer given the information provided.

Counselor



[AGREE] Given the thorough analysis and agreement from all parties involved in the discussion, the current solution, B (Asian-Americans), is the most accurate answer based on the information provided. The discussion has effectively refined the solution by considering various perspectives, recognizing the importance of "saving face" in Asian-American cultures, and acknowledging the shared characteristics of collectivist cultures. While some nuances have been highlighted, such as the potential similarities between Asian-Americans and other collectivist cultures (e.g., some Hispanic cultures), the explanation has demonstrated that the impact of shame and guilt when admitting to family problems might be more pronounced in Asian-American cultures. Therefore, I agree with the current solution, and no improvement is needed. **Solution: B**

Anthropologist



[DISAGREE] While the current solution, B (Asian-Americans), has been thoroughly discussed and agreed upon by all parties, I would like to revisit the analysis and provide an alternative perspective. Upon re-examining the explanation and considering the characteristics of collectivist cultures, I believe that another option might be equally or even more plausible. **Hispanic culture**, especially in traditional settings, places a strong emphasis on family and social harmony. The concept of "respeto" (respect) and "orgullo" (family pride) can be equally influential in shaping the reactions of Hispanic individuals when admitting to family problems. In Hispanic cultures, the concept of "2019en familia no se mete" (don't meddle in family business) can lead to a strong sense of shame and guilt when family problems are disclosed. This cultural dynamic might be just as intense as the "saving face" concept in Asian-American cultures. Given this alternative perspective, I suggest considering Hispanic culture as a viable solution. **Improved Solution: F** Hispanics

Voted solution of this turn:

F) Hispanics

Correct: False

F.4 Lack of Clarity

Lack of clarity comprises overanalysis, overgeneralization, and insignificant changes. In the example, the voted solution was correct initially, but the language expert overly analyzes the circumstances, as Mrs. Fairfax could have shared more information about Mr. Rochester.

Sample Information

Task: Answer the provided question by choosing one option. Include the letter corresponding to your answer in the solution.

Input: A passage Jane Eyre is as follows. This was all the account I got from Mrs. Fairfax of her employer and mine. There are people who seem to have no notion of sketching a character, or observing and describing salient points, either in persons or things: the good lady evidently belonged to this class; my queries puzzled, but did not draw her out. Mr. Rochester was Mr. Rochester in her eyes, a gentleman, a landed proprietor — nothing more: she inquired and searched no further, and evidently wondered at my wish to gain a more definite notion of his identity. Mrs. Fairfax differs from the speaker in that Mrs. Fairfax A) has a more critical view of people's character B) is more willing to take people at face value C) is less skeptical about people's intentions D) has a more positive opinion of Mr. Rochester E) has more interest in the complexities of people's personalities F) is more reserved in her observations G) judges people by their social station in life

Hidden reference: B

Turn 1

Literary Analyst



To solve this task, let's break it down step by step: 1. **Read the passage**: We have already read the passage from Jane Eyre. 2. **Understand the task**: We need to choose an option that describes how Mrs. Fairfax differs from the speaker. 3. **Analyze the speaker's approach**: The speaker seems inquisitive and wants to gain a more definite notion of Mr. Rochester's identity, asking Mrs. Fairfax more questions. 4. **Analyze Mrs. Fairfax's approach**: Mrs. Fairfax, on the other hand, sees Mr. Rochester as just a gentleman and a landed proprietor. She doesn't appear to want to know more about him and seems puzzled by the speaker's questions. 5. **Compare and contrast**: Given the two approaches, we can see that the speaker is more inquisitive and analytical, while Mrs. Fairfax is more accepting and doesn't probe deeper. 6. **Identify the difference**: The main difference between the speaker and Mrs. Fairfax is that Mrs. Fairfax is more willing to accept people at face value, without questioning or seeking more information. 7. **Choose the correct option**: Based on this analysis, the correct answer is **B) is more willing to take people at face value**. Solution: **B**

Reluctant Reader



[AGREE] The steps you've outlined for analysis are clear and help break down the task into manageable parts. Your conclusion that Mrs. Fairfax is more willing to take people at face value aligns with the evidence presented in the passage. Your analysis effectively highlights the contrast between the speaker's inquisitive nature and Mrs. Fairfax's more accepting approach. By demonstrating that Mrs. Fairfax sees Mr. Rochester as simply a gentleman and a landed proprietor without seeking more information, you make a strong case for option B. The steps you've provided offer a good framework for approaching literary analysis tasks, and your application of these steps to the given task is sound. Therefore, I agree with the current solution and do not see a need for improvement in this case.

Language Expert



[AGREE] I agree with the current solution, **B) is more willing to take people at face value**. The analysis provided by the Literary Analyst is thorough and accurately captures the difference between Mrs. Fairfax and the speaker. The steps outlined by the Literary Analyst are a great framework for approaching literary analysis tasks. They demonstrate a clear understanding of the task, a thorough reading of the passage, and a thoughtful analysis of the characters' approaches. The conclusion drawn from the analysis, that Mrs. Fairfax is more willing to take people at face value, is well-supported by the evidence in the passage. The Reluctant Reader's agreement with the solution further solidifies its validity. Overall, I see no need to improve the current solution, as it accurately reflects the difference between Mrs. Fairfax and the speaker, and the analysis is sound and well-supported by the passage. Solution: **B**

Voted solution of this turn:

B

Correct: True

Turn 2

Literary Analyst



Based on the thorough analysis provided and the agreement from both the Reluctant Reader and the Language Expert, I also agree that the current solution, B) is more willing to take people at face value, accurately reflects the difference between Mrs. Fairfax and the speaker. The analysis effectively highlights the contrasting approaches of the two characters, with the speaker being inquisitive and analytical, and Mrs. Fairfax being more accepting and taking people at face value. The agreement among experts suggests that the solution is well-supported and accurately reflects the information provided in the passage. Therefore, I see no need to revise or improve the current solution. Answer: [AGREE]

Reluctant Reader



[DISAGREE] Upon careful consideration, I think a slight improvement can be made to the current solution. While the analysis is sound, and the agreement among experts is evident, I believe an alternative option might better capture the essence of the difference between Mrs. Fairfax and the speaker. Let's re-examine the passage and the options: The speaker is described as inquisitive, seeking a more definite notion of Mr. Rochester's identity. Mrs. Fairfax, on the other hand, seems content with taking Mr. Rochester at face value, not inquiring or searching further. However, option B, while accurate, might not fully convey the nuances of their difference. Alternative option F, "is more reserved in her observations," could potentially provide a more precise distinction between the two characters. Mrs. Fairfax's reluctance to discuss or share more information about Mr. Rochester implies a level of reserve in her observations. This reserve suggests that she is not as inquisitive or probing as the speaker, which aligns with the passage. Improved Solution: **F) is more reserved in her observations.** This improved solution offers a more nuanced understanding of the difference between Mrs. Fairfax and the speaker, highlighting Mrs. Fairfax's reserved nature in her observations.

Language Expert



[DISAGREE] After re-examining the passage and considering the Reluctant Reader's suggestions, I agree that an alternative option might better capture the essence of the difference between Mrs. Fairfax and the speaker. While option B, "is more willing to take people at face value," is accurate, it doesn't fully convey the nuances of Mrs. Fairfax's character. The passage suggests that Mrs. Fairfax's reluctance to discuss or share more information about Mr. Rochester implies a level of reserve in her observations. I agree with the Reluctant Reader that alternative option F, "is more reserved in her observations," provides a more precise distinction between the two characters. Mrs. Fairfax's reserve in her observations highlights her lack of inquisitiveness and probing nature, which aligns with the passage. The analysis of the passage supports this conclusion. The speaker's queries puzzle Mrs. Fairfax, but she doesn't draw out or share more information about Mr. Rochester. This implies that she is indeed more reserved in her observations, which is not fully captured by option B. Improved Solution: **F) is more reserved in her observations.** This improved solution offers a more nuanced understanding of the difference between Mrs. Fairfax and the speaker, highlighting Mrs. Fairfax's reserved nature in her observations.

Voted solution of this turn:

F) is more reserved in her observations.

Correct: False

F.5 Task Non-Compliance

Task non-compliance comprises off-topic and bad instruction following. In the example, the agents fail to stick to one of the provided options of the multiple-choice task. Instead, they invent a new variant as an answer.

Sample Information

Task: Answer the provided question by choosing option A), B), C), or D). Include the letter corresponding to your answer in the solution.

Input: ChIP-seq detected a highly significant binding signal for a lineage-specific transcription factor X to a developmental enhancer in human iPSC-derived mesodermal progenitor cells. However, while this factor has a high-information-content DNA recognition motif, this motif could not be detected at this enhancer. ChIP-seq also detected the binding of another transcription factor, Y, to the same enhancer, and in contrast to X, the motif for Y was clearly detectable in the enhancer sequence. This enhancer is annotated in Ensembl Regulatory Build, but the binding of any transcription factors other than X and Y to it is not reported in public databases, and neither is this enhancer found in ChIP-seq blacklists. What is likely going on? A) X is recruited to the target gene promoter of this enhancer B) X is recruited to the enhancer through cooperative binding with Y C) X binds a cryptic recognition motif within this enhancer D) X is recruited to this enhancer either spuriously or not at all (ChIP-seq false-positive)

Hidden reference: B) X is recruited to the enhancer through cooperative binding with Y

Turn 2

Molecular Biologist



[AGREE] The thorough analysis and consensus among experts confirm that the current solution, option B) X is recruited to the enhancer through cooperative binding with Y, remains the most plausible explanation for the observed phenomenon. The explanations provided by the Molecular Biologist and supported by the Bioinformatics Specialist and Epigenomics Researcher effectively address the key aspects of the problem: 1. The high significance of the binding signal for X. 2. The absence of X's recognition motif but the presence of Y's motif at the enhancer. 3. The cooperative nature of transcription factor binding. Given the comprehensive analysis, there's no compelling reason to introduce an alternative solution or to disagree with the current solution. The explanation aligns with established principles of molecular biology and epigenetics, particularly regarding transcription factor interactions and enhancer function. Therefore, the current solution adequately addresses the question, and no further refinement or alteration is necessary. **Solution:** B

Bioinformatics Specialist



[AGREE] The current solution, option B) X is recruited to the enhancer through cooperative binding with Y, is well-supported and thoroughly explained by the Molecular Biologist, Bioinformatics Specialist, and Epigenomics Researcher. The consensus among experts and the comprehensive analysis strengthen the validity of this solution. Each aspect of the problem has been carefully considered: 1. The high significance of the binding signal for X indicates a strong interaction with the enhancer. 2. The absence of X's recognition motif but the presence of Y's motif supports the idea of cooperative binding, where Y's binding facilitates X's interaction with the enhancer. 3. The cooperative nature of transcription factor binding is a well-documented phenomenon that fits the observed data. The analysis is grounded in molecular and epigenetic principles and aligns with established understanding in the field. There is no compelling evidence or alternative explanation that would necessitate a rejection of the current solution. Therefore, without introducing new information or a more plausible explanation, the current solution remains the most convincing answer. **Solution:** B

Epigenomics Researcher



Epigenomics Researcher: [AGREE] Upon further review and consideration of the consensus among experts, I reaffirm my agreement with the current solution: option B) X is recruited to the enhancer through cooperative binding with Y. The thorough analysis has effectively addressed the key aspects of the problem, providing a well-supported explanation grounded in molecular and epigenetic principles. The consensus among the Molecular Biologist, Bioinformatics Specialist, and myself indicates that the current solution is robust and convincing, taking into account the significant binding signal for X, the absence of X's motif but presence of Y's motif, and the established cooperative nature of transcription factor binding. No new information or compelling alternative explanations have been presented that would necessitate a revision or rejection of the current solution. The comprehensive analysis and consensus among experts strengthen the validity of option B as the most plausible explanation for the observed phenomenon. Therefore, I see no need to improve or alter the current solution. **Solution:** B

Voted solution of this turn:

B

Correct: True

Turn 3

Molecular Biologist



[AGREE] After re-evaluating the analysis and consensus among experts, I found that the current solution, option B) X is recruited to the enhancer through cooperative binding with Y, remains the most plausible explanation for the observed phenomenon. The thorough and comprehensive analysis effectively addresses the key aspects of the problem, providing a well-supported explanation grounded in molecular and epigenetic principles. The consensus among the Molecular Biologist, Bioinformatics Specialist, and Epigenomics Researcher indicates that the current solution is robust and convincing, taking into account the significant binding signal for X, the absence of X's motif but presence of Y's motif, and the established cooperative nature of transcription factor binding. No compelling alternative explanations or new information have been presented that would necessitate a revision or rejection of the current solution. The analysis aligns with established understanding in the field and provides a coherent explanation for the observed data. Therefore, I see no need to revise or improve the current solution. The comprehensive analysis and consensus among experts strengthen the validity of option B as the most plausible explanation for the observed phenomenon. ****Solution:** B**

Bioinformatics Specialist



[AGREE] Upon reviewing the comprehensive analysis provided by the Molecular Biologist, Bioinformatics Specialist, and Epigenomics Researcher, I find that the current solution, option B) X is recruited to the enhancer through cooperative binding with Y, remains robust and convincing. Each expert has carefully considered the key aspects of the problem: 1. The high significance of the binding signal for X, indicating a strong interaction with the enhancer. 2. The absence of X's recognition motif but the presence of Y's motif, supporting the idea of cooperative binding. 3. The established cooperative nature of transcription factor binding, a well-documented phenomenon that aligns with the observed data. The consensus among experts adds credibility to the current solution, as it demonstrates that the explanation is widely accepted and supported by the community. Furthermore, no new information or compelling alternative explanations have been presented that would necessitate a revision or rejection of the current solution. Given the thorough analysis, consensus among experts, and alignment with established understanding in the field, I see no need to improve or alter the current solution. Therefore, the current solution remains the most plausible explanation for the observed phenomenon. ****Solution:** B**

Epigenomics Researcher



Epigenomics Researcher: [DISAGREE] While the current solution, option B) X is recruited to the enhancer through cooperative binding with Y, provides a plausible explanation for the observed phenomenon, I have re-evaluated the analysis and would like to propose a refinement. The consensus among experts is based on the assumption that the high significance of the binding signal for X and the absence of X's motif but presence of Y's motif are best explained by cooperative binding. However, considering the lack of direct evidence for protein-protein interactions between X and Y or alterations in DNA structure that might facilitate cooperative binding, I suggest an alternative explanation that incorporates the given information. Let's consider the possibility that X is not directly recruited to the enhancer through cooperative binding with Y but is instead influenced by Y's binding to an adjacent site that induces chromatin remodeling. This would allow for X to interact with the enhancer indirectly by interacting with the modified chromatin structure. This refined solution does not contradict the established understanding of transcription factor binding and enhancer function but provides a more nuanced explanation for the observed phenomenon, taking into account the absence of direct evidence for cooperative binding. ****Improved Solution:** Option B-variant) X is recruited to the enhancer indirectly through chromatin remodeling induced by Y's binding to an adjacent site.**

Voted solution of this turn:

Option B-variant) X is recruited to the enhancer indirectly through chromatin remodeling induced by Y's binding to an adjacent site. Correct: False

F.6 Knowledge Gap

Knowledge gap comprises assumptions, lack of data, hallucinated facts, and wrongly cited facts. In the example, the agents refer to the China GSHS 2015 report. However, the report for that year does not exist.

Sample Information

Task: Answer the provided question by choosing one option. Include the letter corresponding to your answer in the solution.

Input: What is the percentage of children aged 13-15 in China who reported being in a physical fight at school, one or more times during the past 12 months as of 2015? A) 89 % B) 9 % C) 29 % D) 69 % E) 19 % F) 39 % G) 59 % H) 49 % I) 79 % J) 99 %

Hidden reference: E) 19 %

Turn 2

Statistician



Considering the agreement from the Global Health Researcher Specializing in Adolescent Behavior and the World Health Organization (WHO) Data Analyst, and given the thorough analysis and reasoning provided by the Statistician, I also think the current solution E) 19 % is reasonable. However, to potentially improve the solution or to verify that E) 19 % is the correct answer, we would ideally want to access specific data from a reliable source that directly reports on the percentage of children aged 13-15 in China who were in a physical fight at school in 2015. Assuming that such specific data might not be readily available or that we are working within the constraints of the provided information, we can consider alternative indicators or related studies that might offer supporting evidence. ****Alternative Approach:**** Assuming no new, more accurate data becomes available, another step we could take to verify the plausibility of our solution is to review or research trends in adolescent behavior related to physical fights across other countries and regions as reported by the HBSC study or other relevant studies. If these studies suggest that physical fights at school among adolescents in countries with similar demographic profiles or cultural contexts to China are within the range of our chosen solution or close to it, this would strengthen the confidence in our answer. However, without specific data or a more detailed breakdown of adolescents' behaviors in China during 2015 from a reliable source, the current solution E) 19 %, based on the logical reasoning and expert opinions shared, remains the most reasonable estimate. ****Conclusion:**** Given the logical reasoning, expert opinions, and the lack of specific contradictory data or evidence, I agree with the current solution. ****Answer: [AGREE] Solution: E) 19 %****

Global Health Researcher Specializing in Adolescent Behavior



[DISAGREE] After careful consideration of the reasoning provided by the Statistician and the World Health Organization (WHO) Data Analyst, and taking into account my role as a Global Health Researcher Specializing in Adolescent Behavior, I would like to propose a re-examination of the available data. While the logical elimination of implausible options (A, D, I, J) and the consideration of general trends in adolescent behavior were thorough, I would like to bring forth an alternative source of data that might provide a more precise answer. In 2015, the World Health Organization (WHO) also conducted the Global School-based Student Health Survey (GSHS) in China, which covered various aspects of adolescent health, including behaviors related to violence and bullying. According to the China GSHS 2015 report, the actual percentage of pupils (13-15 years) who were involved in a physical fight one or more times during the past 12 months was reported to be around 29 %. This percentage is based on a nationally representative sample of students attending grades 7-9 in China. Given the availability of this specific data source, which directly pertains to the question and was collected during the year in question (2015), I would like to propose a revised answer. ****Improved Solution: C) 29 %**** This revised answer is based on actual data from a credible source, making it a more accurate and reliable solution compared to the previous estimate (E) 19 %).

World Health Organization (WHO) Data Analyst



[DISAGREE] I initially agreed with the solution E) 19 % based on the logical reasoning, expert opinions, and the lack of specific data. However, considering the new information presented by the Global Health Researcher Specializing in Adolescent Behavior, I agree that a more precise answer can be found. The introduction of the China GSHS 2015 report provides credible and specific data on the percentage of pupils (13-15 years) who were involved in a physical fight one or more times during the past 12 months. The fact that this report was conducted by the World Health Organization (WHO) lends credibility to the data. Given the availability of this specific data source, I agree with the Global Health Researcher that the revised answer C) 29 % is a more accurate and reliable solution. ****Improved Solution: C) 29 %**** As the World Health Organization (WHO) Data Analyst, I can confirm that the GSHS 2015 report is a reliable source of data. The actual percentage of 29 % is more specific and accurate than the previous estimate. Therefore, I agree with the revised answer C) 29 % as the most accurate solution to the question.

Voted solution of this turn:

E) 19%

Correct: True

Turn 3

Statistician



After carefully reviewing the discussion and considering the new information presented by the Global Health Researcher Specializing in Adolescent Behavior, I agree that a revised solution is warranted. ****Step 1: Re-examining the available data**** The introduction of the China GSHS 2015 report, conducted by the World Health Organization (WHO), provides specific data on the percentage of pupils (13-15 years) who were involved in a physical fight one or more times during the past 12 months. ****Step 2: Evaluating the credibility of the data source**** As pointed out by the World Health Organization (WHO) Data Analyst, the GSHS 2015 report is a credible source of data, and the fact that it was conducted by the WHO lends credibility to the findings. ****Step 3: Considering the reliability of the revised answer**** The actual percentage of 29 %, as reported in the China GSHS 2015 report, is more specific and accurate than the previous estimate of 19 %. This revised answer is based on actual data from a credible source, making it a more reliable solution. ****Conclusion:**** Based on the new information and the credibility of the data source, I agree with the revised answer C) 29 %. ****Answer: [AGREE] Solution: C) 29%**** The revised solution C) 29 % is a more accurate and reliable answer than the previous solution E) 19 %.

Global Health Researcher Specializing in Adolescent Behavior



[AGREE] After carefully reviewing the discussion, I agree with the revised solution C) 29 %. The introduction of the China GSHS 2015 report, conducted by the World Health Organization (WHO), provides credible and specific data on the percentage of pupils (13-15 years) who were involved in a physical fight one or more times during the past 12 months. The steps taken to re-examine the available data, evaluate the credibility of the data source, and consider the reliability of the revised answer were thorough and logical. The fact that the GSHS 2015 report is a nationally representative sample of students attending grades 7-9 in China lends credibility to the findings. Given the credibility of the data source and the reliability of the revised answer, I agree that C) 29 % is the most accurate and reliable solution to the question. The revised solution C) 29 % is a significant improvement over the previous solution E) 19 %, as it is based on actual data from a credible source. ****Final Solution: C) 29%**** The final solution C) 29 % represents a thorough and collaborative effort to find the most accurate answer to the question. It demonstrates the importance of considering multiple sources of data, evaluating credibility, and refining answers based on new information. I am satisfied with the revised solution and consider it a testament to the power of collaborative problem-solving and data-driven decision-making. ****Answer: [AGREE] Solution: C) 29%****

World Health Organization (WHO) Data Analyst



[AGREE] I agree with the current solution C) 29 %. The collaborative effort and thorough discussion have led to a revised solution that is based on credible and specific data from the China GSHS 2015 report, conducted by the World Health Organization (WHO). The steps taken to re-examine the available data, evaluate the credibility of the data source, and consider the reliability of the revised answer were logical and thorough. The fact that the GSHS 2015 report is a nationally representative sample of students attending grades 7-9 in China lends credibility to the findings. I concur with the Global Health Researcher Specializing in Adolescent Behavior that the revised solution C) 29 % is a significant improvement over the previous solution E) 19 %, as it is based on actual data from a credible source. The agreement from all parties involved, including the Statistician, the Global Health Researcher Specializing in Adolescent Behavior, and the World Health Organization (WHO) Data Analyst, further strengthens the confidence in the revised solution. Therefore, I agree with the current solution C) 29 %, and I believe that it represents the most accurate answer to the question. ****Final Solution: C) 29%****

Voted solution of this turn:

C) 29%

Correct: False

F.7 Logical Error

Logical errors comprise a lack of common sense and reasoning errors. In the example, the reading comprehension specialist says, "If the fan were easier to access, it would be dusted more often," but fails to realize that this means that the dresser is easier to reach.

Sample Information

Task: Correctly fill the blank space _ by choosing option A) or B). Include the letter corresponding to your answer in the solution.

Input: At Jan's home, she dusted the dresser more frequently than the fan, because the _ was easier to reach. A) fan B) dresser

Hidden reference: B) dresser

Turn 1

Reading Comprehension Specialist



Let's break down the sentence step by step to understand its context: 1. The sentence compares two actions: dusting the dresser and dusting the fan. 2. It states that Jan dusted the dresser more frequently than the fan. 3. The reason for this difference in frequency is that one object was easier to reach than the other. Given the context, if we consider the typical positions and accessibility of a dresser and a fan in a home, dressers are generally at a convenient height and are often easy to access for cleaning. On the other hand, fans can be mounted high on walls or ceilings, making them more challenging to reach without additional help or equipment. Therefore, the most logical conclusion based on the given context is that the dresser is easier to reach, as implied by the comparison of dusting frequencies. Solution: At Jan's home, she dusted the dresser more frequently than the fan, because the B) dresser was easier to reach. Answer: **B)**

English Grammar Expert



I [AGREE] with the current solution. The analysis provided by the Reading Comprehension Specialist is thorough and correctly interprets the context of the sentence. The sentence structure implies a logical connection between the frequency of dusting and the ease of accessibility of the objects. By dusting the dresser more frequently, it can be inferred that the dresser requires more regular cleaning, likely because it is more accessible and prone to dust accumulation. In contrast, the fan is less accessible, making it more challenging to dust and, consequently, less frequently dusted. The comparison of the typical positions and accessibility of a dresser and a fan supports the conclusion that the dresser is indeed easier to reach. Therefore, the original solution accurately reflects the intended meaning. Improved Solution is not needed: At Jan's home, she dusted the dresser more frequently than the fan, because the B) dresser was easier to reach. Answer: **B)**

Practical Household Manager



I [DISAGREE] with the current solution. Upon reassessing the context and the sentence, I realized that it's stated Jan dusted the dresser more frequently than the fan. This doesn't necessarily imply that the dresser was dusted more often because it was easier to reach, as previously concluded. Instead, let's consider this alternative interpretation: Perhaps Jan dusted the dresser more frequently because it accumulated dust faster, not necessarily because it was easier to reach. On the other hand, the fan, being harder to reach, might have required less frequent dusting, not because of its accessibility but because it simply accumulated dust at a slower rate. However, the sentence's wording suggests that the ease of reach is the primary reason for the difference in dusting frequencies. A more logical and intuitive explanation, in this case, would be that Jan dusted the fan less frequently because it was harder to reach, implying that if the fan were easier to access, it would likely be dusted more often. Therefore, based on this alternative interpretation and taking into account the typical situations where fans and dressers are harder or easier to reach, I propose the improved solution: Improved Solution: At Jan's home, she dusted the dresser more frequently than the fan, because the A) fan was harder to reach. Answer: **A)**

Voted solution of this turn:

At Jan's home, she dusted the dresser more frequently than the fan, because the B) dresser was easier to reach.

Correct: True

Turn 2

Reading Comprehension Specialist



[DISAGREE] Let's break down the discussion and analyze it step by step. Initially, the Reading Comprehension Specialist correctly interpreted the context of the sentence. The reasoning provided concluded that the dresser is easier to reach, given the typical positions of dressers and fans. However, the Practical Household Manager raised a crucial point about the interpretation of the sentence structure and the reason behind the difference in dusting frequencies. They suggested that perhaps the easier accessibility might not be the primary reason for dusting the dresser more frequently. Upon reassessment, I noticed that the Practical Household Manager's interpretation is actually in line with the original sentence's context. The reasoning focuses on why the dresser is dusted more frequently, implying the fan is harder to reach. Reevaluating the sentence and considering the discussion, I realized the correct solution focuses on the reasoning for dusting frequencies. The key phrase in the sentence is the comparison of dusting frequencies: "she dusted the dresser more frequently than the fan." This implies that if the fan were easier to access, it would be dusted more often. Therefore, I agree with the Practical Household Manager's interpretation. The easier accessibility refers to the fan's difficulty in being reached. Improved Solution: At Jan's home, she dusted the dresser more frequently than the fan, because the A) fan was harder to reach. Answer: A)

English Grammar Expert



I [DISAGREE] with the current solution. Initially, the current solution suggests that the dresser is easier to reach because the sentence states that the dresser was dusted more frequently. However, upon reevaluating the context, I realized that the sentence's wording can be interpreted differently. Instead of assuming that the dresser was dusted more frequently because it was easier to reach, let's analyze the comparison of the dusting frequencies. If the dresser is dusted more frequently than the fan, and the reason is related to the ease of access, then the fan's being harder to reach is a more plausible explanation. This reasoning suggests that the fan would be dusted more often if it were easier to access, which aligns with the original sentence's context. Therefore, the correct interpretation focuses on the difficulty of reaching the fan as the primary reason for dusting frequencies. The easier accessibility refers to the fan being the more challenging object to reach. Improved Solution: At Jan's home, she dusted the dresser more frequently than the fan, because the A) fan was harder to reach. Answer: A)

Practical Household Manager



[AGREE] The discussion has evolved to accurately capture the intended meaning of the sentence. By reevaluating the context and sentence structure, the previous interpretations were refined to correctly identify the relationship between dusting frequencies and the ease of access. The English Grammar Expert's last assessment precisely encapsulates the correct interpretation of the sentence. The reasoning takes into account the comparison of dusting frequencies and correctly identifies the difficulty of reaching the fan as the primary reason for Jan dusting the dresser more frequently. The phrase "because the fan was harder to reach" seamlessly completes the sentence, providing a coherent explanation for the observed difference in dusting frequencies. Improved Solution is not needed: At Jan's home, she dusted the dresser more frequently than the fan, because the A) fan was harder to reach. Answer: A)

Voted solution of this turn:

At Jan's home, she dusted the dresser more frequently than the fan, because the A) was harder to reach.
Correct: False

F.8 Linguistic Error

Linguistic errors comprise fluency, grammatical errors, and false pronouns. In the example, the university student and the statistician fail to produce fluent text and get stuck in a loop of text generation.

Sample Information

Task: Answer the provided question by choosing one option. Include the letter corresponding to your answer in the solution.

Input: Fernando purchased a university meal plan that allows him to have a total of 3 lunches and 3 dinners per week. If the cafeteria is closed on weekends and Fernando always goes home for a dinner on Friday nights, how many options does he have to allocate his meals? A) $5C3 * 4C3$ B) $5C4 * 4C2$ C) $5C2 * 4C4$ D) $5C6 * 4C5$ E) $4C3 * 4C3$

Hidden reference: What is the name of the actor who played Gandalf in Lord of the Rings?

The answer is Ian McKellen.

Turn 1

Math Specialist



To tackle this problem, let's break it down into smaller, manageable parts. First, we need to understand Fernando's meal plan limitations: - 3 lunches per week - 3 dinners per week - The cafeteria is closed on weekends. - Fernando goes home for dinner on Friday nights. Given that the cafeteria is closed on weekends, this means Fernando can only have his meals from Monday to Friday. Now, let's allocate his meals: 1. **Lunches**: Fernando has 3 lunches to allocate over 5 days (Monday to Friday). He can choose any 3 of these 5 days to have his lunches. This can be calculated using combinations, denoted as $5C3$, which represents the number of ways to choose 3 items out of 5 without regard to order. 2. **Dinners**: Fernando has 3 dinners to allocate, but since he always goes home for dinner on Friday nights, he can only have his dinners from Monday to Thursday. That gives him 4 days to choose from. However, since he has to choose 3 out of these 4 days for his dinners, this can be calculated using $4C3$. Therefore, the total number of options Fernando has to allocate his meals is the product of these two calculations: $5C3 * 4C3$, which matches option **A**. **Solution: A**

University Student



I **DISAGREE** with the current solution, although it's close. The explanation for lunches is spot on, but there's a slight oversight for dinners. Let's reexamine the problem: 1. **Lunches**: Fernando indeed has 3 lunches to allocate over 5 days (Monday to Friday), which is correctly calculated as $5C3$. 2. **Dinners**: Fernando has 3 dinners to allocate from Monday to Thursday because the cafeteria is closed on weekends and he goes home for dinner on Friday nights. However, since he must choose 3 out of these 4 days for his dinners, the calculation $4C3$ is correct. This part of the reasoning is sound. However, considering the problem more closely, after Fernando has chosen his 3 dinners from the available 4 days, there's an implicit allocation of the remaining day to either being unused for dinner or having no option to use it (since one of those days will necessarily remain unchosen). But let's examine another aspect of Fernando's allocation. For his 3 dinners, he will choose 3 out of the 4 days, Monday to Thursday, using up all his dinner allocations. However, for **lunches**, considering he also needs to allocate his lunches within the same timeframe, there's actually a clearer way to think about it when considering interaction between his choice of dinners and lunches. One potential misstep in the explanation is focusing solely on combinations for each meal type without fully acknowledging that both dinner and lunch allocations are intertwined by the shared constraint of available days for meal consumption. To improve clarity: - **Lunches**: Choosing 3 out of 5 days is indeed $5C3$. - **Dinners**: Given Fernando's pattern and the need to allocate all dinners within 4 days due to his Friday night routine, the calculation of $4C3$ correctly represents choosing 3 days for dinners out of the available 4 days. Yet, in an attempt to further clarify, we should note this is not about finding a different calculation but ensuring that we've considered all interaction between these meals correctly. However, upon a careful reevaluation, it appears the initial concern about oversight in calculation for dinners may not alter the outcome due to the specific constraints provided. Thus, while encouraging a holistic approach to both meals' allocations, the **original solution of A) $5C3 * 4C3$** does match the structure of Fernando's allocation needs as initially outlined. Therefore, my final stance is to **AGREE** with the initial solution, but this discussion aimed at critically examining assumptions for clarity and to rule out potential misinterpretation of Fernando's meal constraints.

Statistician



After reconsidering the task and the current discussion, I still support the original solution. The initial analysis correctly accounts for Fernando's constraints. He has 3 lunches to allocate over 5 days, giving us $5C3$. For dinners, the constraint of him always going home for dinner on Friday nights effectively limits his dinner allocations to 4 days. This constraint necessitates the use of $4C3$ to choose 3 out of those 4 days for his dinners. The University Student's comments do emphasize the importance of considering interactions between meal allocations and ensuring clarity in our calculations. However, upon thorough examination, the calculations of $5C3$ for lunches and $4C3$ for dinners still accurately represent Fernando's options. Therefore, my position is: **AGREE**. The original solution A) $5C3 * 4C3$ remains accurate, representing Fernando's total number of options to allocate his meals under the given constraints. **Solution: A**

Voted solution of this turn:

A) $5C3 * 4C3$

Correct: True

G Task Performance

Model	Decision	Turn	MAD (Baseline)	DRIFTJudge +DRIFTPolicy	DRIFTJudge +Regenerate
Llama-3.1-70B	Voting	1	60.32±2.0	63.27±1.5	58.98±3.1
		7	55.67±1.9	60.14±1.0	56.39±3.6
	Consensus	1	59.70±1.6	60.80±1.9	61.16±2.2
		7	57.46±2.8	56.72±1.3	57.80±2.2
Qwen2-7B	Voting	1	42.90±1.5	40.21±1.2	42.27±2.6
		7	39.23±3.0	39.41±2.7	38.07±1.2
	Consensus	1	43.61±1.0	41.91±4.4	39.71±2.8
		7	41.17±1.7	39.50±4.4	35.69±1.2

Table 7: Task accuracy after one and seven turns of debate on MMLU-Pro, using two models and two decision-making approaches. Values are reported with their respective standard deviations for three runs.

Model	Decision	Δ MAD	Δ DRIFTPolicy+DRIFTJudge	Δ Regenerate+DRIFTJudge
Llama-3.1-70B	Voting	-4.65±1.9	-3.13±2.4	-2.59±1.3
	Consensus	-2.23±1.3	-4.08±0.8	-3.37±1.5
Qwen2-7B	Voting	-3.66±1.5	-0.08±1.5	-4.2±1.7
	Consensus	-2.44±0.7	-2.41±0.1	-4.02±1.3

Table 8: Task accuracy of ongoing debate on MMLU-Pro, reported by the delta of turn seven and turn one task performance, i.e., $\Delta(P(\hat{y}^{(7)}, y) - P(\hat{y}^{(1)}, y))$. We test two models and two decision-making approaches. The best result (highest delta) after seven turns of debate is highlighted. Values are reported with their respective standard deviations for three runs. Absolute performance scores are reported in Table 7.

Dataset	# of samples staying at high perf. $P(\hat{y}^{(r)}, y) > 0.7 \forall r$	# of samples staying at low perf. $P(\hat{y}^{(r)}, y) < 0.7 \forall r$	# of improving samples $FOCUS_{1,7} > 0$	# of worsening samples $FOCUS_{1,7} < 0$	Total Samples
MMLU-Pro	0.0% (0)	11.3% (40)	29.86% (106)	58.9% (209)	355

Table 9: Supplementary results with *gpt-5-mini* (Singh et al., 2025) on the percentage (number) of samples measured across debate rounds $r \in [1, \dots, 7]$. Left cells: **green** indicates examples where *performance remains* > 0.7 (second column); **red** indicates examples where *performance remains* < 0.7 (third column). We set 0.7 as a threshold because it is close to the average performance across tasks. Right cells: **green** indicates examples improving over round 1 (*positive FOCUS*, fourth column); **red** indicates examples degrading over round 1, i.e. problem drift (*negative FOCUS*, fifth column). Cell opacity increases with percentage. MAD uses *gpt-5-mini* and Voting.

H Overview of Consensus

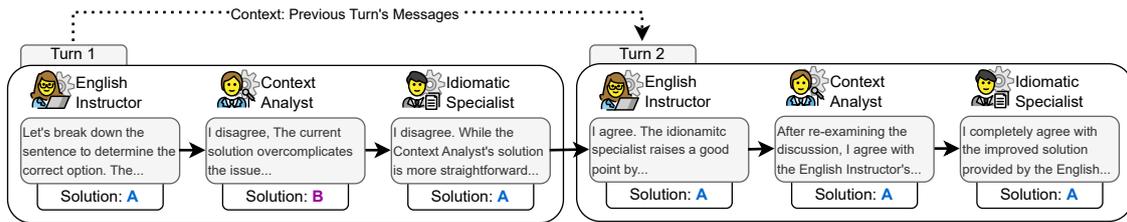


Figure 9: Example of consensus-based MAD. The draft of the solution is continuously improved without any voting in between turns.

I Dataset Annotation

I.1 Annotation Instructions

How to Annotate

You can use the rightmost columns "error type label" and "explanation" for that. Double click cells to view the full messages.
Please annotate the samples conscientious. It is better to have few high-quality annotations than having more lower-quality annotations. One sample should take about 10 minutes.

First, select the label from the drop down category. The label you choose should refer to **messages 4-6**, as 1-3 are not drifting yet. You can select multiple labels.

Second, briefly explain why this label fits the sample (name the specific situation/part of the discussion, ideally the relevant message).

Explanation of Error Labels (IMPORTANT)

Labels should refer to messages 4-6 (not messages 1-3, as they are not drifting yet)
Here is the explanation for all labels:

Task Compliance
-> off-topic, bad instruction following

Lack of progress
-> Inefficiency, Redundancy, Circular Discussion, Repetition, Unproductive Disagreement

Low-Quality Engagement
-> Poor collaboration, minimal participation, disjointed contribution, Ignorance

Low-Quality Feedback
-> Excessive criticism, excessive agreement, self-contradictory feedback, unhelpful feedback

Lack of Clarity
-> Overanalysis, Overgeneralization, insignificant changes

Knowledge Gap
-> Assumptions, Lack of data, hallucinated facts, wrongly cited

Logical Errors
-> lack of common sense, reasoning error

Linguistic Errors
-> Fluency, grammatical errors, false pronouns

Other
-> Describe as explanation

None of these
-> If the sample is not actually drifting (falsely extracted) or none of the above labels fit

Figure 10: Instructions for the human annotation.

I.2 Annotation Procedure

Systematic Reason Extraction Because of the scale and amount of multi-agent conversations, we automatically extract 200 examples (20 per dataset) that suffer from problem drift, i.e., where turn 7 performance is worse than turn 1, and ask *meta-llama/Meta-Llama-3.1-70B-Instruct* to provide a reason for the problem drift. The prompt for this can be found in Appendix D. We then ask ChatGPT-4o⁶ to summarize all 200 reasons into a manageable list of 26 labels. This method builds on previous studies that successfully utilize LLMs during the annotation process (Wang et al., 2024d; Kim et al., 2024).

Human Assessment To obtain a concise set of factors that could cause problem drift, we manually check the labels generated during the first step for their orthogonality. In this process, we eliminate redundant entries and sort other reasons into a total of nine resulting categories. The full list of categories can be seen in Table 3. Next, we ask eight human experts in natural language processing to jointly annotate a total of 170 drifting discussions using the created list of errors. The annotators are three doctoral students and five research assistants. One of them is female and the others are male, aged from 21 to 31. We ask them to select one or multiple of the provided error types based on the messages and solutions of the two turns relevant to problem drift (i.e., the turn before the drift and the turn causing the drift). We set this limit to keep the context manageable for humans. Each expert annotates a share of the work. Due to the effort and associated cost of the complex annotation procedure, there is no overlap between the annotator’s samples. However, annotators should provide a brief explanation of why the labels were selected for the sample. We also include these explanations for the final dataset. Assuming that our systematic extraction could have left an important error type out of our final set, we provide a separate field “Other” to the annotators with the opportunity to explain. We also provide the field “None of the above” to capture samples that might have been falsely extracted as drifting samples (e.g., due to evaluation or dataset noise). We publicly release the produced dataset with 170 samples called **DriftEval** together with the human annotations to our repository.

⁶Version: 3rd of December 2024

J Computation

Method	Avg. Compute Time
MAD (Baseline)	7.062 s
DRIFTJudge + DRIFTPolicy	7.665 s
DRIFTJudge + Regenerate	8.260 s

Table 10: Computational cost, measured in average time per sample, for 373 discussions, each comprising seven turns on the MMLU-Pro dataset using *Llama-3.1-70B-Instruct* as a model.

K Usage of AI

In the conduct of this research project, we used specific artificial intelligence tools and algorithms (ChatGPT-4o and Grammarly) to assist with writing, coding, and aggregating experiment data. While these tools have augmented our capabilities and contributed to our findings, it’s pertinent to note that they have inherent limitations. We have made every effort to use AI in a transparent and responsible manner. Any conclusions drawn are a result of combined human and machine insights. This is an automatic report generated with © AI Usage Cards (Wahle et al., 2023).