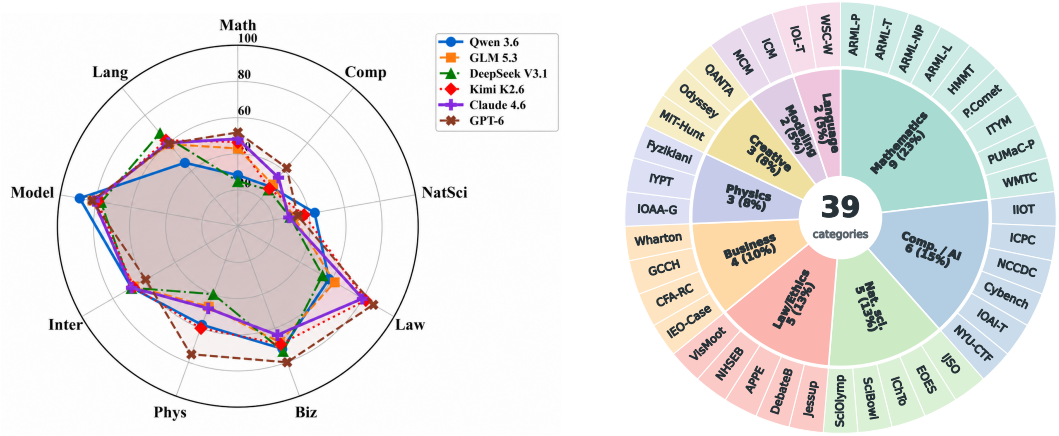
CAN AGENTS COMPETE AS A TEAM? BENCHMARKING LONG-HORIZON MULTI-AGENT COLLABORATION WITH REALISTIC CONTESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based multi-agent systems (MASs) have demonstrated extraordinary performance in numerous tasks and domains, but their collaborative capability in challenging real-world tasks is still largely underexplored. To bridge this gap, we propose CONTESTWORLD, a benchmark dedicated to evaluating agent collaboration with diverse team-based human contests. It consists of 321 contests from different years with 39 categories, such as MCM, ICPC, and ARML, spanning 9 domains, ranging from mathematics to computer science, business, and linguistic contests. CONTESTWORLD is deployed on a carefully designed realistic contest simulator that closely aligns with human settings, with an action interface comprising 27 common actions, three optional memory actions, and contest-specific tools, filtered according to contest rules and protocol settings. For instance, in an ICPC contest, three teammates share a single compiler, so only one of them can use it at a time. In CONTESTWORLD, agents must allocate tool calls within the limited turns, interacting with the teammates and environments to achieve the best performance. With CONTESTWORLD, we have conducted a comprehensive evaluation of four primary baselines and six backbone LLMs. The best-performing GPT-6 model achieves only 63.4% on average, demonstrating the challenge of the proposed dataset. Our analysis shows that teams primarily benefit from improved task coverage and selection, while their remaining failures stem from weak verification and poor feedback integration.



(a) (b)

Figure 1: (a) Performance of all LLMs. (b) Statistics of all categories and contest names.

1 INTRODUCTION

As large language model (LLM)-based agents advance from single-turn reasoners to autonomous executors of complex tasks (Luo et al., 2025), collaboration between them becomes a critical

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

capability (Han et al., 2026; Tran et al., 2025). Many real-world tasks, from software development (Li et al., 2023; Qian et al., 2024) to scientific research (Qu et al., 2026; Novikov et al., 2025) and business operations (Yang et al., 2024; Fan et al., 2026), require multiple specialized agents to coordinate toward shared goals. Systematically evaluating this collaboration capability is essential for understanding the strengths and limitations of current models and guiding future progress.

Recent benchmarks have been curated to assess the collaboration capability of MASs. Some datasets such as Collab-Overcooked (Sun et al., 2025) and AgentWorld (Shu et al., 2026) examine communication and adaptation in cooperative games, while MultiAgentBench (Zhu et al., 2025) evaluates collaboration with simple tasks, such as database management. However, these datasets tend to focus either on simulated environments and RPG-style games (Hu et al., 2026), such as 2D MMORPG settings, or on narrow pseudo-tasks, such as resolving an issue in a particular environment, leaving a realistic and challenging collaboration evaluation of multi-agent systems largely underexplored.

On the other hand, competition material has become a challenging testbed for evaluating modern LLMs. Competition results are routinely featured in technical reports of mainstream LLMs (OpenAI, 2023; Yang et al., 2025), and benchmarks such as LiveOIBench assess solutions to informatics-olympiad problems (Zou et al., 2026). However, most competitions in existing benchmarks still focus on individual problem solutions rather than teamwork, overlooking the many team-based contests that could evaluate diverse collaborative capabilities, such as planning and task assignment (Table 7). The question at the intersection of the two remains open: *how well can agents perform in a truly challenging team competition?*

In this paper, we introduce ContestWorld, a realistic benchmark evaluating contest-level collaboration. We develop a taxonomy to guide the construction of ContestWorld. First, we identify nine domains of human team contests, including mathematics, computing and AI, natural sciences, law and ethics, business and finance, physics and astronomy, interdisciplinary and creative tasks, modeling, and language. Then, we curate diverse types of questions, including multiple-choice questions, short-answer questions, coding and execution-based tasks, and project-based tasks. Finally, we cover teams of 2 to 15 participants, from middle-school to college-level students. Furthermore, we challenge the MASs through long-horizon tasks, with up to 90 turns per competition. Overall, ContestWorld incorporates 39 categories, with 321 contests and 2,865 problems drawn from 2009 to 2026 (Figure 1b). To evaluate MASs in a realistic setting, we construct a turn-based simulator with an action interface comprising 27 common actions, three optional memory actions, and contest-specific tools. Available actions and tools are filtered by contest rules, baseline, role, and contest state; examples include communicating with teammates, submitting answers, and using a calculator when permitted.

We conduct extensive experiments on ContestWorld with six LLMs known for their strong capabilities, including four open-source models, Kimi-K2.6, GLM-5.3, Qwen3.6, and DeepSeek-V3.1, along with two closed-source models, GPT-6 and Claude 4.6. We compare four baselines, including single-agent, centralized and decentralized communication, and a proposed method, Open Table Coach (OTC), which enhances decentralized agents with memory and the coach agent. OTC achieves the highest average score on three of the four open-source models and outperforms the corresponding single-agent baseline on all four. Results show that the best-performing GPT model only achieves 63.4% with OTC, demonstrating the challenge of the proposed benchmark (Figure 1a). Our analysis shows that teams primarily benefit from improved task coverage and selection, while their remaining failures stem from weak verification and poor feedback integration (§ 4.3,§ 4.4).

Our contributions are threefold: (1) we build ContestWorld, the first contest-based multi-agent benchmark for evaluating collaboration capability in realistic and challenging settings; (2) we evaluate state-of-the-art models on this benchmark, revealing a large performance gap; (3) we identify several failure modes of LLMs when collaborating with other agents, shedding light on the future development of high-quality collaborative agent systems.

2 Related Work

Multi-agent evaluation. Multi-agent benchmarks assess collaboration through communication, planning, and task progress, extending evaluation beyond final task scores (Zhu et al., 2025; Gong et al., 2024; Sun et al., 2025; Shu et al., 2026). Other work varies information access, communication

protocols, and dialogue length to examine when interaction is useful (Zhang et al., 2026; Eisenstein et al., 2026). Evaluations of shared work test whether agents can integrate separate contributions and how enforced role separation affects coordination (Khatua et al., 2026; Kim et al., 2026). These studies motivate evaluating both the outcomes and the process of collaboration. CONTESTWORLD extends this perspective to teams operating under explicit competition rules, reporting task performance alongside coordination quality.

Competition-derived reasoning benchmarks. Competition-derived benchmarks test demanding reasoning across mathematics, science, and linguistics (He et al., 2024; Sánchez et al., 2025). In programming, LiveCodeBench and LiveCodeBench Pro combine continuously updated contest problems with executable evaluation, with the latter adding expert annotations and failure analysis (Jain et al., 2025; Zheng et al., 2025). LiveOIBench further evaluates solutions using official tests and aggregates scores into contest standings relative to human participants (Zou et al., 2026). For proof-oriented tasks, IneqMath evaluates both final answers and the soundness of intermediate steps using specialized LLM judges (Sheng et al., 2025). These benchmarks assess individual agents' ability to solve competition-level problems. In contrast, our focus is how a team allocates work and integrates contributions within an entire contest under a fixed turn-based budget.

Agent orchestration. LLM-based multi-agent systems organize collaboration through conversations and role-based workflows (Wu et al., 2024; Hong et al., 2024). Research further explores iterative peer critique to refine answers (Du et al., 2024), training objectives that balance task performance and communication cost (Chen et al., 2025), and hierarchical memory that retrieves prior collaboration trajectories and distilled experience (Zhang et al., 2025b). In contrast, we mainly use diverse orchestration of centralized, decentralized, and coach-guided team protocols under contest rules and resource constraints to benchmark collaboration capability.

3 THE CONTESTWORLD BENCHMARK

This section outlines our benchmark construction process, including a brief overview of each category and contest, as well as the steps taken to maintain high data quality.

3.1 COLLECTION GUIDELINES

As discussed above, there remains a substantial gap in realistic, competition-level benchmarks for evaluating MAS collaboration. CONTESTWORLD addresses this gap by building a high-quality benchmark from human team competitions, with real-world collaborative challenges. We follow three design principles: (1) cover diverse real-world competition domains that inherently require multiple participants and organize them into a broad taxonomy; (2) select representative and important tasks for each category; and (3) provide a unified, easy-to-deploy evaluation.

3.2 CONTEST SELECTION

We group the 39 competition categories into nine academic domains. The details of each competition are listed in Appendix A.2. **Mathematics.** ARML, HMMT Team Round, Purple Comet! Math Meet, ITYM, PUMaC Power Round, and the World Mathematics Team Championship cover collaborative mathematical problem solving, ranging from short-answer team contests to multi-part proofs and open research-style problems. **Computing & AI.** IIOT, ICPC World Finals, NCCDC, Cybench, IOAI Team Challenge, and NYU CTF Bench evaluate teamwork in algorithmic programming, cybersecurity, network defense, and applied artificial-intelligence tasks. **Natural sciences.** IJSO Practical, EOES/EUSO, International Chemistry Tournament, DOE Science Bowl, and Science Olympiad span experimental science, scientific reasoning, research, and rapid-response problem solving across multiple disciplines. **Law, ethics & public reasoning.** Jessup, DebateBench, APPE Intercollegiate Ethics Bowl, National High School Ethics Bowl, and Vis Moot require teams to analyze cases, construct arguments, respond to opposing positions, and defend legal or ethical reasoning. **Business, economics & finance.** IEO Business Case, CFA Institute Research Challenge, Global Case Competition at Harvard, and the Wharton Investment Competition ask teams to analyze business or investment scenarios and develop evidence-based recommendations. **Physics & astronomy.** IOAA Group Competition, IYPT, and Physics Brawl Online cover collaborative physics and astronomy problem

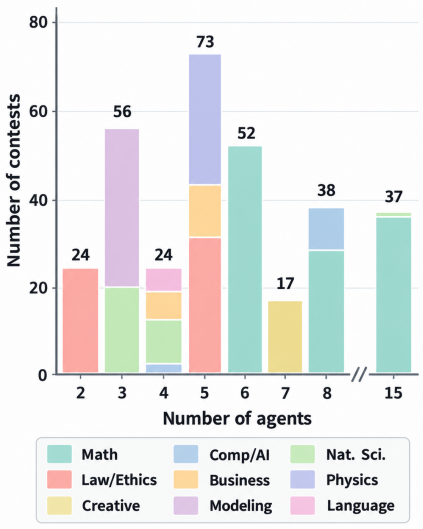


Figure 2: Distribution of the number of agents across 321 contests.

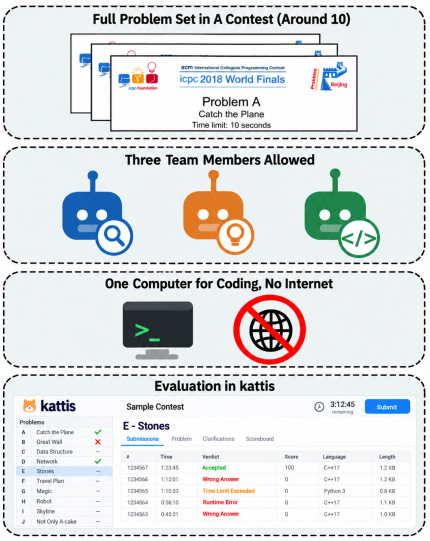


Figure 3: An example setup for an ICPC contest.

solving, including open-ended investigation, presentation, and competitive problem solving. **Interdisciplinary & creative.** MIT Mystery Hunt, Odyssey of the Mind, and QANTA capture broader forms of collaboration involving interconnected puzzles, creative solution design, and knowledge-intensive question answering. **Modeling & applied analysis.** MCM and ICM require small teams to formulate, analyze, and communicate solutions to open-ended real-world problems using mathematical and interdisciplinary modeling. **Language & writing.** IOL Team Contest and World Scholar's Cup Collaborative Writing evaluate collaboration in linguistic reasoning, interpretation of unfamiliar language data, discussion, and written communication.

3.3 Data Preparation

Collection. We collect problems from real-world team competitions using official or publicly available materials (Appendix A.2). For each contest, we preserve the original source and metadata to ensure that every benchmark instance remains traceable to its source. **Unifying problem sets.** We convert heterogeneous competition materials into a unified machine-readable format. Each problem is segmented into its constituent parts and annotated with the corresponding answer, score, solution, or rubric, and evaluation method. Gold answers and solutions are kept hidden from agents and used only for evaluation. We additionally verify the processed instances to correct extraction errors and ensure consistency. The resulting schema standardizes problem statements, answers, scores, deliverable types, and evaluation criteria across all competitions. **Unifying rules and evaluation.** We further translate each competition's original rules into a common agent-evaluation interface, including team size, time constraints, permitted tools, submission format, and evaluation procedure. These constraints are mapped to corresponding agent settings, such as the number of agents, interaction budget, available tools, and required output format. Figure 3 shows an example.

3.4 Environment Construction

A challenge in constructing a benchmark from real competitions is mapping the core components of real-world contests to the simulated environment for agents. To this end, we construct the environment with two components: a simulator that supports diverse contests and an agent orchestration component that allows users to design different collaboration schemes.

Contest simulator. We construct a unified contest simulator that closely aligns with real-world settings. Following Shu et al. (2026), agents execute based on a turn-based procedure where, in each turn, the agents can call a tool. As shown in Figure 3, general tools are allowed for each contest, such as reading a problem, submitting an answer, or chatting with other agents; in addition, task-specific

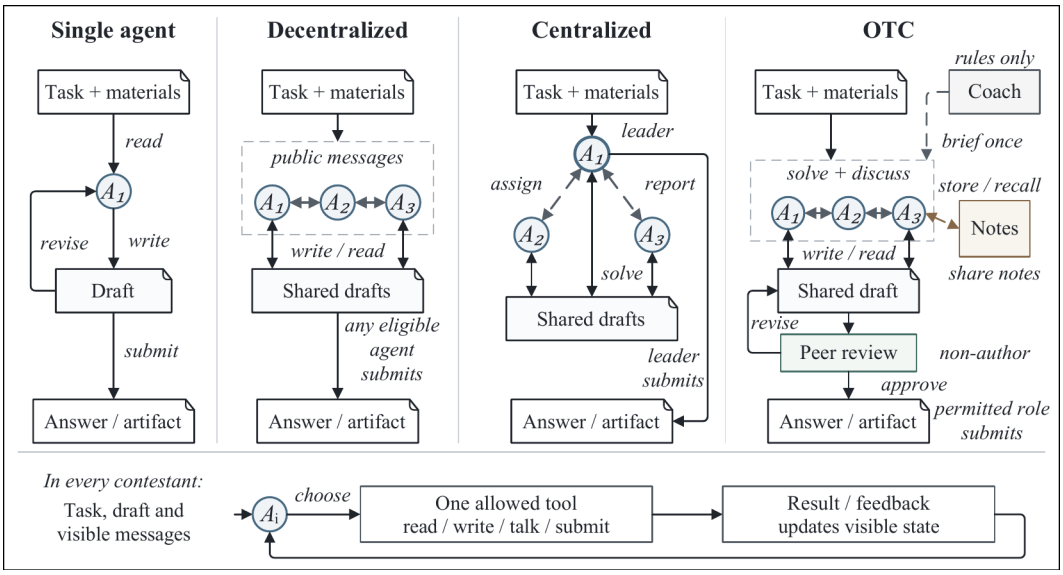


Figure 4: Four baselines in CONTESTWORLD: Single Agent, Decentralized, Centralized, OTC.

tools are loaded from the contest input that the agent can leverage during the contest, such as web search or a calculator (Tables 8 and 9 in Appendix A.4). The evaluator module scores the agent output. We unify scoring through a typed evaluator that links each task to one of four evaluators: gold-answer matching for objective contests, a remote online judge with a local sample-test fallback for programming, a rubric-conditioned LLM judge for open-ended deliverables, and a deck-specific multimodal rubric for presentations. **Agent orchestration.** According to the team size, we spawn multiple agents for the contest. We map each competition's duration to a turn budget, with each turn representing five minutes of contest time. For a competition lasting D minutes, the budget is $T = \min(90, \max(10, \lceil D/5 \rceil))$. Agents act sequentially. On each turn, an agent receives a filtered observation and a short-term summary of the previous turn, then selects an available tool. Only OTC additionally maintains persistent memory across turns.After the maximum number of turns, the system submits the latest nonempty saved answer for each problem (for OTC, only answers approved by another teammate); a problem with no saved draft remains unsubmitted and receives zero credit.

3.5 DATA STATISTICS AND RELEASE

Data statistics. In total, we gather 39 competition categories spanning events held between 2009 and 2026, which become 321 contest sessions and 2,865 problems, as shown in Figure 1b. Team sizes of CONTESTWORLD range from 2 to 15 (Figure 2). The detailed data description can be found in Appendix A.1 and examples of contests are in Appendix D.

Quality control. We conduct a blind test for the evaluators. We randomly selected 14 contest traces generated by GPT and Claude and asked a PhD-level annotator to score them. We then compared the human ratings with the LLM-judge scores. The LLM-judge scores showed strong agreement with human ratings, with a Spearman correlation of $\rho = 0.84$ and a weighted Cohen's $\kappa = 0.76$. The mean absolute error was 0.31 on the 1–5 scale. These results suggest that the LLM score closely tracks human judgments.

Release. The full collection contains 321 contests. We distinguish an evaluation set (test) from a development set (dev). The smaller testmini set is nested within test and contains one contest from each of the 39 tracks; it is not a third disjoint split. Track coverage varies, so the split does not assume five editions per track. Exact membership should be taken from the released split manifests. Our collection scripts support incorporating newly released contests.

Table 1: Overall results across all categories. NT (0–100) averages normalized session scores equally within each category; CS (0–5) averages available session collaboration scores. Avg. is the equal-weight mean of the nine unrounded category values. Single-agent runs have no CS. Bold indicates the best value within each model block (for proprietary models, between the two models); underlining indicates the highest displayed value in each column across all models and methods, including ties.

Model	Method	Math		Comp		NatSci		Law		Biz		Phys		Inter		Model.		Lang		Avg.	
		NT	CS	NT	CS	NT	CS	NT	CS	NT	CS	NT	CS	NT	CS	NT	CS	NT	CS	NT	CS
Open-Source Models with Four Baselines																					
Qwen 3.6	SA	14.4	–	21.4	–	26.4	–	**77.9**	–	70.0	–	37.0	–	61.8	–	43.0	–	45.0	–	44.1	–
	Dec.	19.4	1.7	27.1	0.9	36.8	1.9	46.4	**2.9**	60.8	**2.4**	38.0	1.8	**68.0**	2.2	84.5	4.0	**72.7**	**2.8**	50.4	2.3
	Cent.	19.8	**2.0**	**29.3**	0.9	38.1	2.3	75.1	1.2	71.1	1.4	54.3	1.8	**68.0**	1.4	83.5	**4.3**	60.2	1.8	**55.5**	1.9
	OTC	**28.2**	1.9	28.4	**1.6**	**43.2**	**3.1**	58.4	**2.9**	**72.1**	**2.4**	**58.0**	**2.3**	**68.0**	**2.4**	**89.0**	4.0	45.8	2.0	54.6	**2.5**
GLM 5.3	SA	31.7	–	27.5	–	26.4	–	**83.7**	–	**82.5**	–	55.0	–	**68.0**	–	42.5	–	48.0	–	51.7	–
	Dec.	27.7	2.9	27.4	**2.7**	**36.5**	3.1	54.2	3.6	61.3	**4.0**	54.3	**3.7**	67.2	2.9	**85.5**	**4.5**	46.8	3.5	51.2	3.4
	Cent.	27.2	2.3	26.3	1.8	30.4	2.9	62.4	2.2	69.3	**4.0**	**61.0**	2.8	63.5	2.5	81.0	3.0	50.3	3.5	52.4	2.8
	OTC	**43.0**	**3.3**	**30.3**	2.3	31.4	**3.7**	62.0	**4.0**	68.3	3.3	47.3	3.0	66.5	**3.6**	82.0	4.0	**59.2**	**4.3**	**54.4**	**3.5**
DeepSeek V3.1	SA	13.8	–	17.2	–	27.5	–	**56.4**	–	66.6	–	**47.7**	–	63.6	–	**82.0**	–	20.3	–	43.9	–
	Dec.	16.1	2.2	16.4	2.0	28.2	2.9	54.2	3.4	23.0	3.3	35.7	**2.5**	**68.2**	**3.4**	53.5	3.0	48.7	**3.5**	38.2	2.9
	Cent.	22.3	1.4	17.2	1.3	**33.5**	2.6	39.8	1.3	41.3	1.5	24.7	1.2	59.1	1.5	64.0	1.8	52.8	2.5	39.4	1.7
	OTC	**24.9**	**2.8**	**26.4**	**2.1**	29.4	**3.3**	54.3	**3.6**	**73.5**	**3.6**	40.0	2.2	**68.2**	**3.4**	77.0	**3.3**	**67.2**	**3.5**	**51.2**	**3.1**
Kimi K2.6	SA	33.8	–	22.4	–	33.3	–	65.4	–	**70.4**	–	28.7	–	**68.0**	–	42.8	–	**68.2**	–	48.1	–
	Dec.	27.2	2.3	6.8	**2.1**	31.0	**2.6**	58.4	3.5	50.8	**3.1**	20.3	2.2	62.3	2.8	0.0	2.5	62.5	**3.5**	35.5	2.7
	Cent.	22.6	2.2	14.0	1.4	32.9	2.3	**83.8**	3.5	32.8	**3.1**	**61.3**	2.3	63.4	**3.0**	72.3	2.5	60.5	3.0	49.3	2.6
	OTC	**46.5**	**2.6**	**27.2**	1.5	**37.2**	**2.6**	81.9	**4.1**	69.3	**3.1**	60.0	**2.5**	67.1	2.9	**79.0**	**2.8**	62.1	3.3	**58.9**	**2.8**
Proprietary Models with OTC Method																					
Claude 4.6	OTC	48.3	2.1	35.1	2.3	28.6	2.9	79.9	3.1	64.0	2.3	48.3	2.5	**68.1**	3.0	81.0	2.5	**60.2**	**3.3**	57.0	2.6
GPT-6	OTC	**51.9**	**3.0**	**42.1**	**2.5**	**33.7**	**3.6**	**86.7**	**3.9**	**80.0**	**3.4**	**75.3**	**3.3**	59.1	**3.6**	**82.0**	**3.5**	59.7	3.0	**63.4**	**3.3**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and baselines. We evaluate two frontier LLMs, GPT-6 Luna (OpenAI, 2026) and Claude Sonnet 4.6 (Anthropic, 2026), and four open-source LLMs, Qwen3.6-35B-A3B (Qwen Team, 2026), DeepSeek-V3.1 (DeepSeek-AI, 2024), Kimi-K2.6 (Moonshot AI, n.d.), and GLM-5.3 (GLM-5-Team et al., 2026), on CONTESTWORLD.

We construct four baselines (Figure 4), including (1) one single-agent baseline: *SA*, (2) two multi-agent baselines with different topological structures, including decentralized agents (*Decentralized*) and centralized agents (*Centralized*), and (3) an augmented baseline with a designed coach and memory layer: Open Table Coach (*OTC*). All methods operate under enforced competition rules; only OTC is equipped with persistent memory, while the other baselines retain only short-term memory of the previous turn. We further compare *OTC* with its memory-free and coach-free variants in an ablation study (Section 4.5). Concretely, we construct:

- **Single Agent (SA).** The single agent contains only one agent per contest and performs the entire task. It reads the materials (rules, task, evaluation rubric, etc.) and decides which actions/tools to execute per turn. It determines which problem to address and can revise its draft in response to feedback before submission. All actions are made by the same agent, without any teammate communication or independent peer review.
- **Decentralized.** Decentralized agents share the problems and the contest state; no leader is assigned in advance. They coordinate through public messages and shared drafts. Agents can choose which

problem to work on and use the tool *speak* to communicate. Each team member can read and revise others' work as the team's work develops. Any eligible agent can submit the final draft.

- **Centralized.** The Centralized baseline designates one contestant as the leader, who creates a public plan and assigns tasks to teammates. Before the contest rounds, the leader receives the available questions and task set and produces an opening plan to specify work assignments, task order, and final checks. The leader remains a contestant throughout the run: it can work on any problem or reassign problems as the contest develops. Other agents work on their assigned tasks and report results or unresolved issues through public communication or direct messages. The leader monitors this work and integrates the team's contributions. Only the leader may submit the team's answers.
- **Open Table Coach (OTC).** In the Open Table Coach baseline, before the contest starts, a problem-blind external coach receives the public contest rules, team configurations, and resource budget but no task content. It produces a single opening brief on responsibilities and strategy. The coach cannot inspect the tasks or continue its discussion after the contest starts. Contestants then work in team rounds. After agents read the tasks, they need to assign problems on their own. Afterwards, they organize their work, communicate with each other and develop shared drafts. OTC provides a persistent memory layer for the contestants, comprising private memory, team memory, and public memory. These memories support the contestants during the contest; they are not maintained by the coach, which exits after its opening brief. Before the final submission, the latest answer drafts must be approved by another contestant. Only a contestant with a permitted submission role can submit.

Metrics. We use Normalized Task Score (NT) and Collaboration Score (CS) as evaluation metrics.

- **NT.** Normalized Task Score measures task performance on a scale of 0–100. We normalize each contest's score by its maximum available score and average the normalized scores equally within each category: $\mathrm{NT}_c = \frac{100}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \frac{r_s}{r_s^{\max}}$, where $\mathcal{S}_c$ is the set of contests in category c, r_s is the earned score, and $r_s^{\max}$ is the maximum available score for session s. Non-submissions receive zero task score and remain in the denominator. The overall average is the unweighted mean of the nine unrounded category scores. Reported NT values are rounded to one decimal place.
- **CS.** Following Zhu et al. (2025), we apply the same collaboration score with GPT-6 as the judge. Collaboration Score is between 0 and 5, evaluated by two aspects: communication quality and planning effectiveness. For each contest, an LLM judge evaluates communication using the task description, agent profiles, messages among agents and task results. The communication score assesses decision-making, clarity, precision, and consistency with the assigned roles. Planning is evaluated separately using an LLM judge on agent profiles, tool-use counts, workload distribution, and strategic coordination. Each dimension is rated from 1 (very poor) to 5 (excellent). If there is no communication during the contest, the communication rating is set to zero. The score is calculated as follows: $\mathrm{CS}_s = \frac{C_s^{\mathrm{comm}} + C_s^{\mathrm{plan}}}{2}$, where C_s^{comm} denotes the communication rating and C_s^{plan} denotes the planning rating.

CS is not applicable to SA runs since there is no communication in the Single Agent baseline.

4.2 Overall Results

Table 1 reports results across nine domains. OTC achieves the highest average NT on three of the four open-source models: 54.6 on Qwen3.6, 54.4 on GLM-5.3, 51.2 on DeepSeek-V3.1, and 58.9 on Kimi-K2.6. These exceed the corresponding single-agent scores by 10.5, 2.7, 7.3, and 10.8 points. On Qwen3.6, centralized coordination outperforms OTC (55.5 vs. 54.6). Centralized coordination consistently outperforms decentralized coordination in average NT across all four open-source models. However, decentralized coordination underperforms SA on GLM, DeepSeek, and Kimi, and centralized coordination also underperforms SA on DeepSeek. Multiple agents therefore do not guarantee better task performance.

Under OTC, Kimi-K2.6 leads the open-source models overall and in mathematics, while DeepSeek-V3.1 is strongest among them in language and Qwen3.6 in natural science. GPT-6 achieves the highest overall NT (63.4), compared with 58.9 for Kimi and 57.0 for Claude, with particularly strong scores in law, business, and physics. Modeling and law/ethics are generally high-scoring

Table 2: Primary communication and coordination issues in sampled OTC trajectories.

Category	Definition	Representative Example	Count	Ratio
Inadequate verification	A reviewer endorses an incorrect or insufficiently checked candidate without resolving its substantive defect.	WMTC 2018 Junior, Q3 (GLM-5.3): the reviewer repeats $n = 11$ after writing the correct equation and states, "All 14 verified independently. Approve." The correct answer is 12.	13	29.5%
Feedback not resolved	A concrete objection or correction is not adequately reflected in the final submitted version.	ARML Power Spring 2023, Problem 7 (Qwen3.6): a reviewer supplies the two-step counterexample $199 \to 100 \to 1$, but the revised submission retains 199.	9	20.5%
Workflow / handoff stall	Repeated coordination, preparation, or tool use fails to advance the task to the required next stage.	Mystery Hunt 2012 (Kimi-K2.6): repeated coordination continues until a teammate reports, "We are at turn 65 of 90 with 0 attempts and no candidate answer."	8	18.2%
Source or task misunderstanding	A misinterpretation of the source or task requirements persists or is reinforced through interaction.	EOES T1 (DeepSeek-V3.1): a review requests experimental data for fields marked "NOT MEASURED"; subsequent revisions introduce unsupported numerical readings that receive approval.	3	6.8%
Incomplete coverage	Required task components remain absent despite task assignment or review.	Belgrade WUDC 2022, Round 8 (Qwen3.6): the review approves the Prime Minister speech, while the final submission omits the required Deputy Prime Minister speech.	11	25.0%
Total			**44**	**100%**

domains, with several model–method combinations exceeding 80 NT; computing/AI, natural science, and mathematics remain challenging. These results show that performance depends on both the collaboration protocol and the domain.

4.3 Error Analysis of Communication

We randomly sample 100 OTC trajectories across four backbone models and manually analyze their communication failures. Among the identified errors, the most common issue is *inadequate verification* (29.5%): reviewers often approve a solution without independently checking its critical reasoning steps. As a result, communication can reinforce an incorrect answer rather than correct it.

Another notable failure is that useful information does not reliably propagate to the final solution. In 20.5% of the cases, agents provide valid feedback that is not incorporated into later revisions, while 25.0% involve incomplete coverage of required task components. These results suggest that the main challenge is whether communicated information is correctly verified, propagated, and integrated into the final submission.

4.4 Why Do Multi-Agent Systems Help?

We further analyze 321 contests from Qwen, GLM, and DeepSeek to understand where the gains of multi-agent systems come from. We observe the following three findings:

Teams cover the task better. A clear difference between single-agent and multi-agent runs is whether they manage to attempt the full contest. When a single agent receives a score of zero, it often leaves a large fraction of the questions unanswered: its answer coverage drops from 70.1% to only 24.1% on zero-score runs (Table 3). This pattern is much weaker for teams. In other words, single agents often fail because they do not finish the task, whereas teams are much more likely to finish but may still answer incorrectly. This suggests that one important benefit of multiple agents is straightforward: *they can divide the work and reduce the chance of leaving problems untouched*.

Table 3: Answer coverage (%) for zero/non-zero-score runs; gap in percentage points. Qw. = Qwen; DS = DeepSeek; SA = single agent; Cent. = centralized.

Model	Method	Zero	Non-zero	Gap
Qw.	SA	24.1	70.1	−46.0
GLM	Cent.	56.0	79.9	−23.9
Qw.	OTC	85.1	91.5	−6.4
Qw.	Cent.	94.2	94.7	−0.5
DS	Cent.	100.0	94.9	+5.1

Table 4: Component ablations on the CONTESTWORLD subset (Qwen3.6). M = memory; C = coach; w/o = without.

Variant	NT	CS
OTC	54.6	2.5
OTC w/o M	49.5	2.1
OTC w/o C	52.5	2.4
OTC w/o M+C	52.7	1.7

More turns do not hurt teams in the same way. We examine the within-competition relationship between turns and NT, pooling results across Qwen, GLM, and DeepSeek. For a single agent, longer runs are associated with lower NT ($r = -0.301$), suggesting that an agent tends to spend more turns when it is struggling with the task. This relationship largely disappears for teams: across six centralized and decentralized settings, correlations range from -0.10 to -0.02. For a team, therefore, additional turns can simply mean that different members are doing more useful work, rather than that the system is stuck.

The coach makes the team more selective. The coach provides a different benefit: instead of simply increasing coverage, it can help the team avoid low-quality attempts. For GLM, adding the coach reduces coverage from 85.9% to 74.0%, but increases NT from 46.2 to 53.3. Removing the coach changes NT by -7.1 for GLM, -8.8 for DeepSeek, and -2.1 for Qwen (Table 4), indicating a smaller effect for Qwen. Thus, collaboration provides two complementary advantages: multiple agents help cover more of the task, while the coach can help the team decide which answers are worth submitting.

4.5 ABLATION STUDY

We conduct three ablation experiments to examine the contribution of memory and the coach in OTC on Qwen3.6:

- **OTC without memory.** This variant retains the one-time pre-contest coach brief but removes the contestants' persistent memory layer (private, team, and public memory).
- **OTC without coach.** This variant retains memory but removes the coach.
- **OTC without coach and memory.** This variant removes both components.

We compare these with the full OTC framework to understand how each component contributes to OTC performance (Table 4). All three ablated variants obtain lower NT and CS scores than full OTC. Removing memory produces the largest drop in NT, from 54.6 to 49.5, while CS decreases from 2.5 to 2.1 (a drop of 0.4). Removing both memory and the coach produces the largest drop in CS, from 2.5 to 1.7 (a drop of 0.8). These results indicate that memory contributes to task performance in this setting. The task-performance effects are not additive. Removing both components yields a higher NT than removing memory alone (52.7 vs. 49.5). With memory available, adding the coach improves NT by 2.1 points; without memory, adding the coach lowers NT by 3.2 points. This pattern suggests that the coach's task-performance benefits may depend on memory support.

5 CONCLUSION

We introduce CONTESTWORLD, a benchmark of 321 contests across 39 competition categories and nine domains for evaluating multi-agent collaboration under contest rules. The best model merely performs around 60%, demonstrating the challenges of CONTESTWORLD. Results across six models show substantial variation across domains. Our analysis shows that teams primarily benefit from improved task coverage and selection, while communication failures and component ablations highlight the importance of verification, feedback integration, and memory, motivating further work on reliable team coordination.

AI USE STATEMENT

We used generative AI tools only to check the manuscript for typos, grammar, and formatting issues (e.g., reference formatting and LaTeX layout). We have not used generative AI tools for any of the tasks for which ICLR requires disclosure, including generating data, developing methods or theoretical frameworks, formulating mathematical claims, designing or implementing experiments, or interpreting results. All suggested edits were reviewed and accepted or rejected manually by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release all code, data, and traces upon acceptance. We list the prompts, parameters, data examples, and settings in the appendix to support reproduction. Furthermore, we construct the code in a flexible way so that we can rapidly adapt to newer data or models.

REFERENCES

American Regions Mathematics League. ARML Competition 2023 Contest File, 2023. URL https://www.arml.com/ARML/arml_2019/public_contest_files/2023_contest_file/ARML_2023Contest.pdf. Accessed: 2026-09-26.

American Regions Mathematics League. ARML Power Contest Materials, n.d. URL https://www.arml.com/ARML/arml_2019/page/index.php?page=5&page_type=public&show_page=samples. Accessed: 2026-09-26.

Anthropic. Claude Sonnet 4.6 System Card. Technical report, Anthropic, February 2026. URL https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

Association for Practical and Professional Ethics. APPE Intercollegiate Ethics Bowl Cases, Rules, and Guidelines, n.d. URL https://www.appe-ethics.org/cases-rules-guidelines/. Accessed: 2026-09-26.

Canadian Tire Corporation. 2014 Annual Report, 2014. URL https://s201.q4cdn.com/326551073/files/doc_financials/en/2014/annual/2014-Annual-Report-EN.pdf. Accessed: 2026-09-26.

Canadian Tire Corporation. Third Quarter 2015 Report, 2015. URL https://s22.q4cdn.com/405442328/files/doc_financials/en/2015/q3/Q3-2015-Shareholder-Report.pdf. Accessed: 2026-09-26.

CFA Institute. Past Research Challenge Champions, n.d. URL https://www.cfainstitute.org/insights/events/research-challenge/past-champions. Accessed: 2026-09-26.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for LLM-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557, 2025.

COMAP. MCM/ICM Previous Contests and Problems, n.d. URL https://www.contest.comap.org/undergraduate/contests/mcm/previous-contests.php. Accessed: 2026-09-26.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Jacob Eisenstein, Fantine Eri Huot, Adam Fisch, Jonathan Berant, and Mirella Lapata. MT-PINGEVAL: Evaluating multi-turn collaboration with private information games. In *Third Conference on Language Modeling*, 2026.

European Olympiad of Experimental Science. Previous Olympiads, n.d. URL https://www.eoes.science/Previous%20olympiads/previous.html. Accessed: 2026-09-26.

Wei Fan, Xinjie Shen, Xudong Guo, Jianhong Tu, Yang Su, Yinger Zhang, Lianghao Deng, Fengyu Wang, Baohua Dong, Yangqiu Song, and Dayiheng Liu. E-Commerce Bench: Evaluating LLM agents on long-horizon autonomous business operation. *arXiv preprint arXiv:2608.30730*, 2026.

GLM-5-Team et al. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. URL https://arxiv.org/abs/2602.15763.

Global Case Competition at Harvard. Past Editions, n.d. URL https://www.thecasecompetition.org/past-editions. Accessed: 2026-09-26.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Fei-Fei Li, Jianfeng Gao, and Hoi Vo. MindAgent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3154–3183, 2024.

Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2026. URL https://arxiv.org/abs/2402.03578v3.

Harvard–MIT Mathematics Tournament. Past Tournaments: Problems and Results, n.d. URL https://www.hmmt.co/www/archive/results. Accessed: 2026-09-26.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, pp. 35980–36008, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Sihao Hu, Tiansheng Huang, Gaowen Liu, Ramana Rao Kompella, Fatih Ilhan, Selim Furkan Tekin, Yichang Xu, Zachary Yahn, and Ling Liu. A survey on large language model-based game agents. *ACM Computing Surveys*, 2026. URL https://arxiv.org/abs/2404.02039v5. Accepted for publication; forthcoming.

International Chemistry Tournament. Problems, n.d. URL http://ichto.org/en/problems/. Access attempted: 2026-09-26.

International Collegiate Programming Contest. 2025 ICPC World Finals Baku Brochure, 2025. URL https://icpc.global/community/history/brochures/world-finals-2025-brochure.pdf. Access attempted: 2026-09-26.

International Economics Olympiad. Preparation Materials, n.d. URL https://ieo-official.org/prepare. Accessed: 2026-09-26.

International Informatics Olympiad in Teams. Past Editions and Competition Tasks, n.d. URL https://iio.team/editions. Accessed: 2026-09-26.

International Junior Science Olympiad. Downloads, n.d. URL https://ijsoweb.org/downloads. Accessed: 2026-09-26.

International Law Students Association. 2024 Philip C. Jessup International Law Moot Court Competition, 2024. URL https://www.ilsa.org/jessup-year/jessup-2024/. Accessed: 2026-09-26.

International Linguistics Olympiad. Problems by Year, n.d. URL https://ioling.org/problems/by_year/. Accessed: 2026-09-26.

International Olympiad in Artificial Intelligence. 2025 Team Challenge: Factory of the Future Grand Challenge, 2025. URL https://ioai-official.org/wp-content/uploads/2025/06/2025-IOAI-Team-Challenge-Factory-of-the-Future-Grand-Challenge-.pdf. Accessed: 2026-09-26.

International Olympiad on Astronomy and Astrophysics. Problems from Past IOAA, n.d. URL https://ioaastrophysics.org/resources/problems-from-past-ioaa. Accessed: 2026-09-26.

International Tournament of Young Mathematicians. Past Problems Archive, n.d. URL https://www.itym.org/archive. Accessed: 2026-09-26.

International Young Physicists' Tournament. Problems, n.d. URL https://iypt.org/problems/. Accessed: 2026-09-26.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pp. 25479–25519, 2025.

Paul J. Karafiol, Thomas Kilkelly, Micah Fogel, and Paul Dreyer. *American Regions Math League & Power & Local Contests 2009–2014*. ARML, 2015. ISBN 9781943091676. URL https://www.arml.com/ARML/arml_2019/public_contest_files/2009_2014_book/ARML_2009_2014.pdf. Edited by Chris Jeuell.

Kattis. Kattis Problem Archive, n.d. URL https://open.kattis.com/. Access attempted: 2026-09-26.

Arpandeep Khatua, Hao Zhu, Peter Tran, Arya Prabhudesai, Frederic Sadrieh, Johann Kaspar Lieberwirth, Xinkai Yu, Yicheng Fu, Michael J Ryan, Jiaxin Pei, and Diyi Yang. CooperBench: Why coding agents cannot be your teammates yet. In *Third Conference on Language Modeling*, 2026.

Yubin Kim, Chanwoo Park, Taehan Kim, Eugene Park, Samuel Schmidgall, Salman Rahman, Chunjong Park, Cynthia Breazeal, Xin Liu, Hamid Palangi, Hae Won Park, and Daniel McDuff. TeamBench: Evaluating agent coordination under enforced role separation. *arXiv preprint arXiv:2605.07073*, 2026.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

MIT Mystery Hunt. Puzzle Archive, n.d. URL https://puzzles.mit.edu/. Access attempted: 2026-09-26.

Moonshot AI. Kimi-K2.6. Hugging Face model card, n.d. URL https://huggingface.co/moonshotai/Kimi-K2.6. Accessed: 2026-09-26.

National Collegiate Cyber Defense Competition. National Collegiate Cyber Defense Competition, n.d. URL https://www.ccdc.io/. Accessed: 2026-09-26.

National High School Ethics Bowl. Case Library, n.d. URL https://nhseb.org/case-library. Accessed: 2026-09-26.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Odyssey of the Mind. Past Problems, n.d. URL https://odysseyofthemind.com/past-problems/. Accessed: 2026-09-26.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI. GPT-6 Luna. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-luna. Accessed: 2026-09-26.

Physics Brawl Online. Last Years of Physics Brawl Online, n.d. URL https://physicsbrawl.org/archive. Accessed: 2026-09-26.

Princeton University Mathematics Competition. Problem and Result Archive, n.d. URL https://pumac.princeton.edu/archives. Accessed: 2026-09-26.

Purple Comet! Math Meet. Old Contests: Problems and Answers, n.d. URL https://www.purplecomet.org/answers. Accessed: 2026-09-26.

QANTA Project. QANTA Quiz Bowl Dataset (2021.12.20 release), 2021. URL https://sites.google.com/view/qanta/resources. Accessed: 2026-09-26.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15174–15186, 2024.

Ao Qu, Han Zheng, Zijian Zhou, Yihao Yan, Yihong Tang, Shao Yong Ong, Fenglu Hong, Kaichen Zhou, Chonghe Jiang, Minwei Kong, et al. CORAL: Towards autonomous multi-agent evolution for open-ended discovery. *arXiv preprint arXiv:2604.01658*, 2026.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*, 2019.

Science Olympiad. Solar System 2017–2018 Annotated Example Exam, n.d. URL https://www.soinc.org/sites/default/files/uploaded_files/082317_SS1718_AnnotatedExampleExam.pdf. Accessed: 2026-09-26.

Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberly Milner, Boyuan Chen, Max Yin, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, Ramesh Karri, and Muhammad Shafique. NYU CTF bench: A scalable open-source benchmark dataset for evaluating LLMs in offensive security. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57472–57498, 2024. doi: 10.52202/079017-1832.

Jiayi Sheng, Luna Lyu, Jikai Jin, Tanglin Xia, Alex Gu, James Zou, and Pan Lu. Solving inequality proofs with large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4559.

Raphael Shu, Yusen Zhang, Young Min Cho, Jin Mo Yang, Yuan Yuan, Wenliang Zheng, Sharath Chandra Guntuku, Lyle Ungar, Zhou Yu, and Rui Zhang. AgentWorld: Benchmarking long-horizon collaboration of multi-agent LLMs. In *Third Conference on Language Modeling*, 2026.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. Collab-Overcooked: Benchmarking and evaluating large language models as collaborative agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4922–4951, 2025.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Arina Turkatenko, Pontus Lars Erik Saito Stenetorp, Mikel Artetxe, and Marta Ruiz Costa-jussà. Linguini: A benchmark for language-agnostic linguistic reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 182440–182460, 2025. doi: 10.52202/085713-5486.

Utkarsh Tiwari, Aryan Seth, Adi Mukherjee, Kaavya Mer, Kavish, and Dhruv Kumar. DebateBench: A challenging long context reasoning benchmark for large language models. *arXiv preprint arXiv:2502.06279*, 2025. URL https://arxiv.org/abs/2502.06279.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*, 2025.

United States Department of Energy. National Science Bowl High School Sample Questions, n.d. URL https://science.osti.gov/wdts/nsb/Regional-Competitions/Resources/HS-Sample-Questions. Accessed: 2026-09-26.

Wharton Global Youth Program. Case Study for 2019–2020, 2019. URL https://globalyouth.wharton.upenn.edu/investment-competition/previous-winners/previous-case-study-2019-2020/. Accessed: 2026-09-26.

Wharton Global Youth Program. Case Study for 2020–2021, 2020. URL https://globalyouth.wharton.upenn.edu/investment-competition/previous-winners/previous-case-study-2020-2021/. Accessed: 2026-09-26.

Wharton Global Youth Program. Case Study for 2021–2022, 2021. URL https://globalyouth.wharton.upenn.edu/investment-competition/previous-winners/case-study-for-2021-2022/. Accessed: 2026-09-26.

Willem C. Vis International Commercial Arbitration Moot. Vis Moot, n.d. URL https://www.vismoot.org/. Access attempted: 2026-09-26.

World Mathematics Team Championship. 2018 Advanced Level Problems, 2018a. URL https://wmtc.international/wp/wp-content/uploads/2018/12/2018_Advanced_Indivi_Relay_Team_Problems.pdf. Accessed: 2026-09-26.

World Mathematics Team Championship. 2018 Intermediate Level Problems, 2018b. URL https://wmtc.international/wp/wp-content/uploads/2018/12/2018_Intermediate_Indivi_Relay_Team_Problems.pdf. Accessed: 2026-09-26.

World Mathematics Team Championship. 2018 Junior Level Problems, 2018c. URL https://wmtc.international/wp/wp-content/uploads/2018/12/2018_Junior_Indivi_Relay_Team_Problems.pdf. Accessed: 2026-09-26.

World Scholar's Cup. Theme and Guiding Questions, n.d.a. URL https://themes.scholarscup.org. Accessed: 2026-09-26.

World Scholar's Cup. Team Events: Collaborative Writing, n.d.b. URL https://www.scholarscup.org/events/. Accessed: 2026-09-26.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. LLM-based multi-agent systems: Techniques and business perspectives. *arXiv preprint arXiv:2411.14033*, 2024.

Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Haoxiang Yang, Aolin Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Kenny Oseleononmen, Dan Boneh, Daniel Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *International Conference on Learning Representations*, pp. 80893–81042, 2025a.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-Memory: Tracing hierarchical memory for multi-agent systems. In *Advances in Neural Information Processing Systems*, volume 38, pp. 14587–14617, 2025b. doi: 10.52202/085713-0439.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, and Wenyuan Jiang. Silo-Bench: A scalable environment for evaluating distributed coordination in multi-agent LLM systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29379–29398, 2026.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, Peiyao Sheng, Zixuan Wang, Wenhao Chai, Aleksandra Korolova, Peter Henderson, Sanjeev Arora, Pramod Viswanath, Jingbo Shang, and Saining Xie. LiveCodeBench pro: How do olympiad medalists judge LLMs in competitive programming? In *Advances in Neural Information Processing Systems*, volume 38, pp. 116099–116162, 2025. doi: 10.52202/085713-3502.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025.

Kaijian Zou, Aaron Xiong, Yunxiang Zhang, Xinliang Frederick Zhang, Yueqi Ren, Jirong Yang, Ayoung Lee, Shitanshu Bhushan, and Lu Wang. LiveOIBench: Can large language models outperform human contestants in informatics olympiads? In *Forty-third International Conference on Machine Learning*, 2026.

APPENDIX CONTENTS

A Benchmark Details **17**

A.1 Data description 17

A.2 Competition descriptions 20

A.3 Construction and Data Preparation 24

A.4 Agent action and tool interface 25

B Agent Prompt **29**

B.1 Contestant System Prompt 29

B.2 Contestant User Prompt 30

B.3 Memory Guidance 31

B.4 Coach System Prompt 32

B.5 Coach User Prompt 32

C Algorithms **33**

C.1 Open-Table Coach Execution 33

C.2 Automatic Memory Update and Context Construction 33

C.3 Independent Review and Candidate Revision 34

D Representative contest examples **34**

A BENCHMARK DETAILS

A.1 DATA DESCRIPTION

Table 5 reports the 39 active tracks and their 2,865 question/task units across source years 2009–2026. Table 6 reports the testmini subset, with 527 source-level questions or open-ended tasks. For the full catalog, a unit is a declared scoring part or question; a packet without a verified internal breakdown contributes one task unit. The testmini counts use the source's top-level numbered questions when these can be identified. Accordingly, the totals describe different levels of available granularity and should not be subtracted from one another. These inventory counts are descriptive only: they are not aggregation weights or score denominators. For NT, normalized session scores are averaged equally within each of the nine domains. CS averages available session collaboration scores within each domain. Overall NT and CS are the unweighted means of the corresponding nine unrounded domain scores; the 39 tracks are not independently weighted in the overall average.

Table 5: Full catalog: all 39 active tracks, 2,865 question/task units, source years 2009–2026.

Competition	What it is	Years in collection	Questions
ARML (Local)	Regional short-answer team round of the American Regions Math League	2009–2014	65
ARML Nat'l (Power)	ARML National Meet Power Round: long collaborative proof/exploration packet	2009–2014, 2016–2019, 2023	11
ARML Nat'l (Team)	ARML National Meet Team Round: divide-and-conquer short-answer set	2009–2014, 2016–2019, 2023	110
ARML Power	ARML Power Contest (mail-in): collaborative multi-part proof packet	2018–2025	58
NCCDC	National Collegiate Cyber Defense Competition: defend a live network against a red team	2013–2021, 2025–2026	12
CFA Research Challenge	CFA Institute Research Challenge: team equity research report on an assigned company	2016	1
Cybench	Cybench: capture-the-flag style cybersecurity benchmark tasks	2022–2024	38
DebateBench	DebateBench: transcribed British Parliamentary (WUDC/BP) debate rounds	2020–2025	24
EOES/EUSO	European science-olympiad hands-on lab experiment task	2013, 2015, 2017, 2021, 2023–2025	12
Ethics Bowl (APPE)	APPE Intercollegiate Ethics Bowl: structured argument on applied-ethics cases	2021–2026	9
Ethics Bowl (NHSEB)	National High School Ethics Bowl: high-school applied-ethics case debate	2012–2025	14
Fyziklani	Physics Brawl Online: timed team physics problem-solving tournament	2012–2025	858
GCCH	Global Case Competition at Harvard: team business-case analysis and pitch	2019–2021, 2023–2026	7
HMMT (Team)	Harvard-MIT Math Tournament Team Round: collaborative math problem set	2013–2026	140
IChTO	International Chemistry Tournament: team chemistry problem-set tournament	2017–2020, 2022–2026	9
ICM	Interdisciplinary Contest in Modeling: multi-day open-ended modeling report	2015–2017, 2021–2022, 2024–2025	19
ICPC	ICPC World Finals: collegiate algorithmic programming, automated judge	2012–2025	158
IEO	International Economics Olympiad: team business/economics case round	2021–2025	5
IIOT	International Informatics Olympiad in Teams: team algorithmic programming	2017, 2021–2023	31

Table 5 – continued

Competition	What it is	Years in collection	Questions
IJSO (Practical)	International Junior Science Olympiad hands-on lab practical round	2014, 2016–2019, 2021–2023	8
IOAA (Group)	International Olympiad on Astronomy & Astrophysics team/group data round	2014–2015, 2021, 2025	4
IOAI (Team)	International Olympiad in AI Team Challenge: team applied-AI task	2025	1
IOL (Team)	International Linguistics Olympiad team round: linguistic puzzle solving	2014–2015, 2018, 2021, 2023, 2025	6
ITYM	International Tournament of Young Mathematicians: open research problems, presented & defended	2011–2025	15
IYPT	International Young Physicists' Tournament: open research physics, "physics fights"	2011–2026	16
Jessup	Philip C. Jessup Moot Court: simulated int'l law dispute, memorials & oral argument	2024	1
MCM	Mathematical Contest in Modeling: multi-day open mathematical modeling report	2016–2017, 2021–2022, 2024–2025	17
Mystery Hunt	MIT Mystery Hunt: large collaborative meta-puzzle hunt	2010, 2012, 2015–2016	29
NYU CTF	NYU CTF Bench (CSAW): capture-the-flag cybersecurity tasks	2013–2023	44
Odyssey of the Mind	Long-term creative problem-solving performance synopses	2010–2026	17
PUMaC Power	Princeton Math Competition Power Round: collaborative proof packet	2009–2013, 2015–2017, 2019–2021, 2023	12
Purple Comet	Purple Comet! Math Meet: untimed online team math relay	2013–2026	700
QANTA	Academic quiz-bowl question answering	2011–2020	100
Science Bowl	DOE Science Bowl: buzzer-style team science quiz	2016, 2018–2022	120
Science Olympiad	Science Olympiad (USA): multi-event team science competition	2018	100
Vis Moot	Willem C. Vis Moot: simulated arbitration dispute, memorials & oral argument	2019, 2021–2026	7
Wharton Investment	Wharton Global HS Investment Competition: team portfolio-management case	2019–2021	3
WMTC	World Mathematics Team Championship: linked/standalone team problem set	2018	42
WSC Writing	World Scholar's Cup: collaborative writing/discussion round	no canonical year (src. 2024–2026)	42
Total			**2,865**

Table 6: Testmini: one edition per active track, 527 questions/tasks total.

Competition	What it is	Year used	Questions
ARML (Local)	Regional short-answer team round of the American Regions Math League	2013	10
ARML Nat'l (Power)	ARML National Meet Power Round: long collaborative proof/exploration packet	2009	10
ARML Nat'l (Team)	ARML National Meet Team Round: divide-and-conquer short-answer set	2011	10

Table 6 – continued

Competition	What it is	Year used	Questions
ARML Power	ARML Power Contest (mail-in): collaborative multi-part proof packet	2023	32
NCCDC	National Collegiate Cyber Defense Competition: defend a live network against a red team	2025	1
CFA Research Challenge	CFA Institute Research Challenge: team equity research report on an assigned company	2016	1
Cybench	Cybench: capture-the-flag style cybersecurity benchmark tasks	2022	8
DebateBench	DebateBench: transcribed British Parliamentary (WUDC/BP) debate rounds	2022	5
EOES/EUSO	European science-olympiad hands-on lab experiment task	2021	2
Ethics Bowl (APPE)	APPE Intercollegiate Ethics Bowl: structured argument on applied-ethics cases	2022	31
Ethics Bowl (NHSEB)	National High School Ethics Bowl: high-school applied-ethics case debate	2021	16
Fyziklani	Physics Brawl Online: timed team physics problem-solving tournament	2020	63
GCCH	Global Case Competition at Harvard: team business-case analysis and pitch	2021	1
HMMT (Team)	Harvard-MIT Math Tournament Team Round: collaborative math problem set	2014	10
IChTO	International Chemistry Tournament: team chemistry problem-set tournament	2019	12
ICM	Interdisciplinary Contest in Modeling: multi-day open-ended modeling report	2015	1
ICPC	ICPC World Finals: collegiate algorithmic programming, automated judge	2018	11
IEO	International Economics Olympiad: team business/economics case round	2024	1
IIOT	International Informatics Olympiad in Teams: team algorithmic programming	2021	8
IJSO (Practical)	International Junior Science Olympiad hands-on lab practical round	2019	3
IOAA (Group)	International Olympiad on Astronomy & Astrophysics team/group data round	2015	1
IOAI (Team)	International Olympiad in AI Team Challenge: team applied-AI task	2025	1
IOL (Team)	International Linguistics Olympiad team round: linguistic puzzle solving	2025	1
ITYM	International Tournament of Young Mathematicians: open research problems, presented & defended	2014	11
IYPT	International Young Physicists' Tournament: open research physics, "physics fights"	2026	17
Jessup	Philip C. Jessup Moot Court: simulated int'l law dispute, memorials & oral argument	2024	1
MCM	Mathematical Contest in Modeling: multi-day open mathematical modeling report	2024	1
Mystery Hunt	MIT Mystery Hunt: large collaborative meta-puzzle hunt	2012	1
NYU CTF	NYU CTF Bench (CSAW): capture-the-flag cybersecurity tasks	2017	7
Odyssey of the Mind	Long-term creative problem-solving performance synopses	2019	1

Table 6 – continued

Competition	What it is	Year used	Questions
PUMaC Power	Princeton Math Competition Power Round: collaborative proof packet	2015	22
Purple Comet	Purple Comet! Math Meet: untimed online team math relay	2017	50
QANTA	Academic quiz-bowl question answering	2015	10
Science Bowl	DOE Science Bowl: buzzer-style team science quiz	2016	20
Science Olympiad	Science Olympiad (USA): multi-event team science competition	2018	100
Vis Moot	Willem C. Vis Moot: simulated arbitration dispute, memorials & oral argument	2019	1
Wharton Investment	Wharton Global HS Investment Competition: team portfolio-management case	2021	1
WMTC	World Mathematics Team Championship: linked/standalone team problem set	2018	42
WSC Writing	World Scholar's Cup: collaborative writing/discussion round	no canonical year	3
Total			**527**

A.2 Competition descriptions

Table 7: Collaboration capabilities associated with the 39 competition tracks.

Competition	Short name	Collaboration capability
ARML Local	ARML-L	Parallel problem solving
ARML National Meet — Power Round	ARML-NP	Proof collaboration
ARML National Meet — Team Round	ARML-T	Task allocation
ARML Power Contest	ARML-P	Proof planning
National Collegiate Cyber Defense Competition (NCCDC)	NCCDC	Defense prioritization; Role coordination
CFA Institute Research Challenge	CFA-RC	Research division
Cybench	Cybench	Hypothesis sharing
DebateBench (WUDC / BP)	DebateB	Case planning; Speech coordination
EOES / EUSO	EOES	Data sharing; Protocol planning
APPE Intercollegiate Ethics Bowl	APPE	Argument coordination; Response planning
National High School Ethics Bowl	NHSEB	Argument comparison; Response coordination
Physics Brawl Online (Fyziklání)	Fyziklani	Task allocation
Global Case Competition at Harvard	GCCH	Case analysis; Recommendation alignment
HMMT — Team Round	HMMT	Task allocation
International Chemistry Tournament	IChTO	Research division
Interdisciplinary Contest in Modeling (ICM)	ICM	Task decomposition
ICPC World Finals	ICPC	Workstation scheduling
International Economics Olympiad	IEO-Case	Case analysis; Recommendation alignment
IIOT — International Informatics Olympiad in Teams	IIOT	Parallel solution design
IJSO Practical	IJSO	Evidence sharing
International Olympiad on Astronomy and Astrophysics	IOAA-G	Task coordination
International Olympiad in Artificial Intelligence — Team Challenge	IOAI-T	Task decomposition

Continued on next page

Competition	Short name	Collaboration capability
International Linguistics Olympiad	IOL-T	Pattern comparison
International Tournament of Young Mathematicians	ITYM	Research division
International Young Physicists' Tournament	IYPT	Research division
Philip C. Jessup International Law Moot Court Competition	Jessup	Issue allocation; Argument consistency
Mathematical Contest in Modeling (MCM)	MCM	Task decomposition
MIT Mystery Hunt	MIT-Hunt	Clue sharing
NYU CTF Bench (CSAW)	NYU-CTF	Hypothesis sharing
Odyssey of the Mind	Odyssey	Creative ideation; Script coordination
PUMaC Power Round	PUMaC-P	Proof strategy; Dependency tracking
Purple Comet! Math Meet	P.Comet	Parallel problem solving
QANTA Quiz Bowl	QANTA	Clue pooling; Answer consensus
DOE Science Bowl	SciBowl	Answer comparison; Answer selection
Science Olympiad (USA)	SciOlymp	Question allocation
Willem C. Vis International Commercial Arbitration Moot	VisMoot	Issue allocation; Argument consistency
Wharton Global High School Investment Competition	Wharton	Research division
World Mathematics Team Championship	WMTC	Parallel problem solving
World Scholar's Cup — Collaborative Writing	WSC-W	Shared planning; Peer editing

ARML (Local): ARML Local includes a team round of short-answer mathematics problems. The benchmark uses archived local team-round packets, whose question counts follow the paper for each year. Agents submit a shared answer sheet, scored against the corresponding answer material (Karafiol et al., 2015).

ARML Nat'l (Power): The ARML National Meet Power Round asks a team to work through a connected mathematics packet and justify its solutions. The collection preserves archived Power Round packets as separate sessions from the National Meet Team Round. Agents submit a shared written solution, assessed against available solution material with a proof-oriented rubric (Karafiol et al., 2015; American Regions Mathematics League, 2023).

ARML Nat'l (Team): The ARML National Meet Team Round gives a team a set of short-answer mathematics problems to solve together. CONTESTWORLD collects archived Team Round papers as whole-session tasks. Agents submit one numbered answer sheet, which is checked against the available official answers (Karafiol et al., 2015; American Regions Mathematics League, 2023).

ARML Power: The ARML Power Contest centers on multi-part mathematics problems that call for sustained explanation and proof. The collection contains archived Power Contest packets and their associated solution material where available. Agents produce a shared proof packet, and rubric-based assessment does not reproduce every aspect of the original contest's administration or scoring (American Regions Mathematics League, n.d.).

NCCDC: NCCDC challenges teams to maintain and defend a network while handling operational and business tasks. CONTESTWORLD includes archived team packets and a separate wildcard hardening task. Agents submit written defense, business-response, or hardening plans, which are assessed without a live network or red-team exercise (National Collegiate Cyber Defense Competition, n.d.).

CFA Research Challenge: The CFA Institute Research Challenge asks student teams to analyze a company and prepare an equity-research recommendation. This benchmark uses a historical Canadian Tire issuer packet with an explicit information-date boundary, rather than a complete original 2016 contest packet. Agents prepare an equity-research report, scored with later published written-report criteria whose historical equivalence is not claimed (CFA Institute, n.d.; Canadian Tire Corporation, 2014; 2015).

Cybench: Cybench consists of capture-the-flag cybersecurity challenges drawn from existing contests. CONTESTWORLD evaluates selected challenges individually, with contestant files or services separated from answer material. Agents submit a final flag that is checked against the recorded

answer; the benchmark does not reproduce Cybench's guided subtasks or partial credit (Zhang et al., 2025a).

DebateBench: DebateBench provides motions and transcripts associated with British Parliamentary debates. The benchmark uses selected rounds as prompts for one Opening Government team. Agents submit two labeled written speeches, assessed without live opposition, points of information, oral delivery, or room ranking (Tiwari et al., 2025).

EOES/EUSO: The EOES/EUSO track covers team-based integrated experimental science tasks. CONTESTWORLD represents archived task papers as session-level records and keeps accompanying answer sheets as supporting inputs. Agents submit written analysis based on supplied information, while physical measurements and live laboratory performance are outside the benchmark score (European Olympiad of Experimental Science, n.d.).

Ethics Bowl (APPE): The APPE Intercollegiate Ethics Bowl asks teams to analyze ethical cases and defend their reasoning. CONTESTWORLD includes archived case sets as session-level tasks. Agents produce a written match speech packet, which is scored without live opponent interaction or judge questioning (Association for Practical and Professional Ethics, n.d.).

Ethics Bowl (NHSEB): The National High School Ethics Bowl asks student teams to reason through applied-ethics cases. The collection includes archived national case sets as session-level tasks. Agents produce a written match speech packet, assessed without a live opposing team or judge questioning (National High School Ethics Bowl, n.d.).

Fyziklani: Physics Brawl Online has teams solve physics problems in an online contest. The benchmark collects archived problem papers as session-level tasks, retaining their diagrams and mathematical layout in the original PDFs. Agents submit worked answers, which receive benchmark rubric scores rather than a recreation of the live progressive-release and weighted-scoring system (Physics Brawl Online, n.d.).

GCCH: The Global Case Competition at Harvard asks teams to analyze a business case and recommend a course of action. CONTESTWORLD includes archived case documents from several editions. Agents submit a recommendation deck, which is assessed with case-specific benchmark criteria rather than verified official historical judge weights or a live pitch (Global Case Competition at Harvard, n.d.).

HMMT (Team): The HMMT Team Round is a collaborative mathematics paper whose answer and proof requirements vary by season. CONTESTWORLD includes archived Team Round packets and uses the edition-specific problem text. Agents submit numbered written solutions; verified final answers and proof-based responses are assessed through different benchmark scoring routes (Harvard–MIT Mathematics Tournament, n.d.).

IChTO: The International Chemistry Tournament gives teams open chemistry problems to research and debate. CONTESTWORLD includes archived problem sets as session-level tasks. Agents submit a written chemistry report, scored on scientific content without live debate or the tournament's full presentation and ranking procedure (International Chemistry Tournament, n.d.).

ICM: ICM asks teams to model open-ended problems that connect mathematical methods with other fields. CONTESTWORLD includes archived ICM problem statements as separate session-level tasks. Agents submit a modeling paper and compiled PDF, which receive rubric assessment rather than a reproduction of the original award-panel process (COMAP, n.d.).

ICPC: ICPC World Finals asks three-student university teams to solve algorithmic problems by submitting programs within five hours. The benchmark stores World Finals problems as individual evaluation units with their original task statements. Submitted code is configured for assessment through an authenticated remote programming judge, while the agent environment approximates the original shared-workstation contest setting (Kattis, n.d.; International Collegiate Programming Contest, 2025).

IEO: The IEO Business Case asks teams to analyze a business problem and present a recommendation. CONTESTWORLD includes archived case materials from multiple annual editions. Agents submit a slide deck for assessment; live delivery, jury questions, and speaking balance are not scored (International Economics Olympiad, n.d.).

IIOT: IIOT has school teams solve algorithmic programming problems. CONTESTWORLD stores individual problems from archived International Final task sets, rather than treating each problem as a complete contest session. Agents submit source code, which currently receives benchmark static-review scores instead of an official execution verdict (International Informatics Olympiad in Teams, n.d.).

IJSO (Practical): The IJSO Practical gives teams experimental science tasks involving laboratory work and analysis. CONTESTWORLD represents archived practical papers as session-level tasks and retains marking material separately for assessment. Agents write a report using the supplied task material and any available observations; physical experiments and instrument operation are not assessed (International Junior Science Olympiad, n.d.).

IOAA (Group): The IOAA Group Competition asks an international team to work through an astronomy or astrophysics task. The collection includes archived group or team competition papers from several editions. Agents submit a shared worked answer, assessed against year-specific solutions without reproducing live group-task logistics (International Olympiad on Astronomy and Astrophysics, n.d.).

IOAI (Team): The IOAI Team Challenge asks a team to address an applied artificial-intelligence task. This collection contains the 2025 "Factory of the Future" warehouse-robot challenge guide as one session. Agents submit a Python control program and technical explanation, which are reviewed against task requirements without simulator or physical-robot execution (International Olympiad in Artificial Intelligence, 2025).

IOL (Team): The IOL Team Contest asks a team to solve one linguistic problem using unfamiliar language data. CONTESTWORLD includes archived team problems and the accompanying material needed to interpret them. Agents submit a shared worked answer, which is assessed with available solutions and a benchmark rubric (International Linguistics Olympiad, n.d.).

ITYM: ITYM gives teams open mathematics research problems to develop and defend. The benchmark collects archived annual problem sets as written research tasks. Agents submit a shared solution paper, assessed without the live mathematical fight, opposing team, reviewer exchange, or jury (International Tournament of Young Mathematicians, n.d.).

IYPT: IYPT teams investigate open physics problems and present their findings in physics fights. CONTESTWORLD collects archived annual problem sets as session-level research tasks. Agents submit a written research report, which is evaluated for theory, investigation design, documented evidence, and clarity without a live experiment, opponent, or oral defense (International Young Physicists' Tournament, n.d.).

Jessup: Jessup asks teams to argue a fictional international-law dispute from opposing sides. CONTESTWORLD contains the 2024 case problem as a session-level task. Agents prepare a written memorial, which is evaluated without oral advocacy or judges' questions (International Law Students Association, 2024).

MCM: MCM gives a small team an open-ended mathematical modeling problem and asks for a written solution. The collection includes archived individual contest problems as session-level tasks. Agents submit a modeling paper and compiled PDF, assessed as a report without reproducing the multi-day contest clock or COMAP's award process (COMAP, n.d.).

Mystery Hunt: MIT Mystery Hunt has teams solve interconnected puzzles to advance through a larger hunt. CONTESTWORLD includes selected archived puzzles as separate question-level tasks. Agents submit one answer per puzzle, so the benchmark does not recreate the hunt's full progression or meta-puzzle structure (MIT Mystery Hunt, n.d.).

NYU CTF: NYU CTF Bench contains capture-the-flag challenges associated with CSAW. CONTESTWORLD treats the selected archived challenges as individual question-level tasks and separates contestant inputs from solutions. Teams submit flags checked against recorded answers, rather than receiving a score for a complete CSAW contest (Shao et al., 2024).

Odyssey of the Mind: Odyssey of the Mind asks teams to develop creative solutions that are presented as performances or artifacts. The collection uses public long-term problem synopses from archived seasons. Agents produce written scripts and staging plans for selected synopses, evaluated without a live performance or the unpublished full-problem constraints (Odyssey of the Mind, n.d.).

PUMaC Power: The PUMaC Power Round asks a team to develop proofs for a multi-part mathematics packet. The collection includes archived Power Round papers and published solution material. Agents submit a numbered proof packet, and benchmark rubric grading uses the solutions without reproducing historical hand-grading or the original extended work period (Princeton University Mathematics Competition, n.d.).

Purple Comet: Purple Comet! Math Meet is a timed online team mathematics contest with separate middle-school and high-school papers. CONTESTWORLD includes archived papers from both divisions as complete answer-sheet tasks. Answers are checked against the corresponding keys, while live timing and leaderboard tie-breaks are outside the benchmark evaluation (Purple Comet! Math Meet, n.d.).

QANTA: QANTA contains academic quiz-bowl questions whose clues become progressively more specific. CONTESTWORLD evaluates selected questions individually, using the extracted question text and recorded answers. Agents submit a short answer; buzz timing, interruptions, and full-match scoring are not reproduced (QANTA Project, 2021; Rodriguez et al., 2019).

Science Bowl: DOE Science Bowl asks teams science questions in a buzzer-match format. The benchmark contains extracted questions from archived high-school sample sets as individual evaluation units. Agents submit short answers, which are checked separately without reproducing buzzer timing, consultation rules, or match scores (United States Department of Energy, n.d.).

Science Olympiad: Science Olympiad comprises distinct team events across scientific subjects. This benchmark currently contains one annotated Solar System example exam for Division B, rather than a complete tournament. Agents submit answers to its numbered questions, scored against the printed reference answers (Science Olympiad, n.d.).

Vis Moot: The Vis Moot asks teams to advocate for parties in a fictional international commercial arbitration. CONTESTWORLD collects archived annual case problems as session-level tasks. Agents submit a written memorial, which is evaluated without oral pleading, tribunal questions, or award-panel calibration (Willem C. Vis International Commercial Arbitration Moot, n.d.).

Wharton Investment: The Wharton investment competition asks student teams to design an investment strategy for a client case. CONTESTWORLD includes archived case studies as session-level tasks. Agents submit a written investment policy and strategy report, assessed without live portfolio performance or historical official scoring weights (Wharton Global Youth Program, 2019; 2020; 2021).

WMTC: The World Mathematics Team Championship includes a Team Round in which teammates solve numbered mathematics questions on one paper. CONTESTWORLD uses the archived 2018 Team Round at three competition levels, separate from the event's Individual and Relay rounds. Agents submit a shared answer sheet, checked against the published Team Round answers (World Mathematics Team Championship, 2018c;b;a).

WSC Writing: World Scholar's Cup Collaborative Writing combines team discussion with individually written responses to a shared set of prompts. CONTESTWORLD stores individual guiding questions and groups them into three-author writing tasks in record order; those groupings are not verified original contest packets. Agents write and review essays in stages, which are assessed as written work without claiming to reproduce every official timing or event procedure (World Scholar's Cup, n.d.b;n).

A.3 CONSTRUCTION AND DATA PREPARATION

Collection. We collect problems directly from real competitions. For each contest, we retrieve the official problem material from the organizer's website (Appendix A.2), usually consisting of a problem packet, a problem set, or an official solution document in PDF form; a few contests are available only in other public forms, such as plain-text statements on Kattis, puzzle pages for MIT Mystery Hunt, and existing datasets of prior security benchmarks for Cybench and NYU CTF. We record each item's source URL and raw file so that every processed instance is traceable to an immutable source.

Unifying problem sets. We convert each source document into machine-readable tasks through a uniform pipeline. We extract text from the source PDFs and normalize whitespace, then segment each packet into its numbered problems. An enrichment pass then records, for each part, an expected answer, its point value, the corresponding official solution or rubric, and the evaluation method. To prevent leakage, we strip any appended answer or solution section from the agent-visible statement and keep gold text only in hidden fields; because layouts vary widely, we follow automatic extraction with a curated repair pass to hand-verify answers. The result is a single uniform record schema (problem statement, per-part gold answers, points, deliverable type, and evaluation guideline) shared across all contests.

Converting rules and evaluation. We encode each contest's real-world constraints as simulator-compatible components that specify team size, time limit, permitted materials and tools, and official deliverable. Specifically, we compile these parameters into the agent setting: the time limit becomes a turn and API-call budget through a fixed minutes-per-turn conversion, permitted materials define the available tools (for example, web search, code execution, or asset inspection), team size sets the number of agents, and the deliverable fixes the submission format and evaluation (such as answer sheet, written document, or source code). Because the constraints and the evaluator are stored with the data, each competition is reproducible as an agent task without any per-contest configuration.

A.4 Agent action and tool interface

The contestant interface is defined by a single typed registry (`src/tool_registry.py` and `src/action_modules/`). In the inspected action set (version 17), the registry contains 90 canonical names: 27 common actions, three optional memory actions, and 60 task-specific actions. Of these, 52 are marked as tools because they invoke a computation, resource, or preparation helper; the other 38 operate on contest state, collaboration, memory, or submissions. The same specifications provide argument validation and, when supported by the model provider, function-call schemas. The environment adapter implements 72 names and the contest-session runtime implements all 90. No separate contestant-callable actions belong to the Coach.

This catalog is the union of implemented interfaces, not the menu shown to every agent. At each decision, the runtime filters it by the competition's rule-card permissions (or fallback contest permissions), baseline, assigned role, current phase, visible resources, and live state. For example, only a leader can assign work, memory actions require the memory-enabled condition, and competition-flow actions appear only in their matching stateful contest. A response resolves to one complete, validated action. The short descriptions below state each interface's purpose; the runtime schema and rule card remain authoritative for arguments and availability.

Table 8: Common and optional memory interfaces. "T" marks a tool; "A" marks another action.

Name	Type	Purpose
`select_problem`	A	Claim or switch to a contest problem.
`inspect_problem`	A	Read a problem and its visible draft, review, and submission history without changing the active problem.
`read_problem_pdf`	A	Request approved pages of the original problem PDF when the text or figures are incomplete.
`share_problem_brief`	A	Publish a source-bound problem summary and declare whether its question coverage is complete.
`triage_problem`	A	Set team priority for a problem, including a hopeless designation.
`work`	A	Save a durable candidate answer or draft for a specified or current problem.
`edit_work`	A	Apply bounded, revision-checked edits to a saved artifact source and render the result.
`skip_problem`	A	Release the current problem.
`rest`	A	Pass the current turn.

Name	Type	Purpose
check_budget	A	Inspect remaining turns, calls, tokens, clock, penalties, and uncovered problems.
query_rules	A	Read contest rules and the role-visible rule card.
speak	A	Broadcast a message to the team.
direct_message	A	Send a private message to selected teammate(s).
assign_problem	A	Leader assigns an ordered, enforced work list to a teammate.
request_review	A	Ask for an independent review of the active problem.
review_answer	A	Approve or reject another agent's answer version, optionally pinned by hash.
submit	A	Hand in a final answer for evaluation.
finish_contest	A	End a contest after required submissions and reviews are complete.
propose	A	Open a numbered proposal for team deliberation.
challenge	A	Raise an objection to another agent's open proposal.
provide_evidence	A	Attach a check, derivation, or counterexample to a proposal.
revise	A	Replace the author's current proposal claim.
decide	A	Accept, reject, or defer an open proposal as decision maker.
use_calculator	T	Evaluate a mathematical expression.
execute_code	T	Run code in the contest sandbox; sample checks are local evidence only.
web_search	T	Search web resources permitted by contest rules.
render_pdf	T	Render document or slide source to PDF and page previews.
remember	A	Store a private, optionally problem-tagged note.
recall	A	Retrieve own or shared notes by keyword and optional problem scope.
share_note	A	Publish one of the agent's stored notes to the team.

The three memory actions are enabled only in memory-capable conditions. Their turn cost follows the active turn policy; their API calls and tokens are still accounted for. A private note is neither a team-visible handoff nor a submitted answer until explicitly shared or recorded through the appropriate action.

Table 9: Competition-specific preparation tools. Each returns private feedback about the supplied candidate, not a correctness verdict or submission.

Tool	Purpose
prepare_arml_local_sheet	Assemble numbered ARML Local short answers from the supplied packet.
prepare_arml_team_sheet	Assemble a national ARML Team Round answer sheet.
prepare_wmtc_team_sheet	Preserve WMTC level and numbering in a team answer sheet.
prepare_purple_comet_answers	Check nonnegative integer answers against exact packet numbering.
check_arml_power_proofs	Check numbered ARML Power proofs and part dependencies.
check_arml_national_power_proofs	Check national ARML Power references and circular dependencies.
check_pumac_proof_packet	Check PUMaC proof parts and their earlier-part dependencies.

Tool	Purpose
prepare_hmmt_team_packet	Prepare season-specific HMMT Team answers or proofs.
prepare_guts_set	Assemble a season-specific HMMT Guts set without handing it in.
check_physics_answer_units	Check Physics Brawl values, finite numbers, and units.
check_icpc_io_cases	Organize ICPC stdin/stdout tests and detect contradictory duplicates.
check_iiot_io_cases	Prepare deterministic IIOT input/output cases.
check_linguistic_rule_table	Align proposed rules with examples and counterexamples.
check_ijso_measurements	Check practical measurements, units, uncertainty, and source references.
check_eoes_measurements	Align experimental quantities with units, uncertainty, and provenance.
check_ioaa_observations	Check observations, coordinate frames or units, and source references.
prepare_event_answer_sheet	Prepare a Science Olympiad event sheet for the selected season and division.
prepare_quizbowl_answer	Format a concise quiz-bowl answerline with disambiguation.
prepare_science_bowl_answer	Format a multiple-choice or short Science Bowl answer.
prepare_history_bowl_answer	Prepare one history answer without simulating buzzer or timing rules.
check_bp_speech_order	Check the supplied British Parliamentary speech order and roles.
prepare_appe_case_response	Separate APPE case response, opposing commentary, reply, and judge answers.
prepare_nhseb_case_response	Separate NHSEB presentation, commentary, response, and judge answers.
prepare_chemistry_fight_record	Organize the written chemistry report and documented evidence.
prepare_physics_research_outline	Separate theory, evidence, uncertainty, conclusions, and citations in a deck outline.
check_math_research_dependencies	Check research-paper claims, proofs, and dependency graph.
prepare_mcm_model_record	Link MCM assumptions, variables, model, validation, and sensitivity.
prepare_icm_model_record	Link ICM objectives, model, validation, tradeoffs, and disclosures.
prepare_equity_research_record	Organize thesis, valuation, catalysts, risks, and public sources.
check_portfolio_weights	Check investment weights, client constraints, and duplicate holdings.
prepare_ieo_case_recommendation	Organize IEO alternatives, recommendation, evidence, and implementation.
prepare_ma_case_record	Separate merger target fit, valuation, financing, synergies, and risks.
check_jessup_authority_table	Link each Jessup proposition to authority and packet facts.
check_vis_authority_table	Link each Vis Moot proposition to authority and record facts.
check_puzzle_extraction	Check ordered letter extraction from supplied feeder answers and indices.
prepare_cybench_flag	Pair a recovered Cybench flag candidate with recovery evidence.

Tool	Purpose
prepare_nyu_ctf_flag	Pair a recovered NYU CTF flag candidate with recovery evidence.
prepare_defense_inject_response	Organize a CCDC inject response and verification evidence.
check_robot_mission_sequence	Check a proposed robot mission sequence against a supplied time limit.
check_ioai_artifact_manifest	Check IOAI artifact names, media types, and provenance.
check_creative_constraint_map	Map creative-task constraints to solution elements and evidence.
check_writing_prompt_allocation	Check three distinct WSC prompts, authors, and responses.

Preparation tools operate on contestant-supplied content and the selected contest's output contract. They do not access hidden answers, certify a judge result, execute a robot, or simulate an absent live opponent. A candidate must still be recorded with work and handed in through the applicable submission action.

Table 10: Task-specific submission, resource, and contest-flow interfaces.

Name	Purpose
submit_code	Submit complete source code to the programming evaluator.
read_lab_equipment	Read a declared laboratory equipment resource.
read_star_chart	Read a declared star-chart resource.
start_environment	Start or resume a task's isolated Docker environment.
reset_environment	Discard that task workspace and restart its environment.
inspect_environment	Read container state and network targets without executing commands.
execute_environment_command	Run a shell command inside the task's contestant container.
submit_set	Irreversibly hand in the current HMMT Guts set.
collect_next_set	Reveal the next Guts set after the current hand-in.
buzz	Claim the fully displayed Science Bowl tossup.
answer_tossup	Give the recognized player's first and final tossup answer.
answer_bonus	Give the captain's first and final earned bonus answer.
start_independent_writing	Captain moves WSC writing from planning to private drafting.
write_writing_response	Replace the writer's assigned private response.
read_writing_response	Read one's own draft or assigned peer draft in the permitted phase.
start_peer_review	Captain begins editing after all three drafts exist.
review_writing_response	Replace the response assigned for peer editing.
submit_writing_packet	Captain hands in the three edited responses.

The Guts, Science Bowl, and WSC writing verbs are stateful actions in their respective contest sessions. Their phase and role checks are enforced at call time. The Docker environment tools are likewise restricted to an isolated contestant workspace; inspect_environment does not read files or run commands. In contrast, submit_code is an evaluated submission action, while execute_code is a local execution tool: a successful sample run is not an official accepted verdict.

B AGENT PROMPT

B.1 CONTESTANT SYSTEM PROMPT

OTC Contestant System Prompt

```
You are {agent}, one contestant in a {team_size}-agent team.
You decide which provided function best advances the contest.
Choose exactly one function and call it once.
Do not emit a text answer instead of the function call.

ACTION SCOPE: Advance one assigned question, lemma, or section
per action. Coordinate task division with the team using
available actions; avoid independently regenerating the entire
team deliverable. Keep arguments within the displayed limits
and include the required answer or proof without an unrelated
preamble.

TURN POLICY: This agent turn has one model call and one action,
including PDF, memory and desk tools. The action ends this turn;
its result is delivered at your next turn. There is no separate
private-thinking request or retry request within this turn.

{memory_guidance}

{desk_tool_guidance}

{deadline_rule}

{competition_rule_guidance}

OPEN TABLE COACH PROTOCOL (rule card {rule_id})
Coach gave one problem-blind brief at turn 0 before the clock,
then left permanently. There is no opening summary or Coach
assignment.

Mandatory review: a different teammate must approve the current
answer version before any submission, including final collection
at the end of the contest. A rejection blocks submission; revision
invalidates prior reviews. The controller submits only candidates
whose current versions have valid approval from another teammate.
Candidates without such approval remain unsubmitted.

{additional_rule_card_protocol}

The active task is a shared team cursor, not an ownership lock:
select the problem you intend to act on. Preserve teammates'
drafts and do not overwrite a sound version without a concrete
correction.

Use direct_message for a private question, handoff, or correction
to specific teammates; use speak when the whole team should know.

{task_family_and_review_instructions}

PRE-CONTEST COACH BRIEF
{coach_guidance}

{rule_card_block}

RUNTIME TURN POLICY OVERRIDE: The card's separate private-think
calls and auxiliary-call sequence do not apply to this run.
```

```
Deliberate inside your one action response. Every selected tool
ends the seat turn; there is no additional model generation in
that turn.
```

B.2 CONTESTANT USER PROMPT

OTC Contestant User Prompt

```
CONTEST {session_id}
TASK FAMILY {task_family}
COMPETITION FORMAT {competition_description}

PROBLEM STATEMENT {problem_id}
{problem_statement}

TASK STATUS {task_status}
BUDGET {remaining_budget}

{work_guidance}

{deadline_policy}

{personal_context}

{submission_rule}

{competition_phase}

{programming_gate_guidance}

ACTIVE TASK {problem_id}
({active_task_instruction})

{problem_pdf_context}

{source_requirements_and_current_source}

{review_target_instructions}

SHARED ANSWER REVIEW HISTORY
{answer_version_and_review_history}

YOUR ELIGIBLE PENDING REVIEWS
{eligible_pending_reviews}

{requested_answer_page}

VISIBLE MEMORY
{visible_memory}

{team_handoff_hint}
```

B.3 MEMORY GUIDANCE

OTC Memory Guidance

```
AUTOMATIC CONVERSATION MEMORY: The runtime saves visible
dialogue, your private thinking and candidate claims by task,
then supplies relevant automatic_memory before every decision.
This works without remember or recall. There is no separate
private-thinking request in this mode; source-linked excerpts
from the action and visible dialogue are still saved.

Current-task findings, team_progress and recent_action_errors
are separate. Correct prior failed calls; do not use another
task's result as this task's answer. Old dialogue and agent
summaries are unverified context; the live draft and review
state remain authoritative. The following tools are optional
supplements for curated notes and explicit sharing.

MEMORY DECISION: Before choosing your ordinary action, check
whether useful information should be saved, retrieved, or shared.
These are contest-local scratchpad tools, not external assistance.

MEMORY WRITE: When private reasoning yields a reusable
intermediate result, a failed approach, a corrected assumption,
or an unresolved next step that will matter after a task switch
or handoff, choose remember as this turn's action. An empty note
store is a reason to start a useful note, not to wait for recall.
Tag problem_id and keep a compact finding, evidence or failed
check, uncertainty, and next step. Do not copy every thought or
duplicate a fully visible draft. A saved note is not a verified
answer or submission.

MEMORY READ: Use recall when a prior result, failed attempt,
constraint or handoff is missing from the supplied context and
recall_status reports visible notes that may contain it.
Revisiting a task or repeating a failure is a reason to check,
not an automatic recall. Do not recall when the needed
information is already shown or the note store is empty.
Write focused query keywords and use problem_id to restrict
results to the relevant task. Omit problem_id only when seeking
a method across tasks. Avoid repeating the last query when
notes_changed_since_last_recall is false; an empty result means
rephrase for a different fact or continue solving. recall reads
only saved/shared notes, not all event history; inspect_problem
reads candidate and submission history.

MEMORY SHARE: After remember returns a note_id, use share_note
if that finding, pitfall, or next step is useful to teammates
and not already public. Use an actual id of your own note;
never invent ids. Private notes and private thoughts are not
teammate handoffs until explicitly shared. Memory does not
replace work, independent review, or required discussion.

MEMORY CONTROL: A memory call is your one action and model call
for this agent turn. It ends the turn, and its receipt is
delivered at your next turn. After receiving it, reassess the
next useful action instead of repeating an already completed
save. There is no required memory-call quota: skip memory when
there is nothing useful to preserve, retrieve, or share.
```

B.4 COACH SYSTEM PROMPT

OTC Coach System Prompt

```
You are Coach for a {competition_id} team during the pre-contest
brief. Your only job is to suggest team responsibilities and
competition strategy based on the supplied rules and runtime
constraints.

Return exactly two sections in plain text:

DIVISION OF RESPONSIBILITIES: Name the contestant seats
responsible for each proposed role and the handoffs between
them. Cover parallel work, integration and independent checking
where the rules permit them. Group seats when useful for a
large roster. Respect existing role and submission permissions;
a suggested responsibility grants no new authority.

RULE-BASED STRATEGY: Link each recommendation to a supplied rule
or constraint. Address time allocation, communication, competition
phases, permitted resources and submission readiness as
applicable. Discuss scoring or penalty trade-offs only when the
public rules specify them.

Card advisory topics, within these two sections only:
{advice_scope}

The problems are not available to you. Do not guess their
contents, assign specific problems, propose solution methods,
or write answer content. Leave problem-specific decomposition
to the contestants after they read the task. Do not prescribe
answer length or teach tool syntax, token optimization, memory
APIs, document editing or compiler repair. Do not call actions
or invent rules, participant expertise or permissions.

Keep your own brief concise, at most {brief_char_limit} characters.
```

B.5 COACH USER PROMPT

OTC Coach User Prompt

```
PRE-CONTEST BRIEF
Competition: {competition_id}

Team size: {team_size}; use the named seats and permissions
in the rule card.

Budget: turns={max_turns}, api_calls={max_api_calls};
each contestant seat turn permits at most
{max_api_calls_per_turn} model call(s), a total shared limit
for that seat turn.

Context available for team handoffs: shared context can persist
across team rounds. Use this constraint to plan handoffs, not
to explain memory operations.

This brief happens before the clock starts (turn 0) and costs
the team no contest turn.

You exit after this turn-0 brief and are never called again.
```

```
CONTEST RULE CARD
{public_rule_card}
```

C ALGORITHMS

C.1 OPEN-TABLE COACH EXECUTION

Algorithm 1 Open-Table Coach Execution

Require: Tasks $\mathcal{P}$, public rules $\mathcal{R}$, contestants $\mathcal{A}$, execution budget B
Ensure: Recorded team submissions $\mathcal{Y}$
Initialize contest state S, event history E, and private memories M
$g \leftarrow \textsc{CoachBrief}(\mathcal{R}, \mathcal{A}, B)$ ▷ No task statements
Share g with all contestants
Remove the Coach from subsequent execution
while the contest is active and budget remains **do**
    $a \leftarrow \textsc{NextEligibleContestant}(S)$
    $p \leftarrow \textsc{CurrentTask}(S, a)$
    $C_a \leftarrow \textsc{BuildMemoryContext}(a, p, E, S, M)$
    $\mathcal{U}_a \leftarrow \textsc{PermittedActions}(S, \mathcal{R}, a)$
    $o_a \leftarrow \textsc{Observe}(S, E, a, C_a)$
    $u \leftarrow \textsc{QueryContestant}(a, g, o_a, \mathcal{U}_a)$ ▷ One action
    if u satisfies the current action constraints **then**
        $(S, z) \leftarrow \textsc{ExecuteAction}(S, u)$
    else
        $z \leftarrow$ validation-error receipt
    end if
    Append the action and its outcome z to E
    Update execution budget and contest status
end while
$\mathcal{Y} \leftarrow \textsc{FinalCollection}(S, \mathcal{R})$
return $\mathcal{Y}$

C.2 AUTOMATIC MEMORY UPDATE AND CONTEXT CONSTRUCTION

Algorithm 2 Automatic Memory Update and Context Construction

Require: Contestant a, current task p, event history E, visible task state S, private memories M
Require: Memory-context character budget L
Ensure: Bounded memory context C_a
procedure $\textsc{BuildMemoryContext}(a, p, E, S, M)$
    $E_a \leftarrow \textsc{VisibleSourceEvents}(E, S, a)$ ▷ Exclude judge-only information
    $U \leftarrow \textsc{UnprocessedEvents}(E_a, M[a])$
    for all tasks q represented in U **do**
        $m \leftarrow M[a, q]$
        for all events $e \in U$ associated with q **do**
            Append a bounded excerpt of e to m
            Retain its author and source reference
            Associate the excerpt with the current draft and review state
            if e proposes an answer **then**
                Check whether the proposal became the current draft
                Treat unsupported proposals as unverified claims
            else if e reports an action error **then**
                Retain its failure receipt
            end if
        end for
        Incorporate a valid task-specific summary, if supplied
        Separate task findings from team observations
        Bound retained history and advance the event cursor
        Save m privately as a new memory record

end for
$F \leftarrow$ LATESTVISIBLERECORDS(M, a, S)
for all records $m \in F$ **do**
  Exclude findings whose associated draft or review state has changed
  Attach the current candidate and review status
end for
$C_a \leftarrow$ ASSEMBLECONTEXT(F, p) ▷ Current task, errors, and recent team context
while CHARACTERCOUNT$(C_a) > L$ **do**
  Remove lower-priority context in the prescribed truncation order
  Use compact inspection references if needed
end while
return C_a
end procedure

C.3 INDEPENDENT REVIEW AND CANDIDATE REVISION

Algorithm 3 Independent Review and Candidate Revision

Require: Current candidates D, review histories R
procedure UPDATECANDIDATE(p, c, a)
  Require that problem p is visible and editable
  if candidate content and metadata are unchanged **then**
    return $D[p]$
  end if
  $D[p] \leftarrow$ candidate c authored by a
  for all $r \in R[p]$ **do**
    Mark r as inactive for the current candidate
  end for
  Make $D[p]$ available for independent review
  return $D[p]$
end procedure

procedure RECORDREVIEW(p, a_r, d, b)
  Require an existing, reviewable candidate $D[p]$
  Require $a_r \neq D[p].\text{author}$
  Require $d \in \{\text{approve}, \text{reject}\}$
  Append reviewer a_r, decision d, and explanation b to $R[p]$
  Associate the review with the current candidate
end procedure

function INDEPENDENTLYAPPROVED(p)
  if no candidate exists for p **then**
    return false
  end if
  $R_p \leftarrow$ active reviews of the current candidate
  for all $r \in R_p$ **do**
    if $r.\text{decision} = \text{reject}$ **then**
      return false
    end if
  end for
  for all $r \in R_p$ **do**
    if $r.\text{decision} = \text{approve}$ **and** $r.\text{reviewer} \neq D[p].\text{author}$ **then**
      return true
    end if
  end for
  return false
end function

D REPRESENTATIVE CONTEST EXAMPLES

We provide one figure-style profile for each of the 39 active competition tracks. Each profile presents a contestant-facing task description, catalog coverage, evaluation status and scope, judge-side scoring

reference, tools, team rules, and expected submission. Every profile has an outer figure frame and separately colored panels for its components. Record IDs identify the source records in the dataset. All profiles use the same fixed-size layout, one per page; long fields are truncated with an ellipsis, and the full text is available in the released dataset.

ARML Local (2013)
Example: arml_local_2013
Record ID: `arml_local_2013`

Task prompt

Team Problems 1. Compute the number of digits in the base-10 representation of 2^3 * 3^4 * 4^5 * 5^6. 2. Compute the length of the interval of the set of numbers x that satisfy floor(x^2) = floor((x + 3)^2), where floor(x) denotes the greatest integer less than or equal to x. 3. The sequence x_1, x_2, x_3, ... is formed by setting x_1 = 1, x_2 = -3, and for n > 2, x_n is obtained by cubing one of the previous terms, chosen uniformly at random. Compute the probability (as a fraction) that x_1*x_2*x_3*x_4*x_5 > 1000. 4. Compute the remainder when 8^8 - 4^4 is divided by 4^4 - 2^2. 5. A, B, and C are chosen at random (with replacement) from the set of digits {0, 1, 2 . . .

Track coverage

6 records in this track; source years 2009, 2010, 2011, 2012, 2013, 2014; record unit(s): session.

Evaluation and scope

ready via `gold_answer_v1`; selected record unit: session. *Scope:* No evaluator limitation is stated in this record entry.

Judge-side reference

Team Solutions 1. 2 3344556 = 106 27 34 = 106 128 81 106 128 80 = 107 1024 = 1.024 1010 which has 11 digits. 2. Look at the graph of y = x2 and find plateaus of the graphs that are . . .

Expected submission

Numbered team answer sheet matching the provided packet: 1. [answer] 2. [answer] ... N. [answer] where N is the number of Team problems on the sheet (usually 10 in this corpus; 15...

Track-specific tools

`query_rules`, `prepare_arml_local_sheet`

Team and resource rules

6-person team. The entire six-person team works the Team Round together and files one shared short-answer sheet. Official duration: 45 minutes. Prohibited resources include internet . . .

ARML National Meet: Power Round (2009)
Example: arml_national_power_2009
Record ID: `arml_national_power_2009`

Task prompt

Power Question 2009: Sign on the Label Ann-label is a permutation of the numbers 1 through n. For example, J = 35214 is a 5-label and K = 132 is a 3-label. For a fixed positive integer p, where pn, consider consecutive blocks of p numbers in an n-label. For example, when p = 3 and L = 263415, the blocks are 263, 634, 341, and 415. We can associate to each of these blocks a p-label that corresponds to the relative order of the numbers in that block. For L = 263415, we get the following: 263415 132; 2634 15 312; 26341 5 231; 263415 213. Moving from left to right in the n-label, there are np + 1 such blocks, which means we obtain an (np + 1)-tuple of p-labels. For L = 263415, we get the 4-tuple (132 , 312, 231, 213). We will call this (np + 1)-tuple thep-signature of L (or signature, if p is clear from the context) and denote it by Sp[L]; the p-labels in the signature are called windows. For L = 263415, the windows are 132, 312, 231, and 213, and we write S3[263415] = (132, 312, 231 . . .

Track coverage

11 records in this track; source years 2009, 2010, 2011, 2012, 2013, 2014, 2016, 2017, 2018, 2019, 2023; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* No evaluator limitation is stated in this record entry.

Judge-side reference

Solutions to 2009 Power Question 1. (a) (312 , 123, 231) (b) There are three: 41352 , 42351, 51342. (c) There are four: 352146 , 362145, 452136, 562134. 2. (a) We will prove this by . . .

Expected submission

Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools

`query_rules`, `check_arml_national_power_proofs`

Team and resource rules

15-person team. The entire team produces exactly one Power solution packet. Prohibited resources include internet, calculator, code execution, electronic devices.

ARML National Meet: Team Round (2011)
Example: arml_national_team_2011
Record ID: `arml_national_team_2011`

Task prompt

2011 Team Problems T-1. If 1 ,x,y is a geometric sequence and x,y, 3 is an arithmetic sequence, compute the maximum value of x +y. T-2. Define the sequence of positive integers {an} as follows: { a1 = 1; for n 2, an is the smallest possible positive value of nak2, for 1k <n. For example,a2 = 2 12 = 1, and a3 = 3 12 = 2. Compute a1 +a2 + +a50. T-3. Compute the base b for which 253b 341b = 7 4XY Z b, for some base-b digits X,Y,Z . T-4. Some portions of the line y = 4x lie below the curve y = 10 sin2x, and other portions lie above the curve. Compute the sum of the lengths of all the segments of the graph of y = 4x that lie in the first quadrant, below the graph of y = 10 sin2x. T-5. In equilateral hexagon ABCDEF , mA = 2mC = 2mE = 5mD = 10mB = 10mF , and diagonal BE = 3. Compute [ABCDEF], that is, the area of ABCDEF . T-6. The taxicab distance between points A = (xA,yA) and B = (xB,yB) is defined as d(A,B) = |xA xB| +|yA yB|. Given some s > 0 and . . .

Track coverage

11 records in this track; source years 2009, 2010, 2011, 2012, 2013, 2014, 2016, 2017, 2018, 2019, 2023; record unit(s): session.

Evaluation and scope

ready via `gold_answer_v1`; selected record unit: session. *Scope:* No evaluator limitation is stated in this record entry.

Judge-side reference

T-1. 15/4 T-2. 253 T-3. 20 T-4. (5/4)17 T-5. 9/2 T-6. 96 T-7. 4096 T-8. 70 T-9. 11 T-10. 5/4

Expected submission

Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools

`query_rules`, `prepare_arml_team_sheet`

Team and resource rules

15-person team. The entire team works the Team Round together and files one shared short-answer sheet. Prohibited resources include internet, calculator, code execution, electronic devices.

ARML Power Contest (2023)
Example: arml_power_fall_2023
Record ID: `arml_power_fall_2023`

Task prompt
The Heat Is On! Fall 2023 ARML Power Contest The Background Often, in races such as at track or swim meets, there are more competitors than can compete at one time. For instance, there may be 20 racers but the track only has 8 lanes. The race is accomplished by having several smaller races called heats. Several racers will race in the first heat, then several more in the second heat, and so on until all the racers have raced at least once. These are called the preliminary heats. Then, the fastest racers will race together in the final championship heat. The fastest racers that race in the championship will usually be chosen from the prelim- inary heats by using a stopwatch, and the fastest racers will be chosen regardless of how they did in their particular heat. For example, if the championship heat will have six racers, those six could all come from the same preliminary heat if those six had the fastest times among all racers in all preliminary heats. It is also possible that racer . . .

Track coverage
14 records in this track; source years 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* LLM rubric grading is anchored to the official solution packet; diagram and proof quality . . .

Judge-side reference
The Heat Is On! Fall 2023 ARML Power Contest The Solutions 1. H(n, p, n) is the number of heats needed to determine the fastest p racers out of a total of n racers, when all n racers are . . .

Expected submission
Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools
`query_rules`, `check_arml_power_proofs`

Team and resource rules
15-person team. Contestants collaborate only with teammates on one shared proof packet. This card does not claim a complete current Power Contest administrivia package. Prohibited resources . . .

National Collegiate Cyber Defense Competition (NCCDC) (2025)
Example: NCCDC 2025: Oscorp
Record ID: `ccdc_2025-NCCDC-TeamPacket`

Task prompt
2025 NCCDC Team Packet April 25-27, 2025 2025 Center for Infrastructure Assurance and Security cias.utsa.edu 3 Table of Contents Competition Schedule 4 Sponsors 5 Competition Rules 6 Scoring 8 Password Changes 12 Competition Network Information 13 Team Network Diagram 14 Letter from Oscorp Industries 15 Network Information 16 2025 Center for Infrastructure Assurance and Security cias.utsa.edu 4 Competition Schedule Please note all times are in Central Time. Friday, Apr 25th 10:30 AM 11:25 AM...

Track coverage
12 records in this track; source years 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2025, 2026; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Written planning adaptation of the archived scenario, not an operational cyber-defense . . .

Judge-side reference
data/rubrics/ccdc_open_artifact_llm_20260917_v1.json

Expected submission
BENCHMARK SUBMISSION: Using this team packet, submit a written defense and business-response plan for the described network. Include service priorities, hardening, monitoring, recovery . . .

Track-specific tools
`query_rules, inspect_environment, prepare_defense_inject_response`

Team and resource rules
8-person team. Collaborate only within the team under contest rules. Prohibited resources include internet, calculator.

CFA Institute Research Challenge (2016)
Example: Canadian Tire Corporation, Limited (CTC.A.TO)
Record ID: `cfa_research_challenge_2016`

Task prompt
Prepare an equity research report on Canadian Tire Corporation using the issuer input packet. The benchmark information boundary is 2015-12-31, a dataset choice, not an asserted official 2016 deadline. Use only supplied historical information; state valuation assumptions, missing peer/market inputs and a conditional buy/hold/sell price threshold when needed. Do not import later outcomes or invent historical market prices. Cover business, industry, investment summary, valuation, financial analysis, risks and ESG. The 2025-2026 written-report scorecard is reused explicitly for this historical-source benchmark, not claimed as the 2016 official scorecard. Packet pages 1-12: 2014 annual report original PDF pages 7-13 and 65-69. Pages 13-70: complete 2015 Q3 report. Packet page 3: MANAGEMENTS DISCUSSION AND ANALYSIS 1.7 Key operating performance measures and additional GAAP and non-GAAP financial measures The Company has identified several key . . .

Track coverage
1 record in this track; source years 2016; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Historical issuer inputs with benchmark information boundary; not the original complete . . .

Judge-side reference
data/rubrics/cfa_research_challenge_mini_written_20260917_v1.json

Expected submission
Prepare an equity research report on Canadian Tire Corporation using the issuer input packet. The benchmark information boundary is 2015-12-31, a dataset choice, not an asserted official . . .

Track-specific tools
`query_rules, use_calculator, execute_code, web_search, prepare_equity_research_record`

Team and resource rules
4-person team. Collaborate only within the team under contest rules.

Cybench (2022)

Example: hkcert-ctf/CTF-Challenges/CTF-2022/05-motp

Record ID: `cybench_hkcert-ctf_CTF-Challenges_CTF-2022_05-motp`

Task prompt

You are trying to capture the flag on otp:80. Initial files: - docker-compose.yml - otp/Dockerfile - otp/src/bootstrap.min.css - otp/src/google2fa.php - otp/src/index.php - otp/src/jsonhandler.php - otp/src/login.php - otp/src/main.css

Track coverage

38 records in this track; source years 2022, 2023, 2024; record unit(s): question.

Evaluation and scope

via `gold_answer_v1`; selected record unit: question. *Scope:* Final-flag benchmark only; guided subtasks and their partial credit are not reproduced. . . .

Judge-side reference

hkcert22{mistakes-off-the-page}

Expected submission

Submit the recovered flag(s) / subtask answers clearly labeled.

Track-specific tools

`query_rules`, `inspect_environment`, `start_environment`, `execute_environment_command`, `reset_environment`, `prepare_cybench_flag`

Team and resource rules

5-person team. Collaborate only within the team under contest rules. Prohibited resources include calculator.

DebateBench (WUDC / BP) (2022)

Example: Belgrade WUDC 2022 Round 6

Record ID: `debatebench_Belgrade WUDC 2022 Round 6`

Task prompt

Motion In times of economic crisis, This House Would implement fiscal central banking Information Slide For the purposes of this debate, "fiscal central banking" refers to an economic system that grants central banks the mandate to control circulation of money in the economy through fiscal policy in addition to monetary policy. Under this system central banks can define limitations for the size of the budget, but not its distribution. i.e. they can decide rates for various taxes and limit or increase overall government spending, but can't decide what the money is spent on BENCHMARK SUBMISSION: You are one Opening Government team. Submit two labeled written speeches, PM and DPM, supporting the motion. PM defines and establishes the case; DPM develops and strengthens it. You may anticipate objections explicitly as hypothetical. No opposing speeches or live POIs are supplied; do not invent that they occurred. Only your two speeches are scored, separately out of 100, total 200. This is written preparation, not a four-team ranking or an oral-performance score.

Track coverage

24 records in this track; source years 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope

via `rubric_llm_v1`; selected record unit: session. *Scope:* Text-only preparation; no actual opposition, POIs, oral delivery or room ranking.; Two . . .

Judge-side reference

data/rubrics/debatebench_open_artifact_llm_20260917_v1.json

Expected submission

BENCHMARK SUBMISSION: You are one Opening Government team. Submit two labeled written speeches, PM and DPM, supporting the motion. PM defines and establishes the case; DPM develops and . . .

Track-specific tools

`query_rules, check_bp_speech_order`

Team and resource rules

2-person team. One Opening Government pair prepares PM and DPM written speeches. Prohibited resources include internet, calculator, code execution.

EOES / EUSO (2021)
Example: t1
Record ID: `eoes_t1`

Task prompt
T1 10.05.2021 Intro - 1 Task 1 Acacia Honey photo: Peter Vanko 10. 05. 2021. Szeged Hungary T1 10.05.2021 Intro - 2 Acacia Honey Robinia pseudoacacia is endemic to a few small areas in Northern America but it has been planted and naturalized all over the wor ld and today in some areas it has bec ome an invasive species. In Hungary, it was planted first at the beginning of the 18 th century as a park tree but half a century later Samuel Tessedik discovered that it can live on the sandy, saline soil and can be widely used for reforesting the Great Hungarian Plain. In the 19 th century, it was already considered a Hungarian tree, now its area reaches 4600 km2. In addition, acacia is also very popular with beekeepers. Acacia honey is popular because of its light colour, mild flavour, and acacia blossom scent. There are also many beekeepers in the Szeged area, and the acacia honey used in the measurement comes from here. One of the main advantages of acacia honey is that it flows smoothly . . .

Track coverage
12 records in this track; source years 2013, 2015, 2017, 2021, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Written scientific analysis only; physical measurements and live laboratory performance . . .

Judge-side reference
data/rubrics/eoes_mini_written_20260917_v1.json

Expected submission
Written experimental-analysis benchmark: address the paired task and blank answer sheet by section and question. Solve questions using supplied constants, tables and figures. For questions . . .

Track-specific tools
`query_rules`, `use_calculator`, `inspect_environment`, `check_eoes_measurements`

Team and resource rules
3-person team. Collaborate only within the team under contest rules. Prohibited resources include internet, code execution.

APPE Intercollegiate Ethics Bowl (2022)

Example: 2022-national-case-set

Record ID: `ethics_bowl_appe_2022-national-case-set`

Task prompt

Cases for the 26th APPE Intercollegiate Ethics Bowl National Championship Held in conjunction with the 31st Annual Meeting of the Association for Practical and Professional Ethics Competition to be held online Saturday, February 26 and Sunday, February 27, 2022 Robert Boyd Skipper: Chair, Case Preparation Committee Robert A. Currie Cynthia Jones Heather Pease Prepared by: This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/. 2021 Robert Boyd Skipper, Robert A. Currie, Cynthia Jones, and Heather Pease. Version 1, December, 27, 2021 1 Case 1: Diagnosing Public Figures In January 2020, a professor of psychiatry at Yale University, Dr. Bandy X. Lee, had a few things to say about Alan M. Dershowitzs claims about his sex life. After being questioned in 2019 about his . . .

Track coverage

9 records in this track; source years 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Scores a produced match transcript/speech packet; live opponent interaction . . .

Judge-side reference

Oral case presentation; APPE judging guides

Expected submission

Provide the case response, opposing-team commentary, commentary reply, and judge-question answers as distinct labeled sections exposed by the task.

Track-specific tools

`query_rules`, `prepare_appe_case_response`

Team and resource rules

5-person team. The seated participants may confer only in the designated phase; multiple members may contribute orally, but only one person speaks at a time. Prohibited resources include . . .

National High School Ethics Bowl (2021)

Example: 2021-22-national-case-set

Record ID: `ethics_bowl_nhseb_2021-22-national-case-set`

Task prompt

NATIONAL HIGH SCHOOL ETHICS BOWL 2021-2022 Last Updated: March 1, 2022 National Case Set ABOUT THIS DOCUMENT The cases in this set were compiled for the 2022 National High School Ethics Bowl, and edited for content, clarity, and pedagogical focus by NHSEBs Executive Committee, in editorial collaboration with NHSEBs newly-formed Case Advisory Committee. Unless otherwise noted, these cases are the intellectual property of the Parr Center for Ethics. Please do not edit or reproduce them, in part or whole, without attribution. For more information on citation and attribution, please see licensing information below, or contact NHSEB HQ at ethicsbowl@unc.edu. Contents CASES FOR NATIONAL COMPETITION 01 The Price of Knowledge 02 Banking on Cash 03 Yuccatta Be Kidding Me 04 Married to the Name? 05 Privatized Parenthood 06 Wilding Out 07 CRISPR Conundrum 08 Think Outside the Box 09 Home for the Holidays 10 Death and Taxes 11 Tips for Good Living 12 Unwarranted Interventions? 13 All Eyes on Y ou . . .

Track coverage

14 records in this track; source years 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Scores a produced match transcript/speech packet; live opponent interaction . . .

Judge-side reference

Oral case presentation

Expected submission

Provide the presentation, opposing-team commentary, response to commentary, and answers to judges as distinct labeled sections.

Track-specific tools

`query_rules`, `prepare_nhseb_case_response`

Team and resource rules

5-person team. Only the seated members participate in a match; they may confer in the designated periods, while the non-presenting team remains silent during the other team's speaking . . .

Physics Brawl Online (Fyziklání) (2020)
Example: fyziklani_2020
Record ID: `fyziklani_2020`

Task prompt

Official contestant problem text (PDF pages retain diagrams and typesetting): [PDF page 1] Problem FoL.1 . . . (hopefully just) easy pulleys 3 points Suppose that we have a pulley and some weights hanging from a cord as shown in the figure. Assume that both the cord and the pulley are ideal and massless, M = 2.0 kg and m = 1.0 kg. Find the acceleration of the weight M (in the downwards direction). Lego wants to discover the simplest pulley problem that half of all teams still fail. [PDF page 2] Problem FoL.2 . . . two circuits 3 points Legolas found several old identical resistors and connected them in series. Such a circuit had a total resistance RS = 10.0 k. After that, he decided to connect the resistors in parallel and measured a resistance RP = 1.0 . What was the resistance of a single resistor? Lego made himself a lego out of resistors. [PDF page 3] Problem FoL.3 . . . felling trees 3 points A coniferous tree could be (from a mechanical point of view) approximated by a . . .

Track coverage

14 records in this track; source years 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* LLM proxy scoring does not require recovered numeric gold. Complete contestant task . . .

Judge-side reference

Physics Brawl online judge / published solutions.

Expected submission

Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools

`query_rules, use_calculator, check_physics_answer_units`

Team and resource rules

5-person team. Team members may communicate with one another and with organizers about wording or the contest environment; teachers and other teams are forbidden. Official duration: 180 . . .

Global Case Competition at Harvard (2021)
Example: case-2021
Record ID: `gcch_harvard_case-2021`

Task prompt
HARVARD The competition is brought to you by the IFSA Network and the Harvard Graduate School of Arts and Sciences Business Club. 2 02 1 CAS E GLOBAL CASE COMPETITION AT NVIDIA - AMD Dear Participant, We would like to thank you for taking part in the Global Case Competition at Harvard 2021. When the competition was created in 2015, its aim was to convene the world's talent and give them international exposure. Through the editions, participants were able to meet likeminded individuals and benefit from one another, whilst exhibiting their skills to industry professionals. This is your opportunity to prove your worth to a panel of carefully selected judges, acquire unique industry insight, and share your thoughts and ideas with a broad and impactful audience. Given the global sanitary situation, we are saddened that we cannot offer you all that the competition has to offer by inviting you on campus. However, we sincerely hope that our paths cross in the future editions and we are . . .

Track coverage
7 records in this track; source years 2019, 2020, 2021, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Five recommended solution topics in the original 2021 case operationalized as benchmark . . .

Judge-side reference
data/rubrics/gcch_harvard_open_artifact_llm_20260917_v1.json

Expected submission
Submit a structured report or slide outline covering analysis, recommendation, and evidence.

Track-specific tools
`query_rules, prepare_ma_case_record`

Team and resource rules
4-person team. Collaborate only within the team under contest rules.

HMMT: Team Round (2014)
Example: hmmt_team_2014
Record ID: `hmmt_team_2014`

Task prompt
BENCHMARK SUBMISSION OVERRIDE: Submit numbered plain-text answers with proofs and reasoning. No paper, PDF or paid official grading is required. The controller grades with an LLM using available official proof references. HMMT February 2014 Team Round: 10 proof/justification problems; 250 printed raw points. Submit a numbered proof packet, with a full justification for each attempted problem even when the requested result is numerical. Use the attached original problem PDF as authoritative for formulas, figures and contest instructions; the text below is only a preview. [Archived problem text] The attached original PDF is authoritative for mathematical layout, superscripts, tables and diagrams. The following text is a searchable preview, not a lossless replacement. TEAM This test consists of ten problems to be solved by a team in one hour. The problems are unequally weighted with point values as shown in brackets. They are not necessarily in order of difficulty, though ...

Track coverage
14 records in this track; source years 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Gold parts assess the final answer only, as a user-selected benchmark adaptation ...

Judge-side reference
Use the edition-specific proof rubric and original official solution PDF; no short-answer matching.

Expected submission
Benchmark scoring override: submit numbered final answers for verified short-answer parts; include complete proofs/reasoning for remaining parts. Exact gold is used where a ...

Track-specific tools
`query_rules`, `prepare_hmmt_team_packet`

Team and resource rules
8-person team. All members of one team work together on the Team Round. Official duration: 60 minutes. Prohibited resources include internet, calculator, code execution, outside communication.

International Chemistry Tournament (2019)
Example: IChTo-2019-Problem-Set
Record ID: `ichto_IChTo-2019-Problem-Set`

Task prompt

International Chemistry Tournament 2019 Problem Set Problem 1. Chemical Vacuum Pump Suggest a system comprised of a sealed vessel with no moving mechanical parts and a mixture of compounds that is capable of creating vacuum under appropriate conditions by chemical reaction. The vessel should be loaded at STP (273 K, 1 bar) and the final pressure should be no more than 1 Pa. Explain how the vessel is loaded and reaction is initiated and estimate the final pressure in the system. Problem 2. Reduce, Reoxidise, Recycle Imagine you want to set up the museum of chemistry where different chemical reactions are demonstrated to the public. Suggest a diverse set of reactions for the display bearing in mind that you cannot purchase any reagent except water once you have set up the exhibition. Hint: use of electrolysis is allowed Problem 3. The Periodic Law Illustration One of the most salient features of the periodic table is, well, periodicity at which element p r o p e r t i e s c h a n g e . . . .

Track coverage

9 records in this track; source years 2017, 2018, 2019, 2020, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope

via `rubric_llm_v1`; selected record unit: session. *Scope:* Reuse the 2025 scientific criteria across editions as a 9-point written-report subtotal. . . .

Judge-side reference

Reuse the 2025 scientific criteria across editions as a 9-point written-report subtotal. No live debate, external opponent/reviewer, official TP/RP conversion, or team ranking is reproduced.

Expected submission

One shared written chemistry research report.

Track-specific tools

`query_rules, use_calculator, prepare_chemistry_fight_record`

Team and resource rules

4-person team. Jointly prepare and internally review one shared artifact. Prohibited resources include internet, code execution.

Interdisciplinary Contest in Modeling (ICM) (2015)
Example: icm_2015_D
Record ID: `icm_2015_D`

Task prompt
BENCHMARK SUBMISSION OVERRIDE: Write a complete mathematical modeling paper in self-contained LaTeX and submit its compiled PDF. Include abstract, assumptions, model formulation, derivations, results, validation or sensitivity analysis, limitations, conclusion and references. The benchmark uses an LLM judge; paid official evaluation and membership-only winning papers are not required. 2015 ICM Problem D Is it sustainable? Problem background One of the largest challenges of our time is how to manage increasing population and consumption with the earths finite resources. How can we do this while at the same time increasing equity and eradicating poverty? Since the beginning of the modern environmental movement in the 1960s, balancing human needs with the earths health has been a topic of considerable debate. Are economic development and ecosystem health at odds? In order to reconcile this difficult balance, the concept of sustainable development was introduced in the 1980s. Sustainable . . .

Track coverage
19 records in this track; source years 2015, 2016, 2017, 2021, 2022, 2024, 2025; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Rubric grading evaluates the submitted report; COMAP originality checks and award-level . . .

Judge-side reference
COMAP MCM/ICM report rubric: modeling, analysis, communication; 25 page PDF report.

Expected submission
BENCHMARK SUBMISSION OVERRIDE: Write a complete mathematical modeling paper in self-contained LaTeX and submit its compiled PDF. Include abstract, assumptions, model formulation . . .

Track-specific tools
`query_rules, use_calculator, execute_code, work, render_pdf, prepare_icm_model_record`

Team and resource rules
3-person team. Only registered team members may work on the solution. Official duration: 5940 minutes. Prohibited resources include outside communication.

ICPC World Finals (2018)
Example: Catch the Plane
Record ID: `icpc_wf_2018_catch`

Task prompt
Hide Catch the Plane Your plane to the ICPC Finals departs in a short time, and the only way to get to the airport is by bus. Unfortunately, some of the bus drivers are considering going on strike, so you do not know whether you can get to the airport on time. Your goal is to plan your journey in such a way as to maximize the probability of catching your plane. You have a detailed map of the city, which includes all the bus stations. You are at station 0 and the airport is at station 1 . You also have a complete schedule of when each bus leaves its start station and arrives at its destination station. Additionally, for each bus you know the probability that it is actually going to run as scheduled, as opposed to its driver going on strike and taking the bus out of service. Assume all these events are independent. That is, the probability of a given bus running as planned does not change if you know whether any of the other buses run as planned. If you arrive before the departure time . . .

Track coverage
158 records in this track; source years 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): problem.

Evaluation and scope
remote_judge_ready via `programming_judge`; selected record unit: problem. *Scope:* No evaluator limitation is stated in this record entry.

Judge-side reference
ICPC scoring: ranked by number of problems solved (descending), then by total time penalty (ascending). Each accepted problem earns 1 point; wrong submissions add 20-minute penalty. . . .

Expected submission
Submit a complete stdin/stdout program in an allowed World Finals language (C, C++, Java, Kotlin, or Python 3) through the programming judge. Only an Accepted run solves a problem.

Track-specific tools
`query_rules, execute_code, check_icpc_io_cases`

Team and resource rules
3-person team. Three contestants may collaborate with one another throughout the contest while sharing one workstation. Official duration: 300 minutes. Prohibited resources include . . .

International Economics Olympiad (2024)
Example: ieo_business_case_2024
Record ID: `ieo_business_case_2024`

Task prompt
Housing challenges in Hong Kong introducing solutions to be used by Hong Kong Government IEO 2024 Business Case Task Supporting organization: Hong Kong has been facing a severe housing crisis for years, characterized by high property prices, long waiting times for public housing significant shortage of affordable homes. International Economics Olympiad 2024 Page 2 Background The HKSAR Government has implemented various policies and initiatives to tackle the housing issue, but the problem persists, affecting the quality of life and social stability in the city. The government recognizes the urgent need to develop effective and sustainable solutions to address the housing crisis and ensure that all residents have access to adequate and affordable housing. You are a team of consultants appointed by the HKSAR Government to analyze the housing crisis in Hong Kong and provide recommendations for addressing the issue. Your task is to: International Economics Olympiad 2024 Page 3 Task Analyze . . .

Track coverage
5 records in this track; source years 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope
ready via `slide_deck_v1`; selected record unit: session. *Scope:* The PDF deck is scored; live delivery, jury questions, and team-speaking balance are not reproduced.

Judge-side reference
Evaluation Criteria Clarifying Questions Analytical Problem(s) and goal(s) identification Has the team identified the problems and goals of their client? Information Synthesis Has the team . . .

Expected submission
Submit a structured report or slide outline covering analysis, recommendation, and evidence.

Track-specific tools
`query_rules`, `use_calculator`, `execute_code`, `web_search`, `prepare_ieo_case_recommendation`

Team and resource rules
5-person team. Collaborate only within the team under contest rules.

IIOT: International Informatics Olympiad in Teams (2021)
Example: Chimney Service (chimneys)
Record ID: `iiot_2021_01`

Task prompt
[IIOT logo] International Informatics Olympiads in Teams [book indicator: 4 orange, 1 grey] Online, May 8th 2021 chimneys EN Chimney Service (chimneys) Giorgio is always on the lookout for new business opportu- nities. Today, he acquired a bankrupt company that used to deal with industrial construction material. As soon as he made the purchase, he went into the warehouse and noticed that there were N perfectly preserved chimney parts. A chimney part is, essentially, a pipe (or tube) made of con- crete that is characterized by two diameters,A and B, one for each open side of the pipe, and a lengthL. [Pipe diagram: smaller end diameter A, larger end diameter B, axial length L.] By combining these chimney parts, Giorgio can build and sell chimneys to nearby industries. As you might know, in order to build a functional chimney, the structure should be getting progressively narrower towards the top. This is essential in order for the steam or smoke to naturally flow upward, instead of . . .

Track coverage
31 records in this track; source years 2017, 2021, 2022, 2023; record unit(s): problem.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: problem. *Scope:* LLM static source-review score with published subtask weights and benchmark . . .

Judge-side reference
data/rubrics/iiot_2021_01_source_review_v1.json

Expected submission
Submit source code or the required program output for the programming judge.

Track-specific tools
`query_rules, execute_code, check_iiot_io_cases`

Team and resource rules
4-person team. Four contestants collaborate throughout the contest while sharing two organizer-provided computers; the team leader coordinates but is not a contestant. Prohibited resources . . .

IJSO Practical (2019)

Example: ijso_practical_2019_team_practical_2019

Record ID: `ijso_practical_2019_team_practical_2019`

Task prompt

Practical Exam December 9th, 2019 Please DO NOT turn to next page until the start signal is given. Points: 40 Time: 4 Hours 16th IJSO 2019, Doha, Qatar Practical Exam, Dec. 9th, 2019 Page 2 of 24 IMPORTANT REMARKS AND SAFETY RULES 1. Always follow the instructions. 2. All competitors are expected to work safely, behave responsibly and keep the work environment clean. When carrying out discussions with your teammates, please keep your voice low so as not to disturb others. 3. Safety goggles and lab coats must be worn all the time. You may need to use the face-mask in Tasks number 1 and 3. 4. In case of chemical spills or broken glass, please raise your hand and seek assistance from Lab Assistant. 5. Eating and drinking in the lab are not allowed. If necessary, and for only medical reasons, you may ask a Lab Assistant for permission to take a snack break adjacent to the examination hall. 6. Do not leave the examination hall until you have permission to do so. Ask the Lab Assistant if . . .

Track coverage

8 records in this track; source years 2014, 2016, 2017, 2018, 2019, 2021, 2022, 2023; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Uses the declared benchmark rubric, not a full official competition score.

Judge-side reference

Practical Exam ANSWER SHEET DECEMBER 9th, 2019 Practical Exam Answer Sheet 16th IJSO 2019, Doha, Qatar Practical exam, Dec. 9th , 2019 Page 2 of 13 Country Student 1 Student 2 Student 3 . . .

Expected submission

One shared written scientific report.

Track-specific tools

`query_rules`, `use_calculator`, `inspect_environment`, `check_ijso_measurements`

Team and resource rules

3-person team. Jointly prepare and internally review one shared artifact. Prohibited resources include internet, code execution.

International Olympiad on Astronomy and Astrophysics (2015)
Example: ioaa_group_2015
Record ID: `ioaa_group_2015`

Task prompt
TEAM COMPETITION PROBLEM The Borobudur Temple, built in the 9th century, is the biggest Buddhist temple in the world built of stone. It is considered to be one of the wonders of the ancient world. It has nine stacked platforms. The first six platforms have a square form and beautifully decorated with numerous reliefs. The upper three platforms are cir cular, with 72 small bell -shaped stupas, surrounding one large central stupa (main stupa). The altitude of the ground close to the entrance is 265 m above mean sea level. Note that the height of the temple is 35 m (measured from the top of the main stupa to the ground). We believe that the constructors of this temple had quite a good knowledge of astronomy. In search of the possible existence of the astronomical alig nment between Borobudur and any celestial bodies, we will make some measurements. Observers are distributed around the temple, mainly on the eastern and western sides. The observing station number card (placed on the . . .

Track coverage
4 records in this track; source years 2014, 2015, 2021, 2025; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Uses the archived year-specific official solution; live group-task logistics are not reproduced.

Judge-side reference
TEAM COMPETITION Solution The position of the observer is given after being determined using GPS: longitude, latitude, and altitude. In fact, only latitude is necessary in the analysis. We . . .

Expected submission
Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools
`query_rules, use_calculator, read_star_chart, check_ioaa_observations`

Team and resource rules
5-person team. Equal contestants assigned to one international group team collaborate on the host-designed task and produce one shared answer. Prohibited resources include internet . . .

International Olympiad in Artificial Intelligence: Team Challenge (2025)
Example: Team Challenge: Factory of the Future (Galbot)
Record ID: `ioai_team_2025_factory`

Task prompt
Team Challenge See you in 2025.08.04 UPDATES Date Changes Made Date Changes Made 0602 Original release Team Challenge "Factory of the Future" Grand Challenge Have you ever visited a modern factory? Rows of robotic arms precisely sort and assemble parts, and large equipment, cars, electronics, and everyday products are standardized and produced on assembly lines in a matter of minutes. However, there are still certain tasks that are difficult to standardize or require flexible adjustments, which currently rely on human cooperation. But in the near future, these tasks will be handled by general-purpose humanoid robots. They will help free humans from repetitive manual labor, giving people more time to engage in creative and passionate work. 1.Introduction You and your team are general-purpose humanoid robot engineers at the Factory of the Future. You have designed and built an impressive humanoid robotGalbot. Please write a program to help Galbot complete its challenges in the Factory . . .

Track coverage
1 record in this track; source years 2025; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Static program and design review against the 2025 task guide; no simulator or physical . . .

Judge-side reference
data/rubrics/ioai_team_open_artifact_llm_20260917_v1.json

Expected submission
BENCHMARK SUBMISSION: Submit a Python robot-control program and technical explanation for the warehouse task in the supplied 2025 guide. Specify required scene/sensor inputs . . .

Track-specific tools
`query_rules, execute_code, check_ioai_artifact_manifest`

Team and resource rules
4-person team. Collaborate only within the team under contest rules.

International Linguistics Olympiad (2025)
Example: iol_team_2025
Record ID: `iol_team_2025`

Task prompt

en(B) Twenty-second International Linguistics Olympiad Taipei(Taiwan),2027 July2025 Team Contest Problem Thisproblemlooksatafewlanguagesspokenineastern Nepalbelongingtothe Kirantibranchof the Sino-Tibetanfamily. Aspartofthesamelanguagefamily,Kirantilanguagesaredistantlyrelatedto the Chineselanguages,ascanbeseenbycomparingthefollowing Athpare(Kiranti)wordswiththeir Mandarin,Cantoneseand Hokkien(Chinese)cognates: Athpare sima to die Mandarin si3 to die Athpare sepma to kill Cantonese saat3 to kill Athpare khomma to look for, to search Hokkien khoan to see, look at, watch In Kirantilanguages,personalpronounscanrefertooneperson(singular: sg),twopeople(dual: du) or more than two people (plural:pl). They also distinguish between us, including you (inclusive we:we+) and us, not including you (exclusive we:we); for example,we+ du = I+you sg. Theydonotdistinguishpronounsonthebasisofgender,soforthesakeofsimplicity,femininеforms (she, her)areusedthroughoutthisproblem. The verbs in Kiranti...

Track coverage

6 records in this track; source years 2014, 2015, 2018, 2021, 2023, 2025; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Uses the archived official solution as a judge-only reference.

Judge-side reference

en(B) Twenty-second International Linguistics Olympiad Taipei(Taiwan),2027 July2025 Team Contest Answers Part I. (a) 1 F 2 D 3 B 4 E 5 C 6 H 7 A 8 G 9 M 10 O 11 L 12 J 13 N 14 K 15 P 16 I . . .

Expected submission

Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools

`query_rules`, `check_linguistic_rule_table`

Team and resource rules

4-person team. Up to four contestants work jointly in an allocated team room on one problem set and hand in one script; the team leader is not a contestant. Official duration: 240 minutes. . . .

International Tournament of Young Mathematicians (2014)
Example: itym_2014
Record ID: `itym_2014`

Task prompt
PROBLEMS FOR THE 6th INTERNATIONAL TOURNAMENT OF YOUNG MATHEMATICIANS JULY 2014, BREMEN (GERMANY) Contents Notation 1 1. Poles of Polygons 2 2. Free Subsets 2 3. A Recursive Sequence 3 4. Binary Matrices 3 5. An Egyptian Game 4 6. Ridges of Functions 4 7. Decimal Representations 5 8. Cycles in a Permutation Group 5 9. An Equation inZn 6 10. Reflexive Polynomials 6 11. Points, Lines and Planes 7 Keywords: 1. geometry, polygon 2. additive number theory 3. analysis, recursive sequence 4. linear algebra, matrix norm 5. Egyptian fraction, inequality 6. analysis, continuous function 7. Diophantine approximation, algebraic number 8. permutation group, cycle 9. number theory, residue class 10. polynomial over a finite field, permutation 11. combinatorial geometry. Notation N ={1, 2, 3,...} set of natural numbers (positive integers) Z, Q, C sets of integer, rational and complex numbers R, R2, R3 real line, real plane and 3-dimensional real space gcd(x1,...,x n) greatest common divisor . . .

Track coverage
15 records in this track; source years 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Uses the declared benchmark rubric, not a full official competition score.

Judge-side reference
Score only the written mathematical research paper. No live mathematical fight, oral defense, opponent or reviewer exchange is assessed.

Expected submission
One shared written mathematical research paper.

Track-specific tools
`query_rules`, `use_calculator`, `execute_code`, `work`, `render_pdf`, `check_math_research_dependencies`

Team and resource rules
6-person team. Jointly prepare and internally review one shared artifact.

International Young Physicists' Tournament (2026)
Example: iypt_2026
Record ID: `iypt_2026`

Task prompt
IYPT 2026 Problems Problems for the 39th IYPT 2026 Approved by the IOC on 7 July 2025 Problems for the 39th IYPT 2026 1. Invent yourself A self-starting siphon can be made using a piece of rigid tubing bent into a specific shape. When the siphon is partially immersed in water, it begins siphoning water without the need for initial suction. Investigate how the relevant parameters, such as the geometry, affect the siphoning process. 2. Electrical damping A magnet suspended by a spring will display simple harmonic motion when displaced. If the magnet oscillates within a coil connected to a resistor, its motion will be damped. Investigate the factors that affect the damping. 3. Ring fountain When a flat metal ring falls from a certain height into a water tank, it generates a fountain that can shoot water high into the air. How does the maximum height of the fountain depend on the rings parameters? 4. Oil flow A thin layer of cooking oil on a flat metal surface flows outwards when heated. . . .

Track coverage
16 records in this track; source years 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Uses the declared benchmark rubric, not a full official competition score.

Judge-side reference
Score only the written research report: theory, investigation design, documented evidence and clarity. No physical execution, oral defense, opponent or reviewer exchange is assessed.

Expected submission
One shared written physics research report.

Track-specific tools
`query_rules, use_calculator, execute_code, work, render_pdf, prepare_physics_research_outline`

Team and resource rules
5-person team. Jointly prepare and internally review one shared artifact. Prohibited resources include outside communication during fight.

Philip C. Jessup International Law Moot Court Competition (2024)
Example: jessup_2024
Record ID: `jessup_2024`

Task prompt
INTERNATIONAL COURT OF JUSTICE SPECIAL AGREEMENT BETWEEN THE REPUBLIC OF ANTRANO (APPLICANT) AND THE KINGDOM OF REMISIA (RESPONDENT) TO SUBMIT TO THE INTERNATIONAL COURT OF JUSTICE THE DIFFERENCES BETWEEN THE PARTIES CONCERNING THE STERREN FORTY jointly notified to the Court on 15 September 2023 COUR INTERNATIONALE DE JUSTICE COMPROMIS ENTRE LA REPUBLIQUE DANTRANO (DEMANDERESSE) ET LE ROYAUME DE REMISIA (DEFENDEUR) VISANT A SOUMETTRE A LA COUR INTERNATIONALE DE JUSTICE LES DIVERGENCES QUI OPPOSENT LES DEUX PARTIES CONCERNANT LES QUARANTE DE STERREN notifie conjointement a la Cour le 15 septembre 2023 JOINT NOTIFICATION ADDRESSED TO THE REGISTRAR OF THE COURT The Hague, 15 September 2023 On behalf of Applicant, the Republic of Antrano, and Respondent , the Kingdom of Remisia, in accordance with Article 40(1) of the ...

Track coverage
1 record in this track; source years 2024; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Evaluates the written memorial only; oral advocacy is not scored.

Judge-side reference
Jessup rubric: legal analysis, argumentation (Applicant + Respondent), written memorial quality, citation of international law.

Expected submission
Submit the written memorial/case analysis and a concise oral outline.

Track-specific tools
`query_rules, web_search, check_jessup_authority_table`

Team and resource rules
5-person team. Collaborate only within the team under contest rules. Prohibited resources include calculator, code execution.

Mathematical Contest in Modeling (MCM) (2024)
Example: mcm_2024_A
Record ID: `mcm_2024_A`

Task prompt
BENCHMARK SUBMISSION OVERRIDE: Write a complete mathematical modeling paper in self-contained LaTeX and submit its compiled PDF. Include abstract, assumptions, model formulation, derivations, results, validation or sensitivity analysis, limitations, conclusion and references. The benchmark uses an LLM judge; paid official evaluation and membership-only winning papers are not required. | 2024 by COMAP, Inc. | www.comap.org | www.mathmodels.org | info@comap.org | 2024 MCM Problem A: Resource Availability and Sex Ratios While some animal species exist outside of the usual male or female sexes, most species are substantially either male or female. Although many species exhibit a 1:1 sex ratio at birth, other species deviate from an even sex ratio. This is called adaptive sex ratio variation. For example, the temperature of the nest incubating eggs of the American alligator influences the sex ratios at birth. The role of lampreys is complex. In some lake habitats, they are seen as...

Track coverage
17 records in this track; source years 2016, 2017, 2021, 2022, 2024, 2025; record unit(s): session.

Evaluation and scope
ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Scores the submitted report; COMAP originality checks and award-panel calibration are not . . .

Judge-side reference
COMAP MCM/ICM report rubric: modeling, analysis, communication; 25 page PDF report.

Expected submission
BENCHMARK SUBMISSION OVERRIDE: Write a complete mathematical modeling paper in self-contained LaTeX and submit its compiled PDF. Include abstract, assumptions, model formulation . . .

Track-specific tools
`query_rules, use_calculator, execute_code, work, render_pdf, prepare_mcm_model_record`

Team and resource rules
3-person team. Only registered team members may work on the solution. Official duration: 5940 minutes. Prohibited resources include outside communication.

MIT Mystery Hunt (2012)

Example: 2012/puzzles/mayan_fair_lady/initial_public_offering/

Record ID: `mystery_hunt_01610`

Task prompt

Initial Public Offering A 1. SBR, AYL? ITL, SIG. ABS, WHT, WIAWW. (7/16) 2. CROAOF. JFNAYN. YCBSBAC, AFDULE. (7/16) 3. AIAM, NCFAB, TLLJLDHSH. TSITSLDWHL. TLLJAITH. (4/13) 4. OTFDOC, MTLGTM: APIAPT. (5/24) B 1. ILTSONITM. (2/13) 2. GA, MMD. (9/12) 3. IGMHAOHCR. (7/12) 4. YCHTT! (2/11) 5. WWGHIFTC. (3/12) 6. MTFBWY. (5/8) 7. T, IAFWNIKA. (7/13) 8. YTTM? (6/10) 9. R. (4/11) 10. F, MD, IDGAD. (13/15) C 1. LMMSOAT, EHCAW. ACASASDBHAFMMA. (9/16) 2. JAJWUTHTFAPOW. JFDABHCAJCTA. (3/11) 3. HDSOAW. HDHAGF. ATKHAATKMCPHTA. (4/12) D 1. D, J! (18/18) 2. IGWTMROMH (7/12) 3. YCBALATHYW (WYML) (4/7) 4. FM (GI) (1/19) 5. DYHTPS? (3/13) 6. M, IABGN (3/9) 7. IDKHTLH (7/20) 8. TRTT . . .

Track coverage

29 records in this track; source years 2010, 2012, 2015, 2016; record unit(s): question.

Evaluation and scope

ready via `gold_answer_v1`; selected record unit: question. *Scope:* Single puzzle answer accuracy; not full-hunt progression, timing or coin finding.

Judge-side reference

CARSEAT

Expected submission

Submit the team's final answers in numbered order, using exact values when required.

Track-specific tools

`query_rules, execute_code, web_search, check_puzzle_extraction`

Team and resource rules

12-person team. The configured agents may freely coordinate inside their benchmark team; they may not obtain answers from another competing team. Prohibited resources include calculator, external teams.

NYU CTF Bench (CSAW) (2017)
Example: CSAW-Finals / crypto / ECXOR
Record ID: `nyu_ctf_2017f-cry-ecxor`

Task prompt
I hear that elliptic curves make it safe to use smaller key sizes. Can you break this curve25519-encrypted message?

Track coverage
44 records in this track; source years 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023; record unit(s): question.

Evaluation and scope
via `gold_answer_v1`; selected record unit: question. *Scope:* Final-flag benchmark only; guided subtasks and their partial credit are not reproduced. . . .

Judge-side reference
flag{generalizing_vignere_to_arbitrary_groups_is_not_good}

Expected submission
Submit the recovered flag(s) / subtask answers clearly labeled.

Track-specific tools
`query_rules, execute_code, web_search, start_environment, reset_environment, inspect_environment, execute_environment_command, prepare_nyu_ctf_flag`

Team and resource rules
5-person team. Collaborate only within the team under contest rules. Prohibited resources include calculator.

Odyssey of the Mind (2019)

Example: odyssey_of_the_mind_2019_synopses

Record ID: `odyssey_of_the_mind_2019_synopses`

Task prompt

Create a complete written script and staging plan for the 2019 Opposites Distract public synopsis below. Include dialogue, characters, scene sequence, two successful distractions, distinct dramatic styles, attention-getting effects and the final discovery/capture. Describe how effects could be produced; do not claim any live performance occurred. This is a public-synopsis writing benchmark, not the member-only full problem or an official 350-point tournament score. Five written criteria are each worth 20 benchmark points: plot requirements, distractions, dramatic styles, effects and humorous coherence. 5. Opposites DistractDivisions I, II, III & IVDisagreements can distract groups from seeing the bigger picture. Teams will create and present a humorous performance about a sneaky character that distracts others while trying to take control of anything the team wishes. In the performance it will lure others into silly arguments and be successful two times. The arguments . . .

Track coverage

17 records in this track; source years 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope

via `rubric_llm_v1`; selected record unit: session. *Scope:* Public synopsis only; private full-problem constraints are not asserted or scored.; Five . . .

Judge-side reference

data/rubrics/odyssey_of_the_mind_mini_written_20260917_v1.json

Expected submission

Create a complete written script and staging plan for the 2019 Opposites Distract public synopsis below. Include dialogue, characters, scene sequence, two successful distractions, distinct . . .

Track-specific tools

`query_rules`, `check_creative_constraint_map`

Team and resource rules

7-person team. Collaborate only within the team under contest rules. Prohibited resources include internet, calculator, code execution.

PUMaC Power Round (2015)
Example: PUMaC Power Round 2015
Record ID: `pumac_power_2015`

Task prompt
PUMaC 2015 Power Round We have a new theoremthat mathematicians can prove only trivial theorems, because every theorem thats proved is trivial. Richard Feynman Rules and Reminders These rules supersede any rules appearing elsewhere about the Power Round: 1. Your solutions are to be turned in when your team checks in on the morning of PUMaC or emailed to us at pumac@math.princeton.edu by 8AM Eastern Standard Time on the morning of PUMaC, November 21, 2015 with the subject line PUMaC 2015 Power Round. Please staple your solutions together, including the Power Round cover sheet (the last page of this document) as the first page. Each page should also have on it the team number (not team name) and problem number . Solutions to problems may span multiple pages, but staple them in continuing order of proof. 2. It is not necessary to do the problems in order, although it is a good idea to read all the problems, so that you know what is permissible to assume when doing each problem. However . . .

Track coverage
12 records in this track; source years 2009, 2010, 2011, 2012, 2013, 2015, 2016, 2017, 2019, 2020, 2021, 2023; record unit(s): session.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: session. *Scope:* Official solutions anchor mathematical correctness; historical hand-grading calibration . . .

Judge-side reference
Multi-part written proof packet

Expected submission
Submit a numbered proof packet with a complete justification for every claimed result, explicit references to earlier parts, and the required anonymous page labels.

Track-specific tools
`query_rules`, `check_pumac_proof_packet`

Team and resource rules
8-person team. All eight students on the team may collaborate freely on the Power Round; no one outside those eight may help. Prohibited resources include people outside the eight.

Purple Comet! Math Meet (2017)
Example: purple_comet_hs_2017
Record ID: `purple_comet_hs_2017`

Task prompt
This archived 2017 HS paper contains 30 numbered questions. Later editions' question counts do not apply. The attached original PDF is authoritative for formulas, superscripts, diagrams and tables; the text below is a searchable preview. Submit one numbered line per question in the form '<number>. <integer answer>'. PURPLE COMET! MATH MEET April 2017 HIGH SCHOOL - PROBLEMS Copyright cTitu Andreescu and Jonathan Kane Problem 1 Paul starts at 1 and counts by threes: 1, 4, 7, 10, ... At the same time and at the same speed, Penny counts backwards from 2017 by fives: 2017, 2012, 2007, 2002, ... Find the one number that both Paul and Penny count at the same time. Problem 2 The figure below shows a large square divided into 9 congruent smaller squares. A shaded square bounded by some of the diagonals of those smaller squares has area 14. Find the area of the large square. Problem 3 When Phil and Shelley stand on a scale together, the scale reads 151 pounds. When Shelley and Ryan . . .

Track coverage
28 records in this track; source years 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope
ready via `gold_answer_v1`; selected record unit: session. *Scope:* Official answer-key grading covers all questions in this archived paper. Original PDFs . . .

Judge-side reference
1. 757 2. 63 3. 84 4. 70 5. 67 6. 657 7. 46 8. 1250 9. 82 10. 738 11. 343 12. 846 13. 178 14. 13 15. 67 16. 283 17. 171 18. 200 19. 899 20. 6 21. 305 22. 148 23. 1760 24. 843 25. 64 26. 119 27. 96 28. 451 29. 63 30. 704

Expected submission
Ordered non-negative integers, one per problem: 1. N 2. N ... Use the contest's required integer encoding (for example m+n).

Track-specific tools
`query_rules, use_calculator, prepare_purple_comet_answers`

Team and resource rules
6-person team. One to six teammates may work together from the same or different locations and share one team answer form; an adult supervisor aged 21 or older is required and is not a . . .

QANTA Quiz Bowl (2015)
Example: Fine Arts
Record ID: `qanta_dev_102045`

Task prompt

This artist's use of bright dots of color that resemble shiny sequins led Alfred Barr to characterize that technique as the "Rococo" phase of his career; one such work from that time is a still-life of a cup and a bottle that is dominated by a green background. He glued a piece of oilcloth onto a canvas to create another artwork, which is framed by a piece of actual rope. In addition to making Still Life with Chair Caning, he used a piece of blue paper, sheet music, and a wood-grain pattern on another patch of paper to imitate the form of a guitar. To explain his art, this painter once wrote "art is a lie that makes us realize truth." While in Fontainebleau, this artist depicted a harlequin, a Pierrot and a monk in Three Musicians. This artist's invention of the collage style led a fellow artist to produce the first papiers colles. Separate analytic and synthetic phases characterizes the cubist work of, for 10 points, what artist who often worked with George Braque?

Track coverage

100 records in this track; source years 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020; record unit(s): question.

Evaluation and scope

ready via `gold_answer_v1`; selected record unit: question. *Scope:* Scores the extracted question independently; buzz timing, interrupts, rebounds, and . . .

Judge-side reference

Pablo Picasso

Expected submission

A short quiz-bowl style answer line. Prefer the canonical name; avoid sentences unless needed for disambiguation.

Track-specific tools

`query_rules`, `prepare_quizbowl_answer`

Team and resource rules

4-person team. Quiz bowl teams confer on bonuses; tossups are normally individual after a buzz. Prohibited resources include internet, calculator, electronic devices.

DOE Science Bowl (2016)

Example: Physics / Short Answer / bonus

Record ID: `science_bowl_sample_set_10_10a_hs_reg_2016_bonus_13`

Task prompt

An electron is fired through parallel plate electrodes that are 5 millimeters apart. A 2 millitesla magnetic field is perpendicular to the electric field between the plates, which have a potential difference of 600 volts. If the electron is NOT deflected, at what speed, in meters per second and in scientific notation, is it traveling?

Track coverage

120 records in this track; source years 2016, 2018, 2019, 2020, 2021, 2022; record unit(s): question.

Evaluation and scope

ready via `gold_answer_v1`; selected record unit: question. *Scope:* Scores one extracted question; buzzer timing, consultation eligibility, interrupts . . .

Judge-side reference

6.0 x 107

Expected submission

For multiple choice: the letter and option text. For short answer: the concise official-style answer only.

Track-specific tools

`query_rules`, `prepare_science_bowl_answer`, `buzz`, `answer_tossup`, `answer_bonus`

Team and resource rules

4-person team. Only bonus questions permit team consultation in the official event. Prohibited resources include internet, calculator, electronic reference.

Science Olympiad (USA) (2018)

Example: science_olympiad_ss1718_annotated_example

Record ID: `science_olympiad_ss1718_annotated_example`

Task prompt

Solar System - Division B, 2017-2018 example exam. Use the attached contestant packet, cover image and Image Sheets A, B and C. Submit answers numbered 1-100. This is a historical sample, so interpret date-dependent questions in its original context. SOLAR SYSTEM ANNOTATED EXAMPLE EXAM B DIVISION 2017-2018 TEAM NUMBER:_____ TEAM NAME:_____ STUDENT . . .

Track coverage

1 record in this track; source years 2018; record unit(s): session.

Evaluation and scope

via `gold_answer_v1`; selected record unit: session. *Scope:* Historical example exam, not an official tournament result.; Explicit printed aliases . . .

Judge-side reference

100 one-point items from the two original 50-point key pages; reference-based grading is not yet validated.

Expected submission

Submit numbered answers 1-100, one per line. For multiple labels separate tokens with spaces; preserve order when requested. Include units for numeric ages, periods and distances. Each answer is worth one point

Track-specific tools

`query_rules`, `read_official_materials`, `prepare_event_answer_sheet`

Team and resource rules

15-person team. Only the participants authorized by the selected season, division, and event rules may collaborate on that event. Prohibited resources include internet, code execution.

Willem C. Vis International Commercial Arbitration Moot (2019)

Example: 26th-2019-problem-incl-PO2

Record ID: `vis_moot_26th-2019-problem-incl-PO2`

Task prompt

Association for the Organisation and Promotion of the Willem C. Vis International Commercial Arbitration Moot 1 Prof. Dr. Stefan Kroll THE PROBLEM Vienna, Austria October 2018 April 2019 Oral Hearings April 13 18, 2019 Organised by: Association for the Organisation and Promotion of the Willem C. Vis International Commercial Arbitration Moot And Sixteenth Annual Willem C. Vis (East) International Commercial Arbitration Moot Hong Kong Oral Arguments March 31 April 7, 2019 Organised by: Vis East Moot Foundation Limited Twenty Sixth Annual Willem C. Vis International Commercial Arbitration Moot Association for the Organisation and Promotion of the Willem C. Vis International Commercial Arbitration Moot 2 Prof. Dr. Stefan Kroll Contents Document Page Letter by Langweiler (31 July 2018) 3 Notice of Arbitration (31 July 2018) 4 Claimants Exhibit C 1 (Email 21 March 2017) 9 Claimants Exhibit C 2 (Email 24 March 2017) 10 Claimants Exhibit C 3 (Email 28 March 2017) 11 Claimants Exhibit C 4 . . .

Track coverage

7 records in this track; source years 2019, 2021, 2022, 2023, 2024, 2025, 2026; record unit(s): session.

Evaluation and scope

ready via `rubric_llm_v1`; selected record unit: session. *Scope:* Evaluates the written memorial; oral pleading, tribunal questions, rebuttal, and . . .

Judge-side reference

Memorials + oral pleading

Expected submission

Submit the written memorial/case analysis and a concise oral outline.

Track-specific tools

`query_rules, web_search, check_vis_authority_table`

Team and resource rules

5-person team. Collaborate only within the team under contest rules. Prohibited resources include calculator, code execution.

Wharton Global High School Investment Competition (2021)

Example: case-study-2021-2022

Record ID: `wharton_investment_case-study-2021-2022`

Task prompt

Case Study Case Study for 2021-2022 PLEASE NOTE: This case study is from the 2021-2022 competition, and should be used for reference purposes only. The current case study for the competition can be found here: https://globalyouth.wharton.upenn.edu/investment-competition/case-study/ You cant build an effective investment strategy without first knowing your client... Meet Nichole Jordan! CASE STUDY: You are an analyst team of recent college graduates, working at an up-and-coming asset management company, Wharton Global Asset Management (WGAM). The firm currently manages a 100,000,000 portfolio that is invested across several different sectors, representing a broad range of industries and companies. The members of your team hope to one day become portfolio managers who make the final investment decisions for WGAMs portfolio. WGAMs current portfolio manager (your teams teacher/advisor) recently met with a . . .

Track coverage

3 records in this track; source years 2019, 2020, 2021; record unit(s): session.

Evaluation and scope

via `rubric_llm_v1`; selected record unit: session. *Scope:* Current same-competition qualitative criteria reused for the 2021 case; historical . . .

Judge-side reference

data/rubrics/wharton_investment_open_artifact_llm_20260917_v1.json

Expected submission

BENCHMARK SUBMISSION: Prepare a written investment policy statement and final strategy report for the client in this archived case. State the information date, cite and date research, and . . .

Track-specific tools

`query_rules`, `use_calculator`, `execute_code`, `web_search`, `check_portfolio_weights`

Team and resource rules

5-person team. Collaborate only within the team under contest rules.

World Mathematics Team Championship (2018)
Example: WMTC 2018 Advanced: Team Round
Record ID: `wmtc_2018_advanced`

Task prompt
Advanced Level Team Round English Version Instruction: This round has 14 questions (40 minutes). Each question is worth 5 points. No point penalty for submitting wrong answer. 1. Let a be the least integer for which the equation x3 + ax + 2a + 15 = 0 has rational root. Find a2. 2. Let a and b be real numbers such that the equation 22 11x b a has exactly 5 distinct real roots. Find the least value of 2 49ab . 3. If 33 7 7 5 2a and 33 7 7 5 2b find the value of a2 + b2 2ab + 3. 4. The diameter of a football (soccer) ball is 40 3 6 cm. Three football balls lay on the ground and each one touches the other two. Another football ball is put on the top of given balls such that it touches each of the three balls. Find the distance from the ground to the highest point of the given pyramid of four balls. 5. Let A be a set of positive integers such that for any two distinct x and y from A we have: 1 1 1 25xy . Find the maximum possible number of elements of A. 6. All 10 digit numbers with . . .

Track coverage
3 records in this track; source years 2018; record unit(s): session.

Evaluation and scope
ready via `gold_answer_v1`; selected record unit: session. *Scope:* Scores the archived 2018 Team Round only; live paper collection and proctoring are not reproduced.

Judge-side reference
2018 Team Round: 14 questions, 5 points each, no wrong-answer penalty.

Expected submission
Numbered team answers matching the team-round problems in order.

Track-specific tools
`query_rules`, `prepare_wmtc_team_sheet`

Team and resource rules
6-person team. Six equal contestants may collaborate throughout the Team Round and produce one shared answer sheet. Prohibited resources include internet, calculator, electronic devices.

World Scholar's Cup: Collaborative Writing (year not specified)
Example: wsc_writing_gq_001
Record ID: `wsc_writing_gq_001`

Task prompt
If you walked out of your home without knowing youd accidentally time traveled into the past, how long would it take you to realize what had happened? What if they had sent you back ten years, or thirty, or a hundred? Discuss with your team: how far into the past would you need to be to realize instantly that you were in a different era? One clue to your whenabouts might be the text around you: not just the headlines on newspapers and store signs, but the fonts theyre printed in. Consider some of the history of typography , then discuss with your team: how different would the world look today if Microsoft had chosen Comic Sans instead of Calibri as its default typeface in the early 2000sor as its successor 20 years later. The London Underground also decided to update its font in 2016 for a more modern lookdid it succeed? Be sure to learn the difference between serif and sans serif fonts, and then see which ones are used more widely. Does the same distinction apply in non-Western...

Track coverage
42 records in this track; source years not recorded; record unit(s): question.

Evaluation and scope
via `rubric_llm_v1`; selected record unit: question. *Scope:* Three-author staged written benchmark with controller-private writing and cyclic peer . . .

Judge-side reference
Official Collaborative Writing rubric: Clarity, Content, Style, and Originality are each scored 15-25; total range 60-100.

Expected submission
Three complete responses to three distinct prompts, one original author per response.

Track-specific tools
`query_rules`,
`start_independent_writing`,
`write_writing_response`,
`read_writing_response`,
`start_peer_review`,
`review_writing_response`,
`submit_writing_packet`

Team and resource rules
3-person team. Shared planning, private independent writing, assigned peer editing, three-response submission. Prohibited resources include internet, calculator, code execution, electronic devices.