FATE: FIRST AGENT TEAMS EXAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world professional work requires collaboration among coworkers with distinct responsibilities, information, and permissions, yet existing agent benchmarks provide limited coverage of this setting. We introduce **First Agent Teams Exam (FATE)**, a benchmark for evaluating agent teams on realistic professional tasks. Constructed by 32 domain professionals, FATE comprises 112 tasks across 14 domains and 129 occupations, covering both unimodal and multimodal tasks. Agents use tools and communicate under role-specific resource and permission boundaries to produce intermediate artifacts and final deliverables. We evaluate eight LLMs and compare four communication protocols. Current agent teams achieve a best final evaluation score of only 42.0, with limited collaboration gains at substantial computational cost. These gains vary with communication protocols and execution parallelism, while coordination and artifact delivery remain major bottlenecks to reliable teamwork. We hope FATE will provide a foundation for systematically evaluating agent teams and advancing reliable agent collaboration.

1 INTRODUCTION

Team collaboration is central to complex professional work, where completing a task often requires contributions from members of different professions. By combining complementary expertise, dividing complex tasks, working in parallel, and cross-checking results, teams can accomplish tasks that exceed the capabilities of individual members (Wegner, 1987; Grosz & Kraus, 1996). As shown in Figure 1, assessing the reliability of a cooling system involves hardware engineers defining cooling conditions and clarifying qualification scope, electrical engineers assessing fan sensing, mechanical engineers preparing and running computational fluid dynamics simulations, validation engineers assessing fan-failure risks and finalizing the reliability assessment, and managers coordinating evidence integration and issuing engineering dispositions. These team members exchange information, hand off intermediate artifacts, and coordinate their work to produce final deliverables.

As LLM-based agents are increasingly applied to complex, long-horizon tasks, their applications are expanding from individual execution to collaboration within **agent teams** (Anthropic, 2026b; Agashe et al., 2025). Recent studies apply agent teams to software development (Hong et al., 2024a; Qian et al., 2024), scientific research (Swanson et al., 2025; Gottweis et al., 2026), and medical diagnosis (Chen et al., 2025b). This trend also extends to practical systems: Claude Code supports agent teams with shared task management and direct communication among teammates (Anthropic, 2026b), while Kimi K2.5 introduces dynamically orchestrated agent swarms for parallel task execution (Team, 2026). Looking ahead, agents could become digital coworkers embedded in real-world workflows, sharing professional responsibilities with humans and other agents. Realizing this vision raises a fundamental question: *Can agent teams with distinct occupational responsibilities collaborate reliably to accomplish complex real-world tasks and produce high-quality deliverables?*

However, existing agent benchmarks do not fully address this question. On the one hand, recent benchmarks increasingly evaluate agents on realistic and economically meaningful professional work, broadening coverage across occupations and workplace environments (Patwardhan et al., 2026; Li et al., 2026c; Sun et al., 2026; Hu et al., 2026; Li et al., 2026a; Xu et al., 2026b). However, they primarily focus on individual agents' ability to complete long-horizon tasks. On the other hand, multi-agent benchmarks evaluate general communication and coordination but offer limited coverage of real-world professional collaboration (Zhu et al., 2025; Zhang et al., 2026; Mazumder et al., 2026; Yu et al., 2026; Kim et al., 2026). In particular, they often do not capture how members with distinct occupational responsibilities, input resources, and operational permissions coordinate interdependent work and integrate intermediate artifacts into final deliverables.

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.

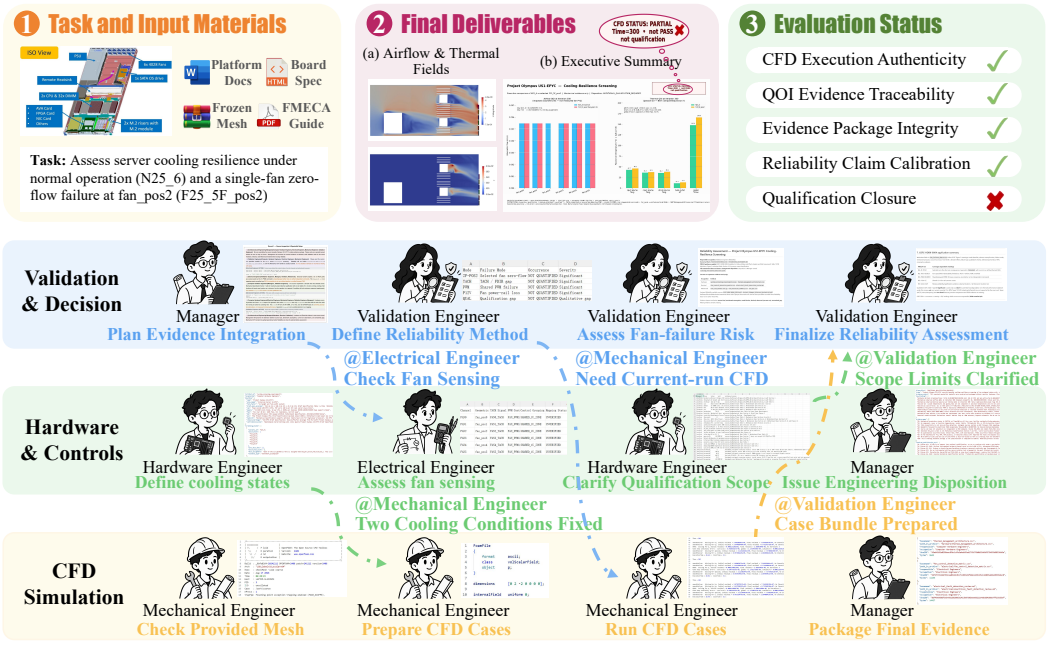


Figure 1: Overview of FATE.

To this end, we introduce **First Agent Teams Exam (FATE)**, a benchmark in which agents with distinct professional roles work as coworkers, using tools and communicating to jointly produce final deliverables. FATE follows two design principles: **(1) Authenticity.** Tasks are grounded in publicly available multimodal input resources from publications, repositories, databases, and other public sources, preserving their original formats where possible. Occupational roles and responsibilities are informed by O*NET (Peterson et al., 2001) and professional practice, while differentiated resource access, operational permissions, and task dependencies reflect real-world collaboration. **(2) Diversity.** FATE covers professional tasks across multiple domains, with varied workflows, collaboration structures, and teams of complementary occupations. Tasks require diverse intermediate artifacts and final deliverables, including documents, spreadsheets, presentations, and code.

Specifically, we recruit 32 domain professionals to construct and annotate tasks, specifying occupational responsibilities, resource access, operation permissions, reference workflows, and atomic evaluation criteria. Each task undergoes a three-stage quality verification process comprising automated validation, expert cross-review, and evaluator testing with manually constructed valid and invalid deliverables. To assess both individual contributions and team outcomes, we use **Member Evaluation (ME)** and **Final Evaluation (FE)** to evaluate members' intermediate artifacts and teams' final deliverables, respectively. The resulting benchmark comprises **112 tasks**, including 74 unimodal and 38 multimodal tasks, spanning **14 domains** and **129 distinct occupations**. On average, each task involves 6.10 roles, 14.81 input resources, 9.51 intermediate artifacts and final deliverables, and 7.95 ME evaluation entries, and takes 1.7 hours to complete.

To assess how well current agent teams handle these professional tasks, we evaluate eight LLMs using Codex CLI, Claude Code, and OpenCode, and compare four communication protocols. Our experiments reveal the following key findings. **(1) Current agent teams struggle with professional tasks.** The best-performing system achieves an FE score of only 42.0, while average FE declines from 39.8 on unimodal tasks to 32.2 on multimodal tasks. **(2) Collaboration offers benefits but incurs substantial costs.** Multi-agent execution yields limited overall gains at higher token and time costs, yet outperforms the single-agent baseline in aggregate across three domains under matched budgets. **(3) Communication and execution structure matter.** Point-to-Point Communication performs best across all three domains examined in the protocol comparison, while greater execution parallelism is associated with larger relative gains over single-agent execution. **(4) Communication and artifact delivery are major sources of failure.** Communication coordination failures and output-interface misalignment account for 34% and 32% of analyzed errors, respectively, revealing persistent difficulties in exchanging information, managing dependencies, and delivering artifacts.

2 THE FATE BENCHMARK

This section introduces the FATE benchmark, and is organized into the following five subsections: **Task Formulation** (§ 2.1), **Evaluation Metrics** (§ 2.2), **Construction Pipeline** (§ 2.3), **Annotation and Verification** (§ 2.4), and **Benchmark Statistics** (§ 2.5).

2.1 TASK FORMULATION

A team task involves agents with real-world occupational roles and separate access to input resources working as coworkers. Team members perform two types of actions, **tool use** and **communication**, to produce **intermediate artifacts** and **final deliverables**. Formally, we define a task as

$$\mathcal{Q} = \left(x, \mathcal{I}, \mathcal{T}, \{(\mathcal{R}_i, \mathcal{C}_i)\}_{i=1}^{N}, \mathcal{C}_F\right), \tag{1}$$

where x specifies the overall task, $\mathcal{I}$ is the set of input resources, and $\mathcal{T} = \{t_1, \ldots, t_N\}$ is an agent team of N members. Each member t_i has a role configuration $\mathcal{R}_i = (j_i, \mathcal{I}_i, \mathcal{O}_i)$, comprising its job responsibilities j_i, initially accessible input resources $\mathcal{I}_i \subseteq \mathcal{I}$, and operation permissions $\mathcal{O}_i$. Strict access isolation prevents members from accessing input resources exclusive to others. $\mathcal{C}_i$ and $\mathcal{C}_F$ specify the criteria for the intermediate artifacts to be produced by member t_i and the final deliverables to be produced by the team, respectively.

Given the task x, a planner $\mathcal{P}$ decomposes it into subtasks and assigns them to team members based on their professions, accessible input resources, and operation permissions, yielding $(x_1, \ldots, x_N) = \mathcal{P}(x, \mathcal{T}, \{\mathcal{R}_i\}_{i=1}^{N})$, where x_i denotes the subtask assigned to member t_i. The planner $\mathcal{P}$ and all team members share the same model. Each member then selects a tool-use or communication action:

$$a_i^s = t_i(x, x_i, \mathcal{R}_i, h_i^s), \qquad a_i^s \in \mathcal{A}_i^{\text{Tool}} \cup \mathcal{A}_i^{\text{Comm}}, \tag{2}$$

where a_i^s denotes the action generated by member t_i at decision step s, and h_i^s includes local observations, previous actions, tool execution results, and received messages and intermediate artifacts.

Tool-use actions involve writing and executing code through agent harnesses such as Codex CLI and Claude Code to carry out subtasks, prepare intermediate artifacts for teammates, and integrate their contributions into final deliverables $\mathcal{D}_F$. Communication actions allow member t_i to share text messages or intermediate artifacts from $\mathcal{D}_i^s$ with teammates, where $\mathcal{D}_i^s$ denotes the intermediate artifacts it produces up to step s. These exchanges support information sharing, subtask coordination, and downstream work, with delivery governed by the communication protocol.

2.2 EVALUATION METRIC

For FATE, we use the **Member Evaluation (ME)** and **Final Evaluation (FE)** to measure the normalized evaluation score of tasks for intermediate artifacts and final deliverables, respectively:

$$\text{ME} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} m_{k,i}, \qquad \text{FE} = \frac{1}{K}\sum_{k=1}^{K} f_k, \tag{3}$$

where K is the number of task instances and N_k is the team size for instance k. Here, $m_{k,i}$ measures how well member i's intermediate artifacts $\mathcal{D}_{k,i}$ satisfy the evaluation criteria $\mathcal{C}_{k,i}$, while f_k measures how well the team's final deliverables $\mathcal{D}_{k,F}$ satisfy the evaluation criteria $\mathcal{C}_{k,F}$.

For an artifact set $\mathcal{D}$ and its corresponding evaluation criterion set $\mathcal{C}$, let $c \in \mathcal{C}$ denote an individual criterion and $\text{Eval}(\mathcal{D}, c) \in [0, 1]$ indicate whether $\mathcal{D}$ satisfies c. We compute the proportions of satisfied criteria as $m_{k,i} = \frac{1}{|\mathcal{C}_{k,i}|}\sum_{c\in\mathcal{C}_{k,i}} \text{Eval}(\mathcal{D}_{k,i}, c)$ and $f_k = \frac{1}{|\mathcal{C}_{k,F}|}\sum_{c\in\mathcal{C}_{k,F}} \text{Eval}(\mathcal{D}_{k,F}, c)$, where $|\mathcal{C}|$ denotes the number of criteria in $\mathcal{C}$. Each criterion is evaluated using rule-based evaluation or LLM-based evaluation. **(1) Rule-based evaluation** uses programmatic checks for objectively verifiable criteria, such as file existence and format compliance. **(2) LLM-based evaluation** uses an LLM judge to assess criteria requiring semantic understanding, such as content correctness and completeness. For LLM-based evaluation, we deliberately design each criterion as an objective question with a clear, unambiguous yes-or-no answer that is easy to judge.

2.3 CONSTRUCTION PIPELINE

As shown in Figure 2, we adopt a five-stage pipeline to transform real-world professional work into agent team tasks. **(1) Domain taxonomy.** We select real-world cowork scenarios that require collaboration across professions, guided by occupational taxonomies such as O*NET (Peterson et al.,

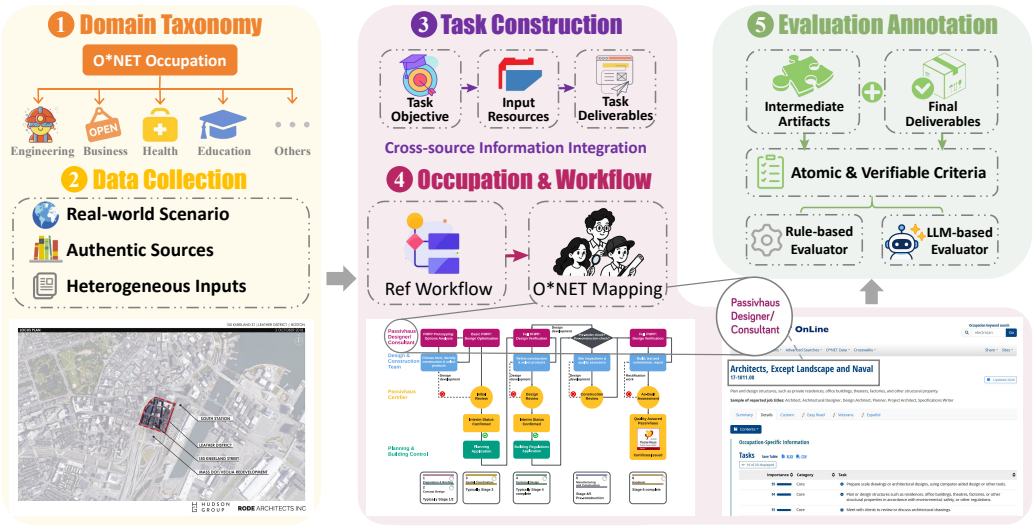


Figure 2: FATE construction pipeline, comprising domain taxonomy, data collection, task construction, occupation and workflow annotation, and evaluation annotation.

2001). We prioritize scenarios with clear input resources, objectives, and final deliverables. **(2) Data collection.** We collect publicly available, non-sensitive multimodal input resources from academic publications, open-access repositories, open databases, and other public web sources, preserving their original formats whenever possible. **(3) Task construction.** Based on the collected materials, we adapt real-world work into benchmark tasks by specifying the task objective, input resources, and final deliverables. We prioritize complex tasks that require cross-source information integration and diverse operations. **(4) Occupation and workflow annotation.** We map team members to O*NET occupations and specify their job responsibilities, accessible input resources, and operation permissions. Based on real-world industry workflows, we annotate a reference DAG to capture task dependencies and opportunities for parallel collaboration. **(5) Evaluation annotation.** We design atomic and verifiable evaluation criteria for both members' intermediate artifacts and final team deliverables, and implement automatic evaluation using rule-based and LLM-based evaluators.

2.4 Annotation and Verification

We maintain strict standards for the domain expertise and data quality through two processes: human annotation and quality verification. For **human annotation**, we recruit 32 professionals across 14 domains, with at least two experts per domain and a minimum of three years of experience in their respective fields. These experts select authentic data sources, construct realistic tasks, and develop accurate evaluation criteria based on their domain expertise. We use real-world occupational data from sources such as the NASA Exoplanet Archive, GovInfo, HCPSS, and NARA Bulletin. All sensitive and private information are removed. The tasks reflect the experts' professional work and are reviewed multiple times to ensure authenticity and clarity. Annotating each task takes an average of 6.3 hours, and we compensate experts at a rate of USD 30 per hour. Experts within each domain cross-check one another's work to verify quality.

For **quality verification**, we use a three-stage validation process. **(1) Automated validation.** Codex checks each task for ambiguity, missing required materials, and conflicting constraints. **(2) Expert verification.** Experts review the task description, input materials, and evaluation criteria, then cross-check one another's assessments. This process verifies correctness and consistency for tasks, materials and criteria while ensuring that the evaluation criteria cover the key task requirements. **(3) Evaluator testing.** Experts manually construct both valid and invalid deliverables to verify the correctness of the evaluator. For rule-based evaluation, experts ensure that valid deliverables pass the checks and invalid deliverables fail. For LLM-based evaluation, all criteria are objective rather than subjective. We use GPT-5.6 Terra as the evaluator. Human validation shows that its judgments agree with human judgments in 92% of criteria from 300 randomly selected criterion instances, indicating that this evaluation method is highly reliable. After three stages of rigorous checks, the final benchmark contains 112 tasks, with a retention rate of 58% from the initial 194 tasks.

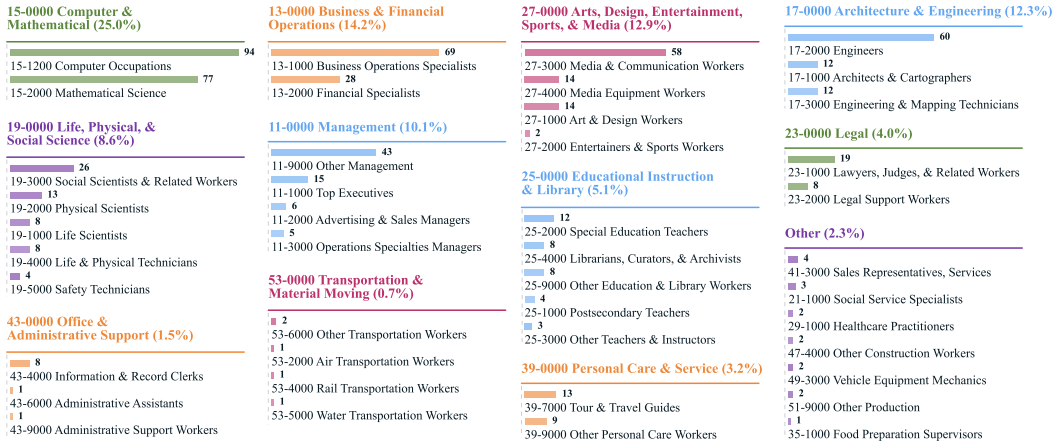


Figure 3: Occupational distribution after mapping to O*NET. Percentages show the proportion of each major group, and bars show the count for each minor group.

2.5 Benchmark Statistics

We present a statistical overview of FATE, demonstrating its **authenticity** and **diversity**. FATE comprises 112 tasks, including 74 unimodal and 38 multimodal tasks, with an average of 6.10 roles, 14.81 input resources, 9.51 artifacts and deliverables, 7.95 ME metrics, and 12.59 FE metrics per task. We illustrate it from two aspects: occupational coverage and data coverage. For **occupational coverage**, FATE captures the complexity and diversity of real-world professional works. **(1) Domain distribution.** FATE covers a total of 14 domains, including *Building*, *Business*, *CFD*, *Education*, *Industrial*, *Finance*, *Health*, *Investigation*, *Legal*, *Media*, *Research*, *Restaurant*, *Social* and *Tourism*. **(2) Occupational distribution.** FATE covers 18 major occupational categories, 43 minor categories, and 129 distinct occupations mapping to the O*NET-SOC hierarchy, as shown in Figure 3 and Table 7, providing broad coverage of professional occupations. For example, the major occupational category *Computer and Mathematical Occupations* encompasses the minor categories *Computer Occupations* and *Mathematical Science Occupations*. Overall, based on the ONET-SOC hierarchy, FATE covers a wide range of real-world occupations across diverse professional fields.

For **data coverage**, FATE encompasses realistic and diverse input resources and deliverables that reflect the demands of real-world professional work. **(1) Real-world data.** FATE contains 1,659 input resources and 1,065 intermediate artifacts and final deliverables derived from actual work scenarios, reflecting the complexity and processing requirements of real-world professional data, as shown in Appendix E. We preserve their original structure and realistic characteristics whenever possible, including missing values, noise, anomalies, irregular layouts, multi-table relationships, and heterogeneous file formats. **(2) Resource and artifact distribution.** As shown in Figure 4, Table 8, and Tables 9, the intermediate artifacts and final deliverables span diverse formats, like *documents*, *code*, *web pages*, *images*, *videos*, and *domain-specific files*. Further statistics on FATE are provided in Appendix B.

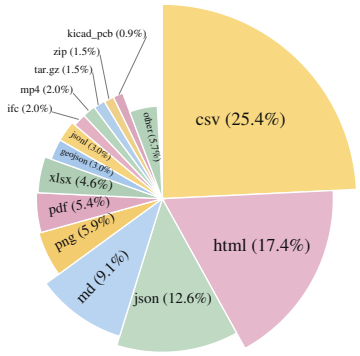


Figure 4: Distribution of final deliverable formats.

3 Experiments

This section introduces the FATE experiments, and is organized into the following three subsections: **Experimental Setup** (§ 3.1), **Main Results** (§ 3.2), and **Analysis** (§ 3.3).

3.1 Experimental Setup

Model and harness. We evaluate eight recent LLMs: Claude Opus 5, Claude Sonnet 5 (Anthropic, 2026a), GPT-5.6 Sol (OpenAI, 2026), GPT-5.6 Terra, GPT-5.5, Gemini 3.8 Flash, Grok 4.6, and DeepSeek V4 Flash (Xu et al., 2026a). We pair GPT and DeepSeek models with Codex CLI (OpenAI, 2025) and Claude models with Claude Code (Anthropic, 2025), using OpenCode (Anomaly, 2026) for the others. All models use High reasoning effort where available and Max otherwise.

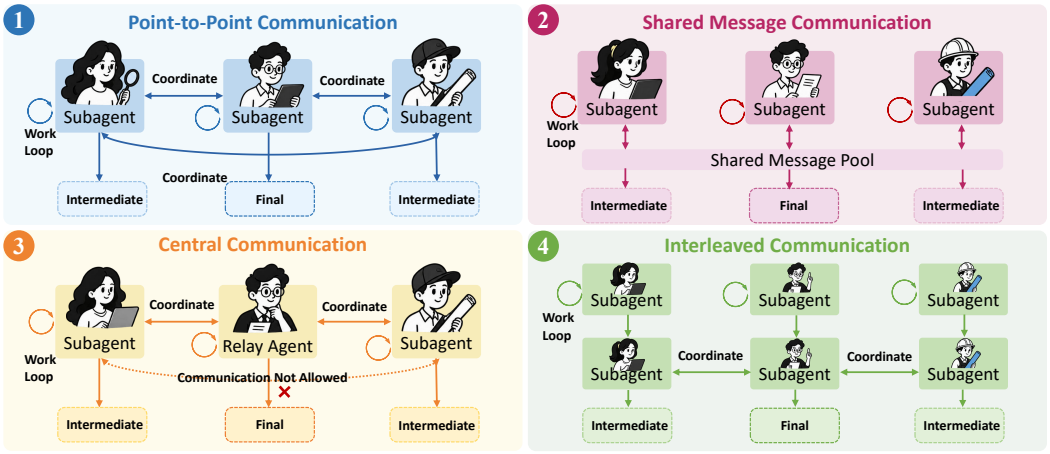


Figure 5: Agent teams communication protocol. Each agent operates in an independent harness session, we design cross-session communication protocols to enable communication among agents.

Communication protocol. The agent team communication protocols are built on harnesses. Each subagent preserves its harness's native work loop, including tool use, code editing, and memory management. We add a communication layer supporting four communication protocols, as shown in Figure 5. **(1) Point-to-Point Communication.** Subagents work on their tasks while exchanging private messages directly with their digital coworkers (Wu et al., 2023; Khatua et al., 2026). **(2) Shared Message Communication.** Subagents publish and read intermediate messages in a shared pool, forming a common collaborative context (Hong et al., 2024b; Salemi et al., 2025). **(3) Central Communication.** A relay agent aggregates and distributes messages, serving as the central hub for all communication between subagents (Wu et al., 2023; Fourney et al., 2024). **(4) Interleaved Communication.** Subagents alternate between distinct work and communication phases. We adopt Point-to-Point Communication for our main experiments, as it achieves the best performance among the four protocols (Chen et al., 2024; 2025a). Detailed comparisons are presented in § 3.3.

Implementation details. **(1) Planning.** The planner decomposes each task into subtasks in a single LLM call, then assigns them to subagents for execution. **(2) Execution and isolation.** Team members run concurrently in independent persistent harness sessions, with information and permission boundaries primarily enforced by session-level sandbox isolation. Agents exchange information through the communication protocols above. **(3) Visual perception.** For text-only models, we use GPT-5.6 Terra to support multimodal tasks. It is exposed to the agent only as a black-box MCP visual-perception tool that accepts natural language instructions and returns visual information. The agent is not informed of the underlying model identity. **(4) Accounting.** Tokens are aggregated from model-API usage statistics, while time is measured as end-to-end whole-program wall-clock time. Full implementation details are provided in Appendix D.

3.2 Main Results

Table 1: Comparison of unimodal and multimodal performance across different task modalities, ranked by overall average FE performance. ME (↑): Member Evaluation; FE (↑): Final Evaluation; Tok. (↓): Token Usage; Time (↓): Execution Time; Freq. (↑): Interaction Frequency. **Bold** and underlined values indicate the best and second-best results.

Model	Harness	Unimodal					Multimodal					Average	
		ME	FE	Tok.	Time	Freq.	ME	FE	Tok.	Time	Freq.	ME	FE
Grok 4.6	OpenCode	**60.9**	**46.0**	42.0M	8.9K	34.0	52.4	34.4	40.9M	11.0K	32.9	**57.9**	**42.0**
GPT-5.6 Terra	Codex CLI	43.0	44.5	25.8M	4.4K	40.2	48.4	**35.6**	21.3M	7.4K	53.1	45.0	41.1
DeepSeek V4 Flash	Codex CLI	57.3	44.6	75.6M	5.6K	**47.2**	**58.2**	32.8	85.1M	6.7K	46.3	57.6	40.5
Gemini 3.8 Flash	OpenCode	56.7	41.9	43.4M	7.7K	26.2	51.3	34.9	29.0M	9.7K	23.8	54.8	39.5
GPT-5.5	Codex CLI	56.3	39.2	**16.5M**	**4.3K**	40.5	53.6	35.4	15.6M	4.9K	41.2	55.3	37.8
GPT-5.6 Sol	Codex CLI	50.7	37.6	17.6M	4.8K	39.6	55.5	33.8	**10.6M**	**3.6K**	**60.6**	52.4	36.3
Claude Sonnet 5	Claude Code	37.1	33.5	17.9M	8.0K	24.5	45.9	25.0	18.6M	9.8K	26.5	40.6	30.1
Claude Opus 5	Claude Code	41.3	31.1	22.6M	5.1K	22.1	49.5	25.2	19.9M	3.9K	23.0	44.1	29.0
Average		50.4	39.8	32.7M	6.1K	34.3	51.9	32.2	30.1M	7.1K	38.4	51.0	37.1

Table 2: Comparison of performance across different task domains.

Model	Harness	Building		Business		CFD		Education		Finance		Health		Industrial	
		ME	FE	ME	FE	ME	FE	ME	FE	ME	FE	ME	FE	ME	FE
Grok 4.6	OpenCode	62.9	19.5	53.3	70.7	**83.9**	**53.8**	35.9	9.4	57.1	17.5	**89.3**	46.8	70.0	75.8
GPT-5.6 Terra	Codex CLI	24.4	16.1	43.7	**72.0**	53.3	42.8	24.7	8.8	32.5	8.0	78.8	68.2	60.0	76.4
DeepSeek V4 Flash	Codex CLI	54.5	16.4	**54.5**	54.4	67.5	43.3	**52.6**	21.1	**62.3**	**20.8**	85.9	**77.1**	**95.0**	**82.5**
Gemini 3.8 Flash	OpenCode	21.7	0.0	44.0	63.2	64.7	46.8	29.6	11.5	48.1	14.5	86.2	44.2	90.0	81.6
GPT-5.5	Codex CLI	64.2	**25.7**	45.0	54.1	59.4	36.7	47.6	**23.7**	52.6	12.2	78.1	52.6	60.0	74.9
GPT-5.6 Sol	Codex CLI	**82.7**	16.8	40.7	51.3	57.5	38.7	40.0	21.4	41.2	8.7	82.5	66.3	40.0	76.4
Claude Sonnet 5	Claude Code	42.6	14.7	34.0	60.6	29.3	9.2	37.1	6.2	49.4	13.5	53.3	31.4	0.0	0.0
Claude Opus 5	Claude Code	54.7	18.9	43.8	61.2	45.8	15.3	40.5	5.6	49.9	10.7	66.1	28.4	0.0	54.5
Average		51.0	16.0	44.9	60.9	57.7	35.8	38.5	13.5	49.1	13.3	77.5	51.9	51.9	65.2

Model	Harness	Investigation		Legal		Media		Research		Restaurant		Social		Tourism	
		ME	FE	ME	FE	ME	FE	ME	FE	ME	FE	ME	FE	ME	FE
Grok 4.6	OpenCode	55.7	**59.2**	15.2	27.3	66.7	55.6	43.5	19.1	88.6	51.0	**76.4**	**64.6**	**51.1**	**45.8**
GPT-5.6 Terra	Codex CLI	**72.0**	56.0	15.6	21.3	63.9	46.3	56.2	35.0	61.5	**70.2**	64.2	52.9	39.4	34.7
DeepSeek V4 Flash	Codex CLI	52.1	39.6	1.7	17.6	77.8	58.3	70.0	**48.7**	91.1	57.5	60.7	57.0	45.8	40.0
Gemini 3.8 Flash	OpenCode	39.2	55.2	16.7	14.1	61.1	58.3	50.5	3.3	**94.7**	42.2	68.2	58.2	44.0	44.3
GPT-5.5	Codex CLI	55.0	51.7	12.3	25.5	80.6	53.7	83.1	21.3	92.0	49.9	64.0	55.7	42.3	35.7
GPT-5.6 Sol	Codex CLI	44.9	56.5	**23.4**	**35.2**	**83.3**	**68.5**	**86.8**	46.1	76.3	48.3	51.5	18.0	46.6	38.7
Claude Sonnet 5	Claude Code	33.3	42.9	5.0	7.9	68.5	44.4	50.0	21.1	59.6	66.4	59.1	37.6	32.8	33.6
Claude Opus 5	Claude Code	41.1	43.5	4.2	17.7	62.0	35.2	51.0	14.4	89.8	40.5	29.8	40.6	35.7	35.1
Average		49.2	50.6	11.7	20.8	70.5	52.5	61.4	26.1	81.7	53.2	59.2	48.1	42.2	38.5

Current systems struggle with realistic professional team tasks. As shown in Table 1 and Table 2, current systems achieve relatively low performance on FATE benchmark. Grok 4.6 + OpenCode achieves the highest FE at only 42.0. GPT-5.6 Terra + Codex CLI and DeepSeek V4 Flash + Codex CLI achieve 41.1 and 40.5, respectively, remaining slightly below the top score. Member level performance is also limited, with the highest ME score reaching 57.9. These results indicate that current systems still struggle with the realistic professional team tasks in FATE, which require agents to work in authentic occupational roles and produce concrete professional deliverables.

Multi-agent coordination provides limited gains at substantial computational cost. As shown in Figure 6, we compare single-agent and multi-agent systems in terms of FE and resource consumption. Overall, multi-agent execution consistently incurs substantially higher computational costs, but these costs do not translate into comparable performance gains. Token consumption increases substantially under multi-agent execution across most model-harness combinations. For example, GPT-5.6 Terra improves in FE from 38.8 to 41.1, while its token consumption increases from 2.4M to 24.3M. Execution time also increases substantially under multi-agent execution.

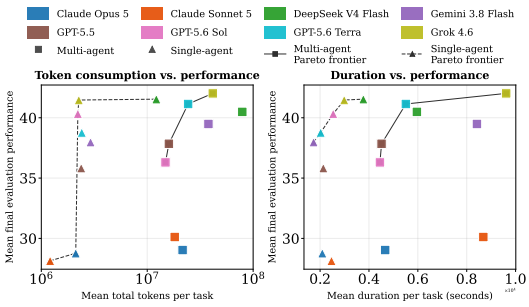


Figure 6: Comparison of single-agent and multi-agent systems in terms of FE performance, token consumption, and execution time.

Some domains and occupational roles remain hard for current systems. As shown in Table 2, Business, Health, Media, and Investigation perform relatively well, whereas Building, Education, Finance, and Legal obtain much lower domain-average FE of 16.0, 13.5, 13.3, and 20.8, respectively. This pattern is consistent with the occupational results in Figure 11: Legal Occupations score 31.8, Business and Financial Operations Occupations 39.4, and Educational Instruction and Library Occupations 40.7. These complementary results suggest that current systems struggle both with coordinating professional workflows and with executing specialized role responsibilities, especially in settings requiring domain-specific knowledge and long-horizon planning.

Multimodal tasks lead to lower final performance. Table 1 further compares system performance on unimodal and multimodal tasks. Even the best performing system on multimodal tasks achieves an FE of 35.6, compared with 46.0 on unimodal tasks. On average, the FE drops from 39.8 on unimodal tasks to 32.2 on multimodal tasks. FE declines across evaluated model-harness combinations, with drops ranging from 3.8 to 11.8. This gap suggests that multimodal tasks remain challenging for current agent systems. The Appendix F provides representative case examples.

Table 3: Comparison of different communication protocols for multi-agent collaboration.

Comm. protocol	Building				Investigation				Research				Average			
	FE	Tok.	Time	Freq.	FE	Tok.	Time	Freq.	FE	Tok.	Time	Freq.	FE	Tok.	Time	Freq.
Point-to-Point Comm.	**14.8**	20.3M	4.3K	51.57	**56.0**	6.6M	**1.5K**	19.00	**30.0**	32.2M	17.9K	**105.33**	**33.6**	19.7M	7.9K	58.63
Shared Message Comm.	0.0	**15.4M**	**2.9K**	13.86	22.2	7.7M	2.9K	18.75	10.2	**23.0M**	**6.8K**	63.17	10.8	**15.4M**	**4.2K**	31.93
Central Comm.	1.3	30.7M	10.6K	**71.14**	34.7	43.9M	8.3K	**62.25**	10.6	79.2M	9.2K	43.50	15.5	51.3M	9.4K	**58.96**
Interleaved Comm.	5.7	29.2M	4.3K	43.43	12.0	**2.6M**	2.2K	30.25	2.2	32.7M	16.0K	69.33	6.6	21.5M	7.5K	47.67

3.3 Analysis

Communication protocol selection. As shown in Figure 5 and Table 3, we analyze the impact of different communication protocols on multi-agent collaboration across Building, Investigation, and Research domains. Point-to-Point Communication achieves the best performance across all three domains, reaching FE of 14.8, 56.0, and 30.0, respectively, with a higher communication frequency of 58.63 on average. Shared Message Communication and Interleaved Communication reduce communication frequency but result in lower performance, while Central Communication incurs higher token consumption and execution time without achieving comparable accuracy.

Matched-budget comparison. To control for the effect of additional computation, we further compare single-agent and multi-agent methods under the same inference budget. Specifically, the single agent executes the task independently n times, after which an additional model call is provided with the outputs from all n runs and instructed to select the result it considers best as the final answer. As shown in Figure 7, the multi-agent method achieves an Overall FE of 28.7, clearly outperforming the same-budget single-agent baseline at 21.8. In the Investigation domain, the multi-agent score increases from 43.7 to 59.2. By comparison, simply increasing the number of single-agent runs only improves the overall FE from 21.6 to 21.8, indicating that the advantage of the multi-agent method does not arise solely from a larger computation budget.

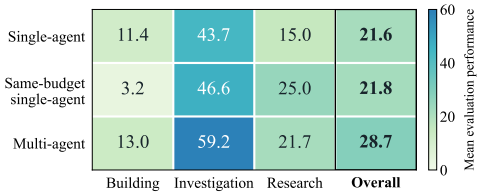


Figure 7: Performance comparison of single-agent and multi-agent settings across domains under matched computation budgets.

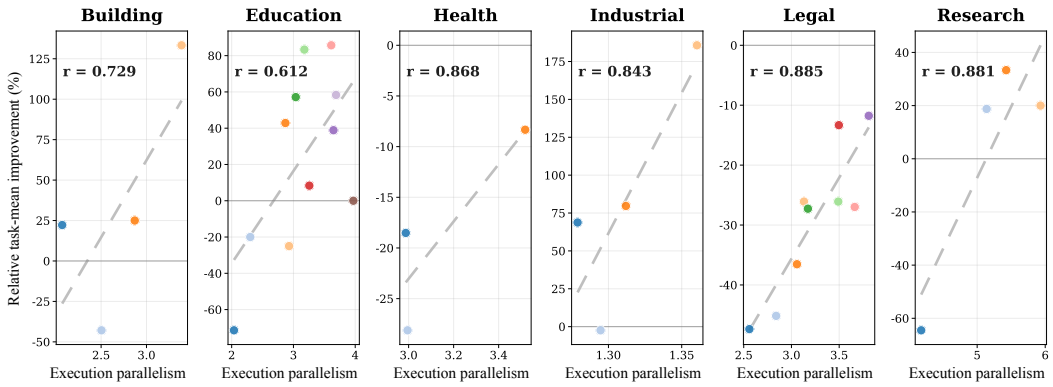


Figure 8: Task-level relationship between execution parallelism and relative performance improvement across six FATE domains. r denotes the Pearson correlation coefficient.

Applicability of multi-agent systems. The benefits of multi-agent systems may vary with the collaboration structure of a task. For tasks with clear specialization and relatively independent subtasks, agents can work concurrently and integrate their artifacts. Thus, subtask dependencies and the potential for concurrent execution are key factors in assessing the suitability of multi-agent systems. To examine this factor, we define the execution parallelism of each multi-agent run as $P_{\text{time}} = \sum_i \tau_i / T$, where τ_i denotes the effective execution time of agent i, and T denotes the total wall-clock time of the task. This metric captures the degree of parallelism realized during execution and reflects both task structure and execution dynamics. As Figure 8 shows, higher execution parallelism is associated with larger relative performance gains from multi-agent systems across multiple domains. For example, in the Building domain, parallelism levels of 2.70 and 3.39 correspond to relative performance gains of 37.5% and 133.3%, respectively. In the Health and Legal domains,

although multi-agent systems still underperform the single-agent baseline, higher execution parallelism generally accompanies a smaller performance gap. These results suggest that multi-agent systems are better suited to tasks requiring complementary expertise and parallel subtasks.

Failure mode analysis. To better understand the limitations of current systems, we conducted an analysis of their failure cases. As shown in Figure 9 and Table 11, communication coordination is the most common failure mode and accounts for 34% of the analyzed failure instances. These failures mainly include information misrouting, ambiguous handoffs, unreported blockages, stale version propagation, and dependency timing violations. The core issue lies in the system's inability to ensure that the right information reaches the right agent at the right time while maintaining a consistent task state. The remaining errors comprise output-interface misalignment (32%), state misjudgment (21%), and role-boundary violations (13%), reflecting failures in artifact delivery, task-state inference, and responsibility ownership, respectively. A detailed failure taxonomy is provided in Appendix G.

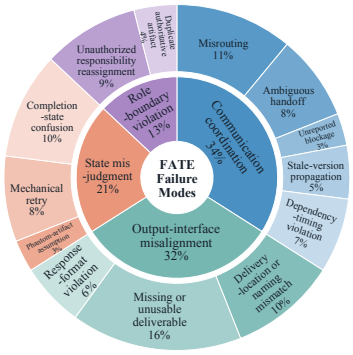


Figure 9: Distribution of failure.

4 Related Work

Professional and Occupational Agent Benchmarks. Recent benchmarks increasingly evaluate agents on realistic professional and workplace tasks (Vidgen et al., 2026; Mazeika et al., 2025; Lau et al., 2026; Drouin et al., 2024; Miserendino et al., 2025; Chan et al., 2025; Wijk et al., 2025; Zhu et al., 2026; Zhong et al., 2026). GDPval (Patwardhan et al., 2026), JobBench (Li et al., 2026c), Agents' Last Exam (Sun et al., 2026), OccuBench (Hu et al., 2026), and PolyWorkBench (Li et al., 2026a) broaden evaluation across occupations and professional tasks, while WorkArena++ (Boisvert et al., 2024), CRMArena (Huang et al., 2025), and TheAgentCompany (Xu et al., 2026b) instantiate realistic knowledge-work and enterprise environments. Most of these benchmarks, however, evaluate an individual agent or workflow rather than occupation-grounded teams with distinct roles, resources, permissions, and intermediate artifacts.

Long-horizon Agent. The Tool Decathlon (Li et al., 2025) evaluates long-horizon tool use across applications, OSWorld 2.0 (Yuan et al., 2026) targets real-world computer-use tasks, and Long-Horizon-Terminal-Bench (Li et al., 2026d) uses fine-grained graded subtasks for dense progress evaluation. Related benchmarks further study long-horizon professional workflows across office, desktop, SaaS, and cross-application environments (Wang et al., 2024; Shi et al., 2026; Li et al., 2026b; Seed, 2026; Zhou et al., 2026; Wang et al., 2026; Wu et al., 2026). FATE complements this direction by distributing complex work across occupational roles, turning task dependencies into requirements for communication, artifact handoff, and downstream coordination.

Agent Teams. MultiAgentBench (Zhu et al., 2025) studies collaboration and competition, SILO-BENCH (Zhang et al., 2026) studies distributed coordination under information silos, and AgentCollabBench (Mazumder et al., 2026) diagnoses failures in multi-hop information transfer. More closely related, EntCollabBench (Yu et al., 2026) evaluates role-specialized collaboration in permission-controlled enterprise workflows, while TeamBench (Kim et al., 2026) enforces Planner–Executor–Verifier separation through operating-system access controls. FATE extends this direction to broader occupation-grounded teams with role-specific resources and permissions, independently evaluable professional artifacts, and team-level deliverables across diverse domains and modalities.

5 Conclusion

We introduced FATE, a benchmark for evaluating agent teams on realistic professional tasks with distinct occupational responsibilities, resource access, and operational permissions. FATE comprises 112 tasks across 14 domains and 129 occupations, evaluating both members' intermediate artifacts and teams' final deliverables. Our evaluation of eight LLMs and comparison of four communication protocols reveal that reliable collaboration remains challenging, with limited overall gains at substantial computational cost. Collaboration gains vary with communication protocols and execution parallelism, while coordination and artifact delivery remain major bottlenecks. We hope FATE will support systematic evaluation of agent teams and guide the development of agent collaboration.

AI USE STATEMENT

During the construction and writing of this work, we used LLM to assist with several tasks, including grammar checking and text polishing. Besides, LLM is also used as components of the benchmark evaluation pipeline for predefined semantic and visual criteria, as described in the paper. In addition, AI assistants were used to improve the presentation of some tables and figures, while all experimental results and data were checked by the authors and remained unchanged.

ETHICS STATEMENT

FATE is constructed from publicly available and lawfully usable materials. We review all collected materials for privacy, licensing, safety, and potential biases before inclusion. Benchmark construction involves paid human annotators and domain experts. All annotators participated voluntarily under agreed annotation terms and compensation arrangements. Experts were compensated at a rate of USD 30 per hour. Their annotations underwent repeated cross-checking to reduce errors and potential biases. Because FATE covers professional domains including health, finance, legal work, and other consequential settings, benchmark outputs **should not be interpreted as professional advice or deployment-ready decisions**. FATE is intended **solely for research on the evaluation of agent team**. The dataset has been prepared for public release under the CC BY 4.0 license.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, evaluation protocol, and experimental settings throughout the main text and appendices. §2 defines the evaluation metrics, benchmark construction pipeline, and annotation procedures. §3 describes the evaluated models, agent harnesses and experimental settings. More detailed reproducibility information is provided in the appendix. Appendix A provides detailed task descriptions and documents the complete construction and quality-control pipeline of FATE. Appendix C specifies the agent team harness and communication protocols. Appendix D describes further implementation details such as harness versions. Besides, our FATE benchmark dataset has been prepared for public release under the CC BY 4.0 license. Together, these materials are intended to support the reproducibility of FATE.

REFERENCES

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 8038–8057. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.448. URL https://doi.org/10.18653/v1/2025.findings-naacl.448.

Anomaly. OpenCode: The open source AI coding agent. https://opencode.ai/, 2026. URL https://opencode.ai/.

Anthropic. Claude Code. https://code.claude.com/docs, 2025. URL https://code.claude.com/docs.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026a. Model announcement.

Anthropic. Orchestrate teams of Claude Code sessions. https://code.claude.com/docs/en/agent-teams, 2026b. Accessed: September 26, 2026.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak,

and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0b82662b6c32e887bb252a74d8cb2d5e-Abstract-Datasets_and_Benchmarks_Track.html.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. Mlebench: Evaluating machine learning agents on machine learning engineering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=EHg5GDnyq1.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=o1Et3MogPw.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025b.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-one: A generalist multi-agent system for solving complex tasks. *CoRR*, abs/2411.04468, 2024. doi: 10.48550/ARXIV.2411.04468. URL https://doi.org/10.48550/arXiv.2411.04468.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=VtmBAGCN7o.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=VtmBAGCN7o.

Xiaomeng Hu, Yinger Zhang, Fei Huang, Jianhong Tu, Yang Su, Lianghao Deng, Yuxuan Liu, Yantao Liu, Dayiheng Liu, and Tsung-Yi Ho. Occubench: Evaluating ai agents on real-world professional tasks via language world models. *arXiv preprint arXiv:2604.10866*, 2026.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. Crmarena: Understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 3830–3850. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.194. URL https://doi.org/10.18653/v1/2025.naacl-long.194.

Arpandeep Khatua, Hao Zhu, Peter Tran, Arya Prabhudesai, Frederic Sadrieh, Johann K. Lieberwirth, Xinkai Yu, Yicheng Fu, Michael J. Ryan, Jiaxin Pei, and Diyi Yang. Cooperbench: Why coding agents cannot be your teammates yet. *CoRR*, abs/2601.13295, 2026. doi: 10.48550/ARXIV.2601.13295. URL https://doi.org/10.48550/arXiv.2601.13295.

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.

Elaine Lau, Markus Dücker, Ronak Chaudhary, Hui Wen Goh, Rosemary Wei, Vaibhav Kumar, Saed Qunbar, Guram Gogia, Yi Liu, Scott Millslagle, Nasim Borazjanizadeh, Ulyana Tkachenko, Samuel Eshun Danquah, Collin Schweiker, Vijay Karumathil, Asrith Devalaraju, Varsha Sandadi, Haemi Nam, Punit Arani, Ray Epps, Abdullah Arif, Sahil Bhaiwala, Curtis G. Northcutt, Skyler Wang, Anish Athalye, Jonas W. Mueller, and Francisco Guzmán. Bankertoolbench: Evaluating AI agents in end-to-end investment banking workflows. *CoRR*, abs/2604.11304, 2026. doi: 10.48550/ARXIV.2604.11304. URL https://doi.org/10.48550/arXiv.2604.11304.

Hongliang Li, Yijin Liu, Zhiwei Zhang, Zihe Liu, Xinyue Lou, Jinan Xu, Fandong Meng, and Kaiyu Huang. Polyworkbench: Benchmarking multilingual long-horizon llm agents. *arXiv preprint arXiv:2607.06008*, 2026a.

Jinchao Li, Yunxin Li, Chenrui Zhao, Zhenran Xu, Baotian Hu, and Min Zhang. Windowsworld: A process-centric benchmark of autonomous GUI agents in professional cross-application environments. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 15262–15280. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.FINDINGS-ACL.750. URL https://doi.org/10.18653/v1/2026.findings-acl.750.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, Junteng Liu, Zhaochen Su, Yiyang Guo, Fan Zhou, Lueyang Zhang, Juan Michelini, Xingyao Wang, Xiang Yue, Shuyan Zhou, Graham Neubig, and Junxian He. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. *CoRR*, abs/2510.25726, 2025. doi: 10.48550/ARXIV.2510.25726. URL https://doi.org/10.48550/arXiv.2510.25726.

Yuetai Li, Yichen Feng, Zhangchen Xu, Zixian Ma, Kaiyuan Zheng, Fengqing Jiang, Xinghua Sun, Rulin Shao, Zichen Chen, Yue Huang, et al. Jobbench: Aligning agent work with human will. *arXiv preprint arXiv:2605.26329*, 2026c.

Zongxia Li, Zhongzhi Li, Yucheng Shi, Ruhan Wang, Junyao Yang, Zhichao Liu, Xiyang Wu, Anhao Li, Yue Yu, Ninghao Liu, Lichao Sun, Haotao Mi, and Leowei Liang. Long-horizon-terminal-bench: Testing the limits of agents on long-horizon terminal tasks with dense reward-based grading. *CoRR*, abs/2607.08964, 2026d. doi: 10.48550/ARXIV.2607.08964. URL https://doi.org/10.48550/arXiv.2607.08964.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, Oliver Ingebretsen, Aditya Mehta, Jean Mottola, John Teichmann, Kevin Yu, Zaina Shaik, Adam Khoja, Richard Ren, Jason Hausenloy, Long Phan, Ye Htet, Ankit Aich, Tahseen Rabbani, Vivswan Shah, Andriy Novykov, Felix Binder, Kirill Chugunov, Luis Ramirez, Matias Geralnik, Hernán Mesura, Dean Lee, Ed-Yeremai Hernandez-Cardona, Annette Diamond, Summer Yue, Alexandr Wang, Bing Liu, Ernesto Hernandez, and Dan Hendrycks. Remote labor index: Measuring AI automation of remote work. *CoRR*, abs/2510.26787, 2025. doi: 10.48550/ARXIV.2510.26787. URL https://doi.org/10.48550/arXiv.2510.26787.

Aritra Mazumder, Shubhashis Roy Dipta, Nusrat Jahan Lia, Tanzila Khan, Kainat Raisa Hossain, Nehaa Shri, Shubhrangshu Debsarkar, Humayra Tasnim, Gour Gupal Talukder Shawon, Debjoty Mitra, Sumaiya Ahmed Rani, Al Jami Islam Anik, and Al Nafeu Khan. Agentcollabbench: Diagnosing when good agents make bad collaborators. *arXiv preprint arXiv:2605.08647*, 2026.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/miserendino25a.html.

OpenAI. Codex CLI. https://github.com/openai/codex, 2025. URL https://github.com/openai/codex.

OpenAI. GPT-5.6 Sol model. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026. Model documentation.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Samuel Miserendino, Gildas Chabot, David Li, Patrick Chao, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks. In *International Conference on Learning Representations*, volume 2026, pp. 24005–24040, 2026.

Norman G Peterson, Michael D Mumford, Walter C Borman, P Richard Jeanneret, Edwin A Fleishman, Kerry Y Levin, Michael A Campion, Melinda S Mayfield, Frederick P Morgeson, Kenneth Pearlman, et al. Understanding work using the occupational information network (o* net): Implications for practice and research. *Personnel psychology*, 54(2):451–492, 2001.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15174–15186. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.810. URL https://doi.org/10.18653/v1/2024.acl-long.810.

Alireza Salemi, Mihir Parmar, Palash Goyal, Yiwen Song, Jinsung Yoon, Hamed Zamani, Hamid Palangi, and Tomas Pfister. Llm-based multi-agent blackboard system for information discovery in data science. *CoRR*, abs/2510.01285, 2025. doi: 10.48550/ARXIV.2510.01285. URL https://doi.org/10.48550/arXiv.2510.01285.

ByteDance Seed. Workflow-gym: Towards long-horizon evaluation of computer-use agentic tasks in real-world professional fields. *CoRR*, abs/2606.11042, 2026. doi: 10.48550/ARXIV.2606.11042. URL https://doi.org/10.48550/arXiv.2606.11042.

Kean Shi, Zihang Li, Tianyi Ma, Zengji Tu, Jialong Wu, Xinbo Xu, Qingyao Yang, Ruoyu Wu, Weichu Xie, Ming Wu, Jason Zeng, Michael Heinrich, Elvis Zhang, Liang Chen, Kuan Li, and Baobao Chang. Saas-bench: Can computer-use agents leverage real-world saas to solve professional workflows? *CoRR*, abs/2605.15777, 2026. doi: 10.48550/ARXIV.2605.15777. URL https://doi.org/10.48550/arXiv.2605.15777.

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

Kimi Team. Kimi K2.5: visual agentic intelligence. *CoRR*, abs/2602.02276, 2026. doi: 10.48550/ARXIV.2602.02276. URL https://doi.org/10.48550/arXiv.2602.02276.

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael Haines, Zach Richards, Chirag Mahapatra, Brendan Foody, and Osvald Nitski. Apex-agents. *CoRR*, abs/2601.14242, 2026. doi: 10.48550/ARXIV.2601.14242. URL https://doi.org/10.48550/arXiv.2601.14242.

Wenkai Wang, Tao Xiong, Jingchen Ni, Yunpeng Bao, Xiyun Li, Tianqi Liu, Hongcan Guo, Zilong Huang, and Shengyu Zhang. Deskcraft: Benchmarking desktop agents on professional workflows and human-in-the-loop collaboration. *CoRR*, abs/2606.03103, 2026. doi: 10.48550/ARXIV.2606.03103. URL https://doi.org/10.48550/arXiv.2606.03103.

Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. Officebench: Benchmarking language agents across multiple applications for office automation. *CoRR*, abs/2407.19056, 2024. doi: 10.48550/ARXIV.2407.19056. URL https://doi.org/10.48550/arXiv.2407.19056.

Daniel M Wegner. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*, pp. 185–208. Springer, 1987.

Hjalmar Wijk, Tao Roa Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Joshua Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Jun Koba Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. Re-bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wijk25a.html.

Jing Wu, Daphne Barretto, Yiye Chen, Nicholas Gyde, Yanan Jian, Yuhang He, and Vibhav Vineet. Os-marathon: Benchmarking computer-use agents on long-horizon repetitive tasks. *CoRR*, abs/2601.20650, 2026. doi: 10.48550/ARXIV.2601.20650. URL https://doi.org/10.48550/arXiv.2601.20650.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155, 2023. doi: 10.48550/ARXIV.2308.08155. URL https://doi.org/10.48550/arXiv.2308.08155.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Frank Fangzheng Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *Advances in Neural Information Processing Systems*, 38, 2026b.

Tao Yu, Hao Wang, Changyu Li, Shenghua Chai, Minghui Zhang, Zhongtian Luo, Yuxuan Zhou, Haopeng Jin, Zhaolu Kang, Jiabing Yang, et al. Beyond the all-in-one agent: Benchmarking role-specialized multi-agent collaboration in enterprise workflows. *arXiv preprint arXiv:2605.08761*, 2026.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lù, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. Osworld2.0: Benchmarking computer use agents on long-horizon real-world tasks. *CoRR*, abs/2606.29537, 2026. doi: 10.48550/ARXIV.2606.29537. URL https://doi.org/10.48550/arXiv.2606.29537.

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, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1354.

Jincheng Zhong, Weizhi Wang, Che Jiang, Kai Tian, Zhenzhao Yuan, Junlin Yang, Dianqiao Lei, and Kaiyan Zhang. Enterpriseclawbench: Benchmarking agents from real workplace sessions. *CoRR*, abs/2606.23654, 2026. doi: 10.48550/ARXIV.2606.23654. URL https://doi.org/10.48550/arXiv.2606.23654.

Jingbo Zhou, Yusai Zhao, Qi Bao, Jingjia Cao, Zhenghai Chen, Chang Gao, Kaiqi Guo, Muxin Guo, Mingxuan Li, Xinjiang Lu, Yanru Ma, Yixiong Xiao, Zenghui Zhang, Le Zhang, and Hua Wu. Omegause-officeval: Benchmarking LLM agents on long-horizon office-suite tasks with economic grounding. *CoRR*, abs/2607.27155, 2026. doi: 10.48550/ARXIV.2607.27155. URL https://doi.org/10.48550/arXiv.2607.27155.

Jian Zhu, Yuzheng Zhang, Zeyao Ma, Bohan Zhang, Armin Schoepf, Daniel Woloch, Peter Yiliu Wang, Guangyu Robert Yang, Samuel Jacob, Siddharth Nagisetty, Abhiram Chundru, Jean Lin, Spencer Mateega, and Jing Zhang. Spreadsheetbench 2: Evaluating agents on end-to-end business spreadsheet workflows. *CoRR*, abs/2606.29955, 2026. doi: 10.48550/ARXIV.2606.29955. URL https://doi.org/10.48550/arXiv.2606.29955.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, et al. 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.

A Data Construction

FATE converts real professional work into auditable and reproducible team work packages. Each benchmark task comprises five connected assets: source-grounded reference materials, occupationally grounded roles with distinct responsibilities, a reference collaboration workflow supported by industry practice or professional standards, role-level intermediate and team-level final deliverables, and executable evaluation criteria for these artifacts.

A.1 Task Sources and Material Provenance

We select candidate scenarios from real-world professional practice and retain them only when their objectives, input materials, constraints, and deliverables can be represented in a controlled digital environment. A candidate must involve substantive collaboration across roles and produce artifacts whose completion can be checked through file content, numerical values, semantic criteria, or visual properties. This excludes prompt-only and single-step tool-use tasks that do not require meaningful division of labor.

Reference materials are drawn from four categories of materials that are either publicly redistributable or otherwise lawfully usable, summarized in Table 4. These sources provide factual evidence, professional and process specifications, technical foundations, and realistic multimodal context.

Table 4: Primary sources of FATE reference materials. "Potential privacy" indicates whether a source category may require redaction or other privacy review before inclusion.

Source	Material forms	Public	Potential privacy	Role in task construction
Government, regulatory, and public-sector sources	Regulations, public filings, statistical databases, technical reports, and official disclosures	Yes	Possible public-person or case-specific information	Establish authoritative facts, compliance constraints, and operational rules
Companies, professional associations, and standards bodies	Guidelines, standards, templates, public project materials, and technical documentation	Yes	Possible author, contact, or project information	Ground workflows, role responsibilities, and deliverable requirements
Academic and open technical resources	Papers, open datasets, software examples, and reproducible experimental assets	Yes	Generally low	Provide data, methodological evidence, and reference results for research and engineering tasks
Public web and multimodal archives	Web archives, maps, satellite imagery, images, videos, and other public records	Yes	Possible incidental personal information	Preserve realistic multimodal evidence and contextual signals

We preserve original file formats, layouts, naming conventions, temporal information, and cross-file relationships whenever they are relevant to the professional task. A package may therefore contain PDFs, CSV/XLSX files, HTML/XML documents, images, videos, code, geospatial data, or other domain-specific formats rather than a flattened text representation. Before inclusion, we review licensing, terms of use, privacy, and confidentiality requirements. Personal or sensitive information is removed, generalized, or replaced without changing the factual relationships required by the task.

A.2 Occupational Taxonomy and Team Formation

We ground each role in a standard occupational taxonomy, primarily O*NET and its corresponding SOC entry. The role's responsibilities, professional judgments, accessible information, and expected artifacts must be supported by occupational descriptions or established industry practice.

We derive roles backward from the task objective and its professional workflow. We identify the substantive subgoals required for completion, such as evidence collection, modeling, compliance review, design, writing, quality assurance, or final integration, and map them to occupations with distinct expertise and work scopes. We retain a role only when it performs a non-trivial responsibility, produces an independent and traceable intermediate artifact, and contributes information or work needed by a downstream role or the final deliverables.

To avoid superficial multi-agent collaboration, we exclude candidate designs that exhibit any of the following patterns:

- **Mechanical file splitting:** dividing the same file by fields or pages among otherwise identical data-processing roles. If removing role titles leaves every role performing the same extraction or aggregation operation, the division of labor is not occupationally meaningful.
- **Unsupported role titles:** introducing roles that cannot be mapped to a standard occupational taxonomy and lack real-world responsibilities.
- **Unreproducible core activities:** centering a task on physical work, unavailable proprietary equipment, or other core activities that cannot be reasonably represented in the controlled digital environment.
- **Unverifiable team outputs:** requiring only a short final answer without role-level artifacts or member-level evaluation.

A.3 Data Annotation Instructions

1. **Realism**: Tasks, reference materials, professional roles, workflows, and final deliverables should have identifiable counterparts in real-world industries and should avoid unnecessary fabrication.
2. **Diversity**: Tasks should reflect collaboration across different professions.
3. **Verifiability**: Tasks should have clearly defined intermediate outputs, final deliverables, and evaluation criteria, such that task completion does not depend solely on an overall subjective assessment.
4. **Long-Horizon Collaboration**: Tasks should be of sufficient scale to support parallel or interleaved work across multiple roles.

Data and Sources

Data Sources. Tasks and reference data should be derived from actual work in real-world settings. Annotators should consult materials that are publicly redistributable or otherwise lawfully usable and make use of artifacts produced, used, or delivered by human teams in real-world projects. The data used should comply with applicable conditions governing use, reproduction, and redistribution. Before data are included in the dataset, relevant copyright and licensing conditions, privacy requirements, and related restrictions should be reviewed.
Tasks should preferably exhibit the following characteristics:

- **Diverse data types**: Materials include information of different types, from different sources, or covering different time periods.
- **Diverse actions**: Tasks naturally require multiple types of professional operations, such as evidence retrieval and verification, structured data transformation, quantitative analysis, constraint-based planning, and domain-specific artifact construction.

Reference Task Selection Process.

- Identify a specific domain and a specific real-world scenario. Search the public Internet for tasks that genuinely occur in that scenario. Prefer cases for which both the reference materials required to perform the task and a human-produced deliverable are available. For example, in an aviation accident investigation scenario, it is preferable to identify a real accident, its corresponding accident investigation report, and the real-world reference materials required to produce that report.
- Adapt the original task into a benchmark task and organize the reference materials into a task file directory. Use human-produced deliverables to identify key task requirements and verifiable elements, and design evaluation criteria accordingly. Avoid treating incidental structure, wording choices, or nonessential details of a particular reference deliverable as the sole correct standard.

Professional Roles and Workflow

Role Definition and Occupational Realism. Role names should correspond to real-world occupations and should preferably be mapped to standardized occupational classifications such as O*NET. Domain-specific or region-specific occupational classifications may be used when necessary.

Role Division. The division of professional responsibilities should be grounded in standard workflows or project management practices used in real-world industries and should not be fabricated. Different professions should exhibit genuine differences in domain expertise and scope of responsibility.

Deliverables and Evaluation

Deliverables. Each task should include both intermediate outputs and a final deliverable. Each role on the team should leave an independent, traceable, and verifiable work record or intermediate result that makes clear what work the role actually performed. The final deliverable should correspond to an artifact that would actually be delivered in real-world professional practice.

Evaluation. Evaluation comprises Member Evaluation (ME; `member_eval`) for role-level intermediate outputs and Final Evaluation (FE; `final_eval`) for the team's final deliverable.

Evaluation Methods. Evaluation is divided into Rule-based evaluation and LLM-based evaluation.
Rule-based evaluation uses fixed programmatic logic to assess factual or structural conditions that can be checked through code, such as format and quantity.
LLM-based evaluation supplements semantic relationships that are difficult to capture with rules. Evaluation prompts should focus on observable judgments, such as "whether an item exists," "whether certain information is mentioned," or "whether a particular step was performed." Subjective or highly knowledge-dependent judgments, such as "whether something is excellent," "whether something is aesthetically pleasing," or "whether something complies with local law," should be avoided.

Evaluation Composition. The final evaluation may include evaluation criteria implemented through both Rule-based and LLM-based methods.

A.4 Quality-Control Procedure

After annotation, each task undergoes automated validation, expert verification, and evaluator testing. Automated checks verify the required task structure and artifacts, expert review examines the task definition, role assignments, and deliverable requirements, and evaluator testing checks that the associated evaluation criteria operate as intended. Issues identified during these checks are revised before inclusion in the benchmark.

B Benchmark Statistics

B.1 Benchmark Comparison

We compare FATE with existing single-agent and multi-agent benchmarks along several dimensions, including occupational coverage, communication protocols, member-level evaluation, role isolation, task scale, and output artifact types. As shown in Table 5, FATE supports a broad range of occupational domains and subdomains while explicitly evaluating both team members and role-isolated interactions across diverse deliverable formats.

B.2 Task Resources

FATE contains 112 tasks, 683 roles, 1,659 input resources, 1,065 artifacts and deliverables, 890 ME entries, and 1,410 FE entries. Table 6 further reports task counts and the mean numbers of roles, inputs, deliverables, ME, and FE per task for each domain. On average, a task contains 6.10 roles, 14.81 input resources, 9.51 artifacts and deliverables, 7.95 ME, and 12.59 FE.

Task-resource requirements vary substantially across domains, reflecting differences in material density, team organization, and deliverable forms in real professional work. For example, Research tasks contain an average of 12.00 roles and 39.00 input resources, while Finance and Restaurant tasks contain relatively large numbers of deliverables and FE entries, respectively.

Table 5: Comparison of **FATE** with existing agent benchmarks. **Setting** denotes a single agent or an agent team. **Dom.** and **Subdom.** report the numbers of covered occupational domains and occupational subdomains. **Comm.** indicates the number of communication protocols supported. **ME** indicates whether individual team members' performance is explicitly evaluated. **Role Iso.** indicates whether agents have isolated permissions and access controls, and **Tasks** reports the number of benchmark tasks, and **Products** summarizes the evaluated output artifact types.

Benchmark	Setting	Dom.	Subdom.	Comm.	ME	Role Iso.	Tasks	Products
Single-Agent Benchmarks								
JobBench (Li et al., 2026c)	Single	10	35	✗	✗	✗	130	TXT, Excel, PDF, PNG
ALE (Sun et al., 2026)	Single	13	55	✗	✗	✗	1,490	TXT, PDF, PNG, Video, 3D, Code
PolyWorkBench (Li et al., 2026a)	Single	5	-	✗	✗	✗	67	TXT, Excel
OccuBench (Hu et al., 2026)	Single	10	65	✗	✗	✗	382	TXT
GDPval (Patwardhan et al., 2026)	Single	9	44	✗	✗	✗	1,320	TXT, PDF, PNG, Excel, PPT, Video
Multi-Agent Benchmarks								
MultiAgentBench (Zhu et al., 2025)	Team	6	-	✓(4)	✗	✗	500	TXT, Code
TheAgentCompany (Xu et al., 2026b)	Team	1	7	✓(1)	✗	✗	175	TXT, Code
EntCollabBench (Yu et al., 2026)	Team	6	11	✓(1)	✗	✓	300	Code
TeamBench (Kim et al., 2026)	Team	3	19	✓(1)	✗	✓	931	TXT, Code
SILO-BENCH (Zhang et al., 2026)	Team	1	-	✓(3)	✓	✓	30	TXT
AgentCollabBench (Mazumder et al., 2026)	Team	3	-	✓(5)	✗	✓	900	TXT
FATE (Ours)	**Team**	**14**	**41**	✓(4)	✓	✓	112	TXT, Excel, PDF, PPT, PNG, Video, 3D, HTML

Table 6: Basic statistics of FATE benchmark tasks by domain. **Tasks** is the number of tasks; **Roles** is the mean number of entries in teams per task; **Inputs** is the mean number of input resources per task; **Deliv.** is the mean number of unique intermediate artifacts and final deliverables per task; **ME** is the mean number of ME entries per task; and **FE** is the mean number of FE entries per task.

Domain	Tasks	Roles	Inputs	Deliv.	ME	FE
Building	7	5.57	5.43	7.71	5.57	11.00
Business	13	5.15	13.54	6.08	5.15	11.77
CFD	7	5.43	12.00	13.71	11.43	8.29
Education	12	6.42	20.33	6.42	5.83	9.50
Finance	11	6.09	12.27	15.91	6.55	1.55
Health	3	5.33	22.33	12.67	10.00	10.00
Industrial	4	5.00	8.00	9.00	5.00	22.00
Investigation	4	5.75	12.75	10.25	9.75	10.25
Legal	10	6.10	12.30	8.30	5.20	12.40
Media	3	5.67	12.33	9.00	10.00	7.00
Research	6	12.00	39.00	12.33	11.33	13.17
Restaurant	10	6.60	10.90	7.70	6.60	30.00
Social	10	3.90	13.30	8.40	10.20	15.00
Tourism	12	6.75	16.33	10.33	12.92	13.17
Avg.	8	6.10	14.81	9.51	7.95	12.59

B.3 Occupational Coverage

Table 7 lists the O*NET-SOC occupations represented in the FATE benchmark tasks across major, minor, broad, and detailed occupational levels.

Table 7: O*NET-SOC occupations represented in the FATE benchmark tasks.

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
11-0000					Management Occupations
	11-1000				Top Executives
		11-1020			General and Operations Managers
			11-1021	11-1021.00	General and Operations Managers
	11-2000				Advertising, Marketing, Promotions, Public Relations, and Sales Managers
		11-2010			Advertising and Promotions Managers
			11-2011	11-2011.00	Advertising and Promotions Managers
		11-2020			Marketing and Sales Managers
			11-2021	11-2021.00	Marketing Managers
	11-3000				Operations Specialties Managers
		11-3070			Transportation, Storage, and Distribution Managers

Continued on next page

Table 7 continued from previous page

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
			11-3071	11-3071.00	Transportation, Storage, and Distribution Managers
			11-3071	11-3071.04	Supply Chain Managers
	11-9000				Other Management Occupations
		11-9030			Education and Childcare Administrators
			11-9032	11-9032.00	Education Administrators, Kindergarten through Secondary
		11-9040			Architectural and Engineering Managers
			11-9041	11-9041.00	Architectural and Engineering Managers
		11-9050			Food Service Managers
			11-9051	11-9051.00	Food Service Managers
		11-9080			Lodging Managers
			11-9081	11-9081.00	Lodging Managers
		11-9140			Property, Real Estate, and Community Association Managers
			11-9141	11-9141.00	Property, Real Estate, and Community Association Managers
		11-9160			Emergency Management Directors
			11-9161	11-9161.00	Emergency Management Directors
		11-9190			Miscellaneous Managers
			11-9199	11-9199.02	Compliance Managers
13-0000					Business and Financial Operations Occupations
	13-1000				Business Operations Specialists
		13-1020			Buyers and Purchasing Agents
			13-1023	13-1023.00	Purchasing Agents, Except Wholesale, Retail, and Farm Products
		13-1030			Claims Adjusters, Appraisers, Examiners, and Investigators
			13-1031	13-1031.00	Claims Adjusters, Examiners, and Investigators
		13-1040			Compliance Officers
			13-1041	13-1041.00	Compliance Officers
			13-1041	13-1041.01	Environmental Compliance Inspectors
			13-1041	13-1041.07	Regulatory Affairs Specialists
		13-1080			Logisticians and Project Management Specialists
			13-1081	13-1081.00	Logisticians
			13-1082	13-1082.00	Project Management Specialists
		13-1110			Management Analysts
			13-1111	13-1111.00	Management Analysts
		13-1120			Meeting, Convention, and Event Planners
			13-1121	13-1121.00	Meeting, Convention, and Event Planners
		13-1150			Training and Development Specialists
			13-1151	13-1151.00	Training and Development Specialists
		13-1160			Market Research Analysts and Marketing Specialists
			13-1161	13-1161.00	Market Research Analysts and Marketing Specialists
	13-2000				Financial Specialists
		13-2010			Accountants and Auditors
			13-2011	13-2011.00	Accountants and Auditors
		13-2050			Financial Analysts and Advisors
			13-2051	13-2051.00	Financial and Investment Analysts
		13-2060			Financial Examiners
			13-2061	13-2061.00	Financial Examiners
15-0000					Computer and Mathematical Occupations
	15-1200				Computer Occupations
		15-1210			Computer and Information Analysts
			15-1211	15-1211.00	Computer Systems Analysts
			15-1212	15-1212.00	Information Security Analysts
		15-1240			Database and Network Administrators and Architects
			15-1243	15-1243.00	Database Architects
		15-1250			Software and Web Developers, Programmers, and Testers
			15-1252	15-1252.00	Software Developers
			15-1253	15-1253.00	Software QA Analysts and Testers
			15-1254	15-1254.00	Web Developers
			15-1255	15-1255.00	Web and Digital Interface Designers
		15-1290			Miscellaneous Computer Occupations
			15-1299	15-1299.02	Geographic Information Systems Technologists and Technicians
			15-1299	15-1299.03	Document Management Specialists
	15-2000				Mathematical Science Occupations
		15-2010			Actuaries
			15-2011	15-2011.00	Actuaries
		15-2030			Operations Research Analysts
			15-2031	15-2031.00	Operations Research Analysts
		15-2040			Statisticians
			15-2041	15-2041.00	Statisticians
			15-2041	15-2041.01	Biostatisticians
		15-2050			Data Scientists

Continued on next page

Table 7 continued from previous page

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
			15-2051	15-2051.00	Data Scientists
			15-2051	15-2051.02	Clinical Data Managers
17-0000					Architecture and Engineering Occupations
	17-1000				Architects, Surveyors, and Cartographers
		17-1010			Architects, Except Naval
			17-1011	17-1011.00	Architects, Except Landscape and Naval
		17-1020			Surveyors, Cartographers, and Photogrammetrists
			17-1021	17-1021.00	Cartographers and Photogrammetrists
			17-1022	17-1022.00	Surveyors
	17-2000				Engineers
		17-2010			Aerospace Engineers
			17-2011	17-2011.00	Aerospace Engineers
		17-2030			Bioengineers and Biomedical Engineers
			17-2031	17-2031.00	Bioengineers and Biomedical Engineers
		17-2040			Chemical Engineers
			17-2041	17-2041.00	Chemical Engineers
		17-2050			Civil Engineers
			17-2051	17-2051.00	Civil Engineers
		17-2060			Computer Hardware Engineers
			17-2061	17-2061.00	Computer Hardware Engineers
		17-2070			Electrical and Electronics Engineers
			17-2071	17-2071.00	Electrical Engineers
			17-2072	17-2072.00	Electronics Engineers, Except Computer
		17-2110			Industrial Engineers, Including Health and Safety
			17-2111	17-2111.00	Health and Safety Engineers
			17-2111	17-2111.02	Fire-Prevention and Protection Engineers
			17-2112	17-2112.00	Industrial Engineers
			17-2112	17-2112.02	Validation Engineers
			17-2112	17-2112.03	Manufacturing Engineers
		17-2120			Marine Engineers and Naval Architects
			17-2121	17-2121.00	Marine Engineers and Naval Architects
		17-2130			Materials Engineers
			17-2131	17-2131.00	Materials Engineers
		17-2140			Mechanical Engineers
			17-2141	17-2141.00	Mechanical Engineers
		17-2190			Miscellaneous Engineers
			17-2199	17-2199.03	Energy Engineers, Except Wind and Solar
			17-2199	17-2199.11	Mechatronics Engineers
	17-3000				Drafters, Engineering Technicians, and Mapping Technicians
		17-3010			Drafters
			17-3011	17-3011.00	Architectural and Civil Drafters
			17-3012	17-3012.00	Electrical and Electronics Drafters
		17-3020			Engineering Technologists and Technicians, Except Drafters
			17-3027	17-3027.00	Mechanical Engineering Technologists and Technicians
		17-3030			Surveying and Mapping Technicians
			17-3031	17-3031.00	Surveying and Mapping Technicians
19-0000					Life, Physical, and Social Science Occupations
	19-1000				Life Scientists
		19-1020			Biological Scientists
			19-1023	19-1023.00	Zoologists and Wildlife Biologists
			19-1029	19-1029.01	Bioinformatics Scientists
		19-1040			Medical Scientists
			19-1041	19-1041.00	Epidemiologists
	19-2000				Physical Scientists
		19-2010			Astronomers and Physicists
			19-2011	19-2011.00	Astronomers
			19-2012	19-2012.00	Physicists
		19-2020			Atmospheric and Space Scientists
			19-2021	19-2021.00	Atmospheric and Space Scientists
		19-2040			Environmental Scientists and Geoscientists
			19-2041	19-2041.00	Environmental Scientists and Specialists, Including Health
	19-3000				Social Scientists and Related Workers
		19-3020			Survey Researchers
			19-3022	19-3022.00	Survey Researchers
		19-3030			Psychologists
			19-3032	19-3032.00	Industrial-Organizational Psychologists
		19-3040			Sociologists
			19-3041	19-3041.00	Sociologists
		19-3050			Urban and Regional Planners
			19-3051	19-3051.00	Urban and Regional Planners
		19-3090			Miscellaneous Social Scientists and Related Workers
			19-3092	19-3092.00	Geographers
			19-3094	19-3094.00	Political Scientists
			19-3099	19-3099.01	Transportation Planners
	19-4000				Life, Physical, and Social Science Technicians

Continued on next page

Table 7 continued from previous page

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
		19-4020			Biological Technicians
			19-4021	19-4021.00	Biological Technicians
		19-4090			Miscellaneous Life, Physical, and Social Science Technicians
			19-4099	19-4099.00	Life, Physical, and Social Science Technicians, All Other
			19-4099	19-4099.01	Quality Control Analysts
	19-5000				Occupational Health and Safety Specialists and Technicians
		19-5010			Occupational Health and Safety Specialists and Technicians
			19-5011	19-5011.00	Occupational Health and Safety Specialists
21-0000					Community and Social Service Occupations
	21-1000				Counselors, Social Workers, and Other Community and Social Service Specialists
		21-1010			Counselors
			21-1015	21-1015.00	Rehabilitation Counselors
23-0000					Legal Occupations
	23-1000				Lawyers, Judges, and Related Workers
		23-1010			Lawyers and Judicial Law Clerks
			23-1011	23-1011.00	Lawyers
			23-1012	23-1012.00	Judicial Law Clerks
	23-2000				Legal Support Workers
		23-2010			Paralegals and Legal Assistants
			23-2011	23-2011.00	Paralegals and Legal Assistants
25-0000					Educational Instruction and Library Occupations
	25-1000				Postsecondary Teachers
		25-1120			Arts, Communications, History, and Humanities Teachers, Postsecondary
			25-1124	25-1124.00	History Teachers, Postsecondary
	25-2000				Preschool, Elementary, Middle, Secondary, and Special Education Teachers
		25-2020			Elementary and Middle School Teachers
			25-2022	25-2022.00	Middle School Teachers, Except Special and Career/Technical Education
		25-2030			Secondary School Teachers
			25-2031	25-2031.00	Secondary School Teachers, Except Special and Career/Technical Education
		25-2050			Special Education Teachers
			25-2057	25-2057.00	Special Education Teachers, Middle School
	25-3000				Other Teachers and Instructors
		25-3010			Adult Basic Education, Adult Secondary Education, and English as a Second Language Instructors
			25-3011	25-3011.00	Adult Basic Education, Adult Secondary Education, and English as a Second Language Instructors
	25-4000				Librarians, Curators, and Archivists
		25-4010			Archivists, Curators, and Museum Technicians
			25-4011	25-4011.00	Archivists
			25-4012	25-4012.00	Curators
		25-4020			Librarians and Media Collections Specialists
			25-4022	25-4022.00	Librarians and Media Collections Specialists
	25-9000				Other Educational Instruction and Library Occupations
		25-9030			Instructional Coordinators
			25-9031	25-9031.00	Instructional Coordinators
27-0000					Arts, Design, Entertainment, Sports, and Media Occupations
	27-1000				Art and Design Workers
		27-1020			Designers
			27-1024	27-1024.00	Graphic Designers
			27-1025	27-1025.00	Interior Designers
	27-2000				Entertainers and Performers, Sports and Related Workers
		27-2010			Actors, Producers, and Directors
			27-2012	27-2012.00	Producers and Directors
	27-3000				Media and Communication Workers
		27-3020			News Analysts, Reporters and Journalists
			27-3023	27-3023.00	News Analysts, Reporters, and Journalists
		27-3030			Public Relations Specialists
			27-3031	27-3031.00	Public Relations Specialists
		27-3040			Writers and Editors
			27-3041	27-3041.00	Editors

Continued on next page

Table 7 continued from previous page

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
			27-3042	27-3042.00	Technical Writers
			27-3043	27-3043.00	Writers and Authors
		27-3090			Miscellaneous Media and Communication Workers
			27-3091	27-3091.00	Interpreters and Translators
			27-3092	27-3092.00	Court Reporters and Simultaneous Captioners
	27-4000				Media and Communication Equipment Workers
		27-4010			Broadcast, Sound, and Lighting Technicians
			27-4011	27-4011.00	Audio and Video Technicians
			27-4013	27-4013.00	Radio Operators
			27-4014	27-4014.00	Sound Engineering Technicians
		27-4020			Photographers
			27-4021	27-4021.00	Photographers
		27-4030			Television, Video, and Film Camera Operators and Editors
			27-4032	27-4032.00	Film and Video Editors
29-0000					Healthcare Practitioners and Technical Occupations
	29-1000				Healthcare Diagnosing or Treating Practitioners
		29-1030			Dietitians and Nutritionists
			29-1031	29-1031.00	Dietitians and Nutritionists
		29-1210			Physicians
			29-1229	29-1229.05	Preventive Medicine Physicians
35-0000					Food Preparation and Serving Related Occupations
	35-1000				Supervisors of Food Preparation and Serving Workers
		35-1010			Supervisors of Food Preparation and Serving Workers
			35-1011	35-1011.00	Chefs and Head Cooks
39-0000					Personal Care and Service Occupations
	39-7000				Tour and Travel Guides
		39-7010			Tour and Travel Guides
			39-7011	39-7011.00	Tour Guides and Escorts
			39-7012	39-7012.00	Travel Guides
	39-9000				Other Personal Care and Service Workers
		39-9030			Recreation and Fitness Workers
			39-9032	39-9032.00	Recreation Workers
41-0000					Sales and Related Occupations
	41-3000				Sales Representatives, Services
		41-3040			Travel Agents
			41-3041	41-3041.00	Travel Agents
43-0000					Office and Administrative Support Occupations
	43-4000				Information and Record Clerks
		43-4030			Court, Municipal, and License Clerks
			43-4031	43-4031.00	Court, Municipal, and License Clerks
		43-4050			Customer Service Representatives
			43-4051	43-4051.00	Customer Service Representatives
	43-6000				Secretaries and Administrative Assistants
		43-6010			Secretaries and Administrative Assistants
			43-6012	43-6012.00	Legal Secretaries and Administrative Assistants
	43-9000				Other Office and Administrative Support Workers
		43-9110			Statistical Assistants
			43-9111	43-9111.00	Statistical Assistants
47-0000					Construction and Extraction Occupations
	47-4000				Other Construction and Related Workers
		47-4010			Construction and Building Inspectors
			47-4011	47-4011.00	Construction and Building Inspectors
49-0000					Installation, Maintenance, and Repair Occupations
	49-3000				Vehicle and Mobile Equipment Mechanics, Installers, and Repairers
		49-3010			Aircraft Mechanics and Service Technicians
			49-3011	49-3011.00	Aircraft Mechanics and Service Technicians
51-0000					Production Occupations
	51-9000				Other Production Occupations
		51-9060			Inspectors, Testers, Sorters, Samplers, and Weighers
			51-9061	51-9061.00	Inspectors, Testers, Sorters, Samplers, and Weighers
53-0000					Transportation and Material Moving Occupations
	53-2000				Air Transportation Workers
		53-2010			Aircraft Pilots and Flight Engineers
			53-2011	53-2011.00	Airline Pilots, Copilots, and Flight Engineers
	53-4000				Rail Transportation Workers
		53-4010			Locomotive Engineers and Operators
			53-4011	53-4011.00	Locomotive Engineers
	53-5000				Water Transportation Workers

Continued on next page

Table 7 continued from previous page

Major	Minor	Broad	Occupation	O*NET-SOC	O*NET-SOC Title
		53-5020			Ship and Boat Captains and Operators
			53-5021	53-5021.00	Captains, Mates, and Pilots of Water Vessels
	53-6000				Other Transportation Workers
		53-6050			Transportation Inspectors
			53-6051	53-6051.01	Aviation Inspectors

B.4 Input and Output Modalities

Tables 8 and 9 report the modality distributions of input resources and task-level final deliverables, respectively. Inputs span PDFs, CSV files, Markdown, HTML, JSON, text, images, spreadsheets, geospatial data, code, and domain-specific formats, with PDF, CSV, Markdown, and HTML as the most common modalities. Outputs include structured, document, visual, spatial, and multimedia artifacts such as CSV, HTML, JSON, Markdown, PNG, PDF, XLSX, GeoJSON, IFC, and MP4. These distributions show that FATE requires information processing, professional production, and cross-role integration over multimodal evidence and diverse file deliverables rather than reducing collaboration to text-only question answering.

Table 8: Modalities of input resources in FATE.

Modality	Count	Percentage
pdf	349	21.04%
csv	261	15.73%
md	234	14.10%
html	169	10.19%
json	154	9.28%
txt	141	8.50%
jpg	86	5.18%
xlsx	40	2.41%
png	39	2.35%
htm	24	1.45%
jsonl	24	1.45%
geojson	23	1.39%
py	21	1.27%
dat	13	0.78%
tar.gz	11	0.66%
ifc	10	0.60%
xml	8	0.48%
tsv	7	0.42%
xpt	7	0.42%
vtt	6	0.36%
zip	5	0.30%
kicad_pcb	4	0.24%
mp4	4	0.24%
gz	3	0.18%
nmf	3	0.18%
srt	2	0.12%
xls	2	0.12%
yaml	2	0.12%
db	1	0.06%
docx	1	0.06%
epub	1	0.06%
f	1	0.06%
pptx	1	0.06%
stl	1	0.06%
wav	1	0.06%
Total	**1659**	**100.00%**

Table 9: Modalities of task-level final deliverables in FATE.

Modality	**Count**	**Percentage**
csv	117	25.38%
html	80	17.35%
json	58	12.58%
md	42	9.11%
png	27	5.86%
pdf	25	5.42%
xlsx	21	4.56%
geojson	14	3.04%
jsonl	14	3.04%
ifc	9	1.95%
mp4	9	1.95%
tar.gz	7	1.52%
zip	7	1.52%
kicad_pcb	4	0.87%
py	4	0.87%
sh	4	0.87%
docx	3	0.65%
ics	3	0.65%
svg	3	0.65%
txt	3	0.65%
r	2	0.43%
vtt	2	0.43%
yaml	2	0.43%
pptx	1	0.22%
Total	**461**	**100.00%**

C COMMUNICATION PROTOCOLS

We define a *communication event* as one agent transmitting a text message of at most 60,000 characters together with up to 10 attachments through the active communication protocol. For directed communication, sending the same content to multiple recipients counts as one communication event per recipient. Communication frequency (Freq.) is the average number of communication events per task. Cross-agent information exchange occurs through explicitly communicated messages and attachments, while workspace isolation is preserved throughout execution. Each team run uses one communication protocol, fixed for the duration of the run.

In multi-agent execution, READY is a coordination action indicating that an agent has completed its current team responsibilities. A team task is considered complete when all agents are simultaneously in the READY state. Protocol-specific communication may subsequently return a ready agent to active collaboration.

Point-to-Point Communication. Each agent has a private inbox and may contact one or more specific teammates by sending a text message with optional attachments. Each delivery is placed in the corresponding recipient's inbox, and the recipient explicitly reads the inbox before incorporating the received information into its context. Replies follow the same mechanism. Communication relationships are formed dynamically during execution according to the recipients selected by the sender. A newly delivered message returns a ready recipient to active collaboration.

Shared Message Communication. Agents publish requests and optional attachments to a shared message pool. Other agents inspect pending requests and may *claim* one for execution. Each agent may hold at most one active claim at a time, with eligible claimants drawn from teammates distinct from the publisher. After completing a claimed request, the agent returns its answer and files to the original publisher. The shared pool retains the request, claim, and result state throughout this process. Shared requests are presented anonymously in the public pool. Newly published requests become available when agents next inspect the shared pool.

Central Communication. This protocol introduces an additional *relay agent* through which worker-to-worker consultations are routed. Worker agents send consultation requests to the relay agent, which receives the original task, role descriptions, task assignments, and consultation text and uses this context to determine the appropriate routing. The relay agent uses this coordination context for routing decisions, and worker agents perform task execution in their own workspaces. Responses are routed through the relay agent to the appropriate worker.

Interleaved Communication. All worker agents proceed within their own task workflows. When an agent encounters a question requiring another agent's expertise, it explicitly selects that agent and sends a consultation request with optional attachments. The requester pauses the corresponding work until the consultation result is returned, then incorporates the returned answer and files and continues its task. The consulted agent preserves its primary workflow while handling the consultation in a separate working copy derived from its current context. The consulted agent may answer directly or consult a third agent in the same manner, allowing a consultation chain to form. Results propagate back along the consultation chain to the original requester. In our experiments, consultation chains have a maximum recursion depth of 2.

D IMPLEMENTATION DETAILS

D.1 RUNTIME AND AGENT CONTEXT

Each agent runs in an independent harness session with a sandboxed private workspace. The role configuration $\mathcal{R}_i$ specifies the agent's occupational responsibilities j_i, accessible resources $\mathcal{I}_i$, and permitted operations $\mathcal{O}_i$. In our implementation, $\mathcal{O}_i$ is instantiated by the workspace-scoped operations exposed by the agent's harness, with web search and subagent delegation removed from the default harness capabilities.

Table 10: Harness versions and runtime integrations used in our experiments.

Harness	Harness version	Integration
Codex CLI	0.147.0	`openai-codex` 0.147.0
Claude Code	2.1.149	`claude-agent-sdk` 0.2.149
OpenCode	1.18.27	HTTP server 1.18.27

For text-only models, visual perception is provided through a black-box MCP tool backed by GPT-5.6 Terra. The perception backend is presented to the agent only through this tool interface. Token usage internal to the MCP perception backend is excluded from the reported token usage.

At initialization, each agent receives the overall task description, its own subtask and file specification, and the names, roles, and responsibility summaries of its teammates. Private task context and workspace contents remain isolated across agents. The planner $\mathcal{P}$ is invoked once at initialization.

Tool-use actions $\mathcal{A}_i^{\text{Tool}}$ are provided by the agent's harness and task environment, while communication actions $\mathcal{A}_i^{\text{Comm}}$ follow one of the communication protocols described in Appendix C. Each sandbox is provisioned before evaluation with the software dependencies required by its task.

D.2 Single-agent Setting

FATE supports both multi-agent and single-agent execution of the same benchmark task. In single-agent execution, one agent receives the complete task specification and all resources required for completion, organizes the full task within a single harness session, and produces the same task-level final deliverables as the multi-agent team. Single-agent runs are evaluated with Final Evaluation (FE), and comparisons between single-agent and multi-agent execution use FE as the common task-level metric.

E TASK EXAMPLES

Project Olympus Cooling-Resilience Screening

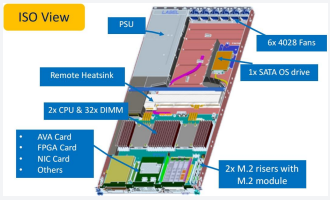
Five-Occupation Team

◆ **CFD** · Server Thermal Management · Reliability Engineering

TASK DESCRIPTION

Assess the cooling resilience of a Project Olympus US1-EPYC server under normal six-fan operation and a selected single-fan zero-flow failure at `fan_pos2`. The team integrates hardware and electrical evidence, fresh OpenFOAM airflow and passive-thermal screening, and reliability analysis to produce an auditable management disposition while preserving unresolved evidence and qualification limits.

SYSTEM CONTEXT AND INPUTS



US1-EPYC system and six-fan cooling layout.

INPUT MATERIALS

Platform documents (.html, .pdf). Open Compute reference, US1-EPYC brief, and 1U mechanical specification; define the platform, cooling architecture, and component layout.

Motherboard documents (.pdf). AMD EPYC motherboard specification and US1-EPYC schematic; provide BMC, PWM/TACH, fan-channel, and connectivity evidence.

Frozen CFD mesh (.tar.gz). provided-mesh.tar.gz; supplied server mesh for fresh normal/failure OpenFOAM runs without remeshing.

FMECA guidance (.pdf). GSFC-HDBK-8004; failure-mode, effect, and reliability-assessment guidance.

MULTI-OCCUPATION TEAM

Computer Hardware Engineers
*O*NET 17-2061.00*
Establish the hardware, cooling, workload, and system-assumption basis while distinguishing source-backed facts from benchmark assumptions.

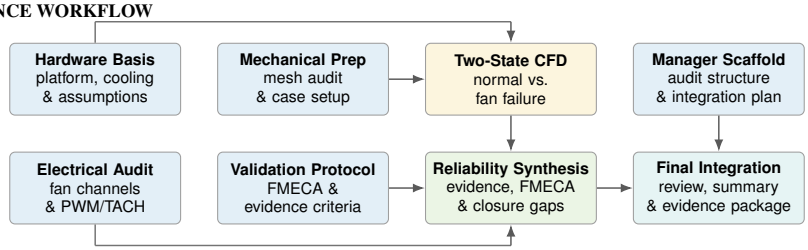
Electrical Engineers
*O*NET 17-2071.00*
Map the six fan channels and BMC/PWM/TACH connectivity; assess fan-control capability and whether signals can detect airflow loss.

Mechanical Engineers
*O*NET 17-2141.00*
Audit the frozen mesh, prepare normal/failure OpenFOAM cases, execute fresh simulations, and extract airflow and passive-air thermal QOIs.

Validation Engineers
*O*NET 17-2112.02*
Integrate specialist handoffs into FMECA and reliability assessment while preserving evidence status, uncertainty, and remaining gaps.

Architectural and Engineering Managers
*O*NET 11-9041.00*
Audit cross-role consistency, reconcile scope and evidence, and integrate the final review, evidence package, and disposition.

REFERENCE WORKFLOW



Specialist evidence is produced in parallel and transferred through explicit handoffs; the Engineering Manager reconciles the resulting evidence during final integration.

REQUIRED TEAM OUTPUTS

Integrated Engineering Review
`Olympus_Cooling_Resilience_Review.html`
Evidence-backed assessment, limitations, and final disposition.

Executive Comparison Figure
`Olympus_Cooling_Resilience_Summary.png`
Normal-versus-failure airflow and passive-air thermal QOI comparison.

Evidence Package
`Olympus_Cooling_Resilience_Evidence.tar.gz`
Specialist artifacts, CFD evidence, reliability outputs, and provenance.

EVALUATION EXAMPLES

CFD execution authenticity. Verify supplied-mesh identity, real OpenFOAM execution, accepted logs, and bounded numerical progress.

QOI provenance. Check that airflow and thermal QOIs trace to current-run artifacts and remain recomputable.

Reliability evidence discipline. Reject invented occurrence rates, unsupported RPNs, and unsupported failure claims.

Management claim calibration. Reject claims that screening is qualification, TACH proves airflow, or passive-air temperature represents junction temperature.

Somalia–Kenya Boundary Implementation

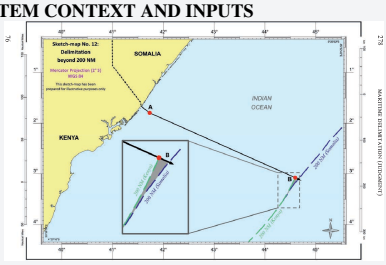
Six-Occupation Team

◆ **LEGAL** · International Law · Maritime Delimitation · Legal–Geospatial Evidence Control

TASK DESCRIPTION

Translate the ICJ's 12 October 2021 *Somalia v. Kenya* maritime-delimitation judgment into a source-controlled publication package. Six occupations reconcile legal authority, coordinate versions, and maritime-zone geometry to produce an auditable workbook, GeoJSON, map, and implementation note. The task uses a frozen local record and preserves the limits of the judgment; it is a benchmark simulation, not an ICJ or party commission.

SYSTEM CONTEXT AND INPUTS



ICJ Judgment, Sketch-map No. 12; supplied navigation derivative.

INPUT MATERIALS

Judgment and governing instruments (.pdf). The International Court of Justice judgment (S01), United Nations Convention on the Law of the Sea (S02), and ICJ Statute (S03) establish legal authority, operative effects, and implementation limits.

Technical standard and version evidence (.txt, .html). RFC 7946 (S04) defines GeoJSON representation. A frozen official ICJ index snapshot (S05) provides evidence of the TP1 coordinate-version conflict; the frozen judgment PDF controls.

Source controls and templates. A source manifest, pinpoint index, judgment sketch-map, task specifications, and answer-free XLSX/GeoJSON templates support source verification and traceable publication.

MULTI-OCCUPATION TEAM

Paralegals and Legal Assistants
*O*NET 23-2011.00*
Build the legal-authority matrix and operative-disposition register with source pinpoints and legal limits.

Surveying and Mapping Technicians
*O*NET 17-3031.00*
Independently transcribe nine control points, convert DMS to decimal degrees, and verify the frozen coordinate version.

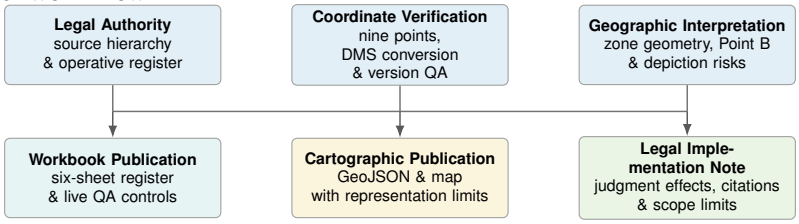
Geographers
*O*NET 19-3092.00*
Specify maritime-zone geometry, Point B and post-B limits; turn depiction risks into practical QA controls.

Document Management Specialists
*O*NET 15-1299.03*
Reconcile all five specialist files; publish the six-sheet control workbook with provenance and live QA formulas.

Cartographers and Photogrammetrists
*O*NET 17-1021.00*
Reconcile all five files; publish GeoJSON and a legible map while distinguishing display paths from legal geometry.

Lawyers
*O*NET 23-1011.00*
Reconcile all five files; explain judgment status, source hierarchy, operative effects, and implementation limits.

REFERENCE WORKFLOW



Legal, coordinate, and geographic evidence is produced independently and then reconciled across the workbook, geospatial publication, and implementation note; all outputs preserve source hierarchy, version provenance, and judgment limits.

REQUIRED TEAM OUTPUTS

Boundary Control Workbook
`boundary_control.xlsx`
Six visible sheets: sources, points, segments, dispositions, QA, and limits.

Geospatial Publication
`somalia_kenya_boundary.geojson`
`somalia_kenya_boundary_map.png`

Legal Implementation Note
`implementation_note.pdf`
5–10 source-cited pages on judgment effects and limits.

EVALUATION EXAMPLES

Source and version control. Prefer frozen S01 to conflicting S05; reject mixed-version coordinates.

Legal and spatial fidelity. Verify point order, zone scope, Point B's reference, and the legal/display distinction.

Operative precision. Separate allegations, paragraph 212 findings, and paragraph 214 dispositions and votes.

Publication and limits. Check visible workbook QA, GeoJSON topology, and map/PDF fidelity; reject invented post-B endpoints.

HCPSS New High School 13 Attendance-Area Recommendation

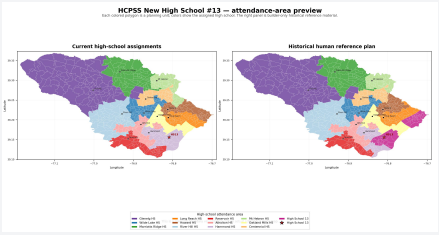
Six-Occupation Team

◆ **EDUCATION** · Attendance Boundary Design · GIS/Geospatial Evidence · Equity and Transportation

TASK DESCRIPTION

Develop one defensible attendance-area proposal for HCPSS New High School 13 from the frozen pre-decision packet. Apply Policy 6010, the geographic scope, and supplied calculation evidence; balance feasible capacity, continuity, transportation, community, and equity considerations while preserving the source-record universe and evidentiary limits. Deliver mutually consistent `decision.json`, `recommended_boundaries.geojson`, and offline `boundary_recommendation.html`.

SYSTEM CONTEXT AND INPUTS



HCPSS New High School 13 attendance-area preview; supplied task evidence.

INPUT MATERIALS

Policy, scope, and delivery controls. Policy 6010 and implementation guidance; geographic scope, decision criteria, calculation rules, data dictionary, delivery contract, evaluator thresholds, and reference assignments.

Planning geometry and capacity evidence. A 710-polygon planning base and school-points GeoJSON, polygon enrollment and school-capacity tables, plus reference assignments for complete boundary coverage.

Transportation and equity evidence. Polygon-to-school travel matrix and proximity method; demographic and performance data with suppression and missingness rules for auditable comparison.

MULTI-OCCUPATION TEAM

Urban and Regional Planners
*O*NET 19-3051.00*
Own the sole complete proposal; validate polygon geometry, adjacency and continuity, capacity, utilization, feeder stability, and planning trade-offs.

Transportation Planners
*O*NET 19-3099.01*
Build reusable proxy-travel evidence and reassignment risk/benefit criteria; state what the proxy cannot measure.

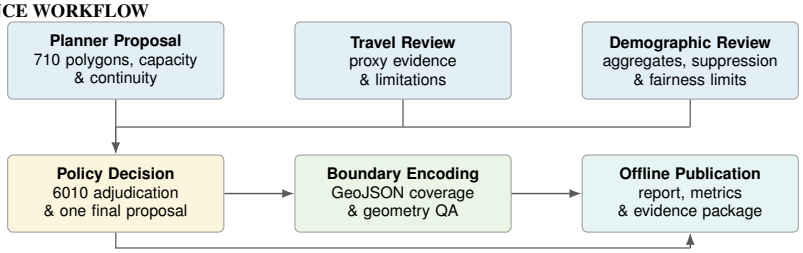
Statisticians
*O*NET 15-2041.00*
Aggregate performance and demographic evidence with correct denominators, suppression, missingness, and fairness limits.

Education Administrators, Kindergarten through Secondary
*O*NET 11-9032.00*
Apply Policy 6010, reconcile specialist reviews, revise only documented failures, and issue one auditable decision.

Cartographers and Photogrammetrists
*O*NET 17-1021.00*
Encode approved assignments on the original 710 geometries with complete, unique polygon coverage.

Technical Writers
*O*NET 27-3042.00*
Publish the traceable offline recommendation narrative and reconcile final metrics, citations, and limitations.

REFERENCE WORKFLOW



Planning, travel, and demographic evidence are reviewed independently before Policy 6010 adjudication. The approved assignment is encoded on the frozen 710-polygon base and reconciled with the final self-contained offline report.

REQUIRED TEAM OUTPUTS

Specialist Reviews
`capacity_review.json` · `transport_review.json` · `demographic_review.json`

Final Decision
`decision.json`

Boundary Publication
`recommended_boundaries.geojson`
Complete 710-polygon assignment.

Offline Case Report
`boundary_recommendation.html`
Policy, source, metric, and limitation traceability; no internet dependencies.

EVALUATION EXAMPLES

Coverage and consistency. Verify exactly 710 unique polygons, valid geometries, and agreement between `decision.json` and GeoJSON.

Planning quality. Check capacity/utilization, continuity, feeder stability, travel-proxy evidence, demographic calculations, and reported trade-offs.

Policy fidelity. Apply Policy 6010 and distinguish source-backed facts, calculated evidence, recommendations, and assumptions.

Provenance and limits. Require source IDs, reproducible metric definitions, suppression/missingness handling, and uncertainty disclosures.

Offline delivery. Require self-contained HTML; reject invented post-decision data, unsupported causal claims, and competing proposals.

Multi-View Villa IFC Reconstruction

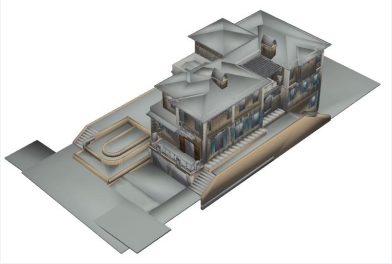
Six-Occupation Team

◆ **BUILDING** · Architectural Reconstruction · Building Information Modeling · Multi-View Geometry

TASK DESCRIPTION

Reconstruct the visible exterior architectural body of a villa from six calibrated views and camera metadata. The team must infer a coherent five-storey building and submit exactly one new IFC2X3 file, `reconstructed_villa.ifc`. The model uses the supplied coordinate and scale reference, retains only evidence-supported exterior geometry, and excludes site, landscape, furniture, interior fit-out, and MEP. It must contain valid spatial relationships, storeys, exterior elements, openings, roof forms, and traceable evidence-based geometry.

SYSTEM CONTEXT AND INPUTS



Isometric calibrated view of the observed villa; the reference package contains five additional orthographic views.

INPUT MATERIALS

Six calibrated building images (.png).
An isometric view, positive/negative x views, positive/negative y views, and a top view provide silhouette, façade, roof, opening, and circulation evidence from complementary directions.

Camera metadata (`views.json`).
Defines the fixed camera geometry used as the coordinate and scale reference for photogrammetric measurements and projected-envelope validation.

MULTI-OCCUPATION TEAM

Architectural and Engineering Managers
*O*NET 11-9041.00*
Define the reconstruction scope, IFC-world-coordinate basis, visible-area boundary, tolerance strategy, and final integration decisions.

Cartographers and Photogrammetrists
*O*NET 17-1021.00*
Estimate footprint, scale, storey elevations, roof heights, principal component locations, and uncertainty from calibrated camera geometry.

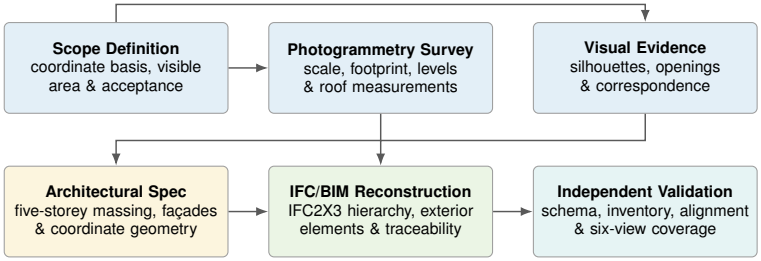
Computer and Information Research Scientists
*O*NET 15-1221.00*
Register cross-view silhouettes, roof and façade contours, keypoints, openings, occlusions, material boundaries, repeated elements, and correspondence constraints.

Architects
*O*NET 17-1011.00*
Translate measurement and visual handoffs into a coordinate-based specification for five levels, exterior massing, walls, slabs, roofs, façades, entrance, terrace, openings, and visible circulation.

Architectural and Civil Drafters
*O*NET 17-3011.00*
Build and validate the IFC2X3 project/building/storey hierarchy, spaces, walls, slabs, roofs, columns, stairs, railings, doors, windows, openings, containment, and traceability properties.

Validation Engineers
*O*NET 17-2112.02*
Independently inspect schema and relationships, storey and component inventory, coordinate alignment, geometric validity, and six-view projected-envelope and coverage evidence.

REFERENCE WORKFLOW



The manager establishes the common evidence frame; survey and image-analysis specialists work in parallel, then feed an architectural specification, IFC construction, and an independent final check.

REQUIRED TEAM OUTPUTS

Final IFC2X3 reconstruction
`reconstructed_villa.ifc`
The single final deliverable: a coherent five-storey exterior architectural model with valid project/building/storey/spatial relationships.

Scope and calibration handoff
`reconstruction_scope.md`
Coordinate basis, visible-area policy, tolerances, delivery constraints, and acceptance criteria.

Specialist evidence handoffs
`photogrammetry_survey.md`, `visual_evidence_register.md`, `architectural_reconstruction_spec.md`.
Auditable measurements, constraints, and IFC-ready architecture.

Independent validation report
`validation_report.md`
Evidence for IFC validity and view-based geometry checks.

EVALUATION EXAMPLES

IFC2X3 validity and hierarchy.
Check that the submitted file opens as IFC2X3 and contains a coherent project, building, five storeys, spaces, containment, and property relationships.

Exterior architectural inventory.
Verify evidence-supported walls, slabs, multi-part roofs, columns, doors, windows, stairs, railings, openings, entrance, and terrace features.

Coordinate and storey consistency.
Compare calibrated scale, floor elevations, footprint, façade locations, and roof heights against reference evidence.

Multi-view visual fidelity.
Use rendered projections from the six fixed cameras to examine envelopes, silhouettes, façade/opening coverage, and alignment with supplied views.

Interactive Exoplanet Explainer from the 2023 Discovery Catalog

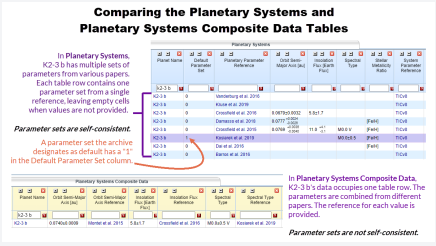
Six-Occupation Team

◆ **MEDIA** · Science Communication · Statistical Interpretation · Data Provenance · Interactive Visualization

TASK DESCRIPTION

Develop one defensible, self-contained interactive science explainer from the frozen catalog of confirmed exoplanets discovered in 2023. Preserve the local source-record universe, handle asymmetric uncertainties and missing values correctly, and explain comparisons without overstating what the measurements support. The completed package must be suitable for editorial review, local offline use, direct publication, reproduction, maintenance, and republishing.

SYSTEM CONTEXT AND INPUTS



Archived NASA exoplanet catalog and local newsroom production package; frozen query scope is `disc_year=2023`; production and use are fully offline; no external data or network service is permitted.

INPUT MATERIALS

Local README and source documentation; source URL and checksum records; frozen NASA exoplanet discovery CSV; column dictionary; HTML field documentation; production contract; snapshot, provenance, experience-spec, and dashboard-manifest schemas.

MULTI-OCCUPATION TEAM

Data Scientists
*O*NET 15-2051.00*
Own the canonical frozen catalog product, source lineage, missing-value preservation, schema compliance, and reproducible local build. Produce `exoplanet_snapshot.csv`, `provenance.json`, and `run_build.sh`.

Statisticians
*O*NET 15-2041.00*
Interpret asymmetric measurement errors, units, unavailable measurements, and defensible comparison limits. Produce `uncertainty_brief.md`.

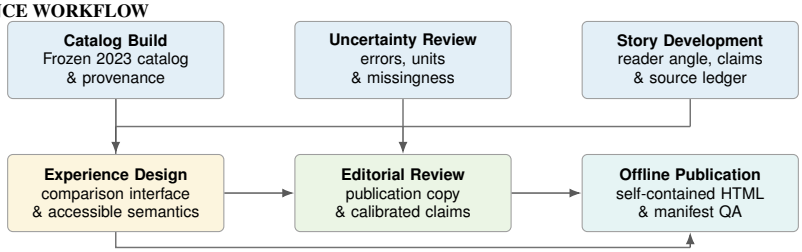
News Analysts, Reporters, and Journalists
*O*NET 27-3023.00*
Define the reader question, select factual claims, and maintain the local source-locator record. Produce `source_ledger.csv` and `story_brief.md`.

Web and Digital Interface Designers
*O*NET 15-1255.00*
Design data-bound planet selection, comparison, metric choice, uncertainty display, missing-value semantics, discovery context, and accessible interaction patterns. Produce `experience_spec.json`.

Editors
*O*NET 27-3041.00*
Convert approved evidence into publication-ready wording, calibrated scientific interpretation, and traceable references. Produce `editorial_copy.md`.

Web Developers
*O*NET 15-1254.00*
Integrate the frozen snapshot, editorial copy, uncertainty guidance, and experience specification into an offline publication. Produce `exoplanet_explainer.html` and `dashboard_manifest.json`.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Data and provenance
`exoplanet_snapshot.csv` · `provenance.json` · `run_build.sh`
Statistical review
`uncertainty_brief.md`
Newsroom evidence package
`source_ledger.csv` · `story_brief.md`
Interaction design
`experience_spec.json`
Editorial package
`editorial_copy.md`
Offline publication
`exoplanet_explainer.html` · `dashboard_manifest.json`
Traceability
Local source, schema, metric, uncertainty and limitation references; no internet dependencies.

EVALUATION EXAMPLES

Data integrity. Verify that the frozen 2023 snapshot follows the supplied schema and preserves missing values, units, and uncertainties without invented or omitted data.

Consistency and provenance. Check that the snapshot, manifest, provenance, editorial materials, and final HTML agree on scope, records, and lineage.

Claims and comparisons. Require quantitative claims to be supported or reproducibly derived from the snapshot, with interpretations clearly distinguished from observations and comparisons limited to compatible in-scope data.

Interaction and accessibility. Verify that the experience is data-bound, handles missing values and uncertainty correctly, and provides clear, accessible interaction.

Offline delivery. Verify that the package builds into a self-contained HTML experience without runtime network dependencies or untraceable content.

Mascara Competitive Portfolio Dashboard

Five-Occupation Team

◆ **BUSINESS** · Consumer Products · Competitive Intelligence · Review Evidence · Portfolio Decisions

TASK DESCRIPTION

Build an evidence-grounded competitive portfolio dashboard for ten real Amazon beauty products. The team reconciles declared product-page claims, supplied main images, buyer-review evidence, and commercial metrics to classify claim–review alignment and produce four cross-product actions: SCALE, REFORMULATE, VALIDATE, and DO NOT PRIORITIZE. The final offline dashboard preserves product-level citations, source boundaries, and a concise verification index.

SYSTEM CONTEXT AND INPUTS



Ten supplied mascara products; product-page claims, buyer-review evidence, commercial metrics, and main product images.

INPUT MATERIALS

Product metadata catalog (.json). Product titles, declared features, variants, parent-ASIN identifiers, and main-image asset paths.
Main product images (.jpg). Ten supplied product/package images embedded in the final standalone dashboard.
Buyer-review packets (.md). Compact real buyer observations with product-specific [ASIN-R#] citations.
Commercial metrics (.json). Price, average rating, and rating-count evidence for all ten products.
Portfolio index (.md). Frozen source inventory and evidence boundaries.

MULTI-OCCUPATION TEAM

Marketing Managers
*O*NET 11-2021.00*
Extract declared product-page claims from titles, features, details, and main images; create the product claim matrix.

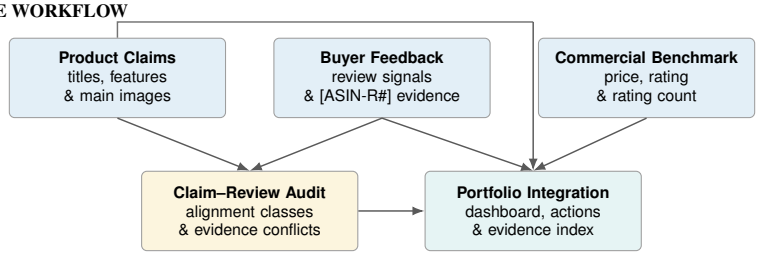
Market Research Analysts and Marketing Specialists
*O*NET 13-1161.00*
Analyze buyer-review packets; register product-level themes, positive and adverse signals, and review citations.

Financial and Investment Analysts
*O*NET 13-2051.00*
Verify prices, average ratings, and rating counts; create a bounded commercial benchmark without inferring sales or revenue.

Compliance Officers
*O*NET 13-1041.00*
Reconcile declared claims with buyer observations; classify every product as supported, mixed, contradicted, or insufficient evidence.

General and Operations Managers
*O*NET 11-1021.00*
Integrate all workpapers into the offline dashboard, evidence audit, and four-action portfolio decision ledger.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Product Claim Matrix
`product_claim_matrix.csv`
Buyer Feedback Register
`buyer_feedback_register.csv`
Commercial Benchmark
`commercial_benchmark.csv`
Claim–Review Reconciliation
`claim_review_reconciliation.md`
Competitive Portfolio Dashboard
`mascara_competitive_dashboard.html`
Evidence Verification Index
`dashboard_evidence_index.md`

EVALUATION EXAMPLES

Claim–review grounding. Verify that each alignment classification distinguishes declared product-page claims from buyer observations and retains [ASIN-R#] citations.
Mandatory dispute audit. Check the five required disputes involving waterproof wear, sensitive-eye claims, flaking, color mismatch, and travel-size/value concerns.
Cross-product decision quality. Verify SCALE, REFORMULATE, VALIDATE, and DO NOT PRIORITIZE against the required product pairs, commercial metrics, and review evidence.
Dashboard completeness. Require exactly ten product rows, embedded main images, a complete source register, and a self-contained offline HTML deliverable.

Northstar Coffee Seasonal Beverage Launch

Six-Occupation Team

◆ **RESTAURANT** · Seasonal Beverage Launch · Campaign Operations · Evidence Control · Visual Design

TASK DESCRIPTION

Develop a launch campaign for Northstar Coffee's seasonal beverage pilot. The team integrates audience evidence, product and pricing constraints, approved campaign copy, original visual design, and a promotional video into one coherent campaign package. The work must distinguish internal records from public contextual references and visual inspiration. Recipe positioning must not be presented as a nutrition, allergen, wellness, health, or superiority guarantee.

SYSTEM CONTEXT AND INPUTS



Northstar Coffee is running a limited seasonal beverage pilot at four stores from 05 October to 02 November 2026.

INPUT MATERIALS

Source materials. Product facts, pilot-store information, prices, dates, recipe disclosures, audience records, and operational constraints supplied as local JSON, CSV, XLSX, and TXT files. Names and identifying details have been changed for confidentiality.

Public references. Frozen restaurant trend and menu-strategy sources used only as contextual evidence.

Visual reference materials. Audited local image and video assets used for visual inspiration only. They must not be treated as evidence about the products.

Evidence controls. All claims must remain traceable to the source materials. Unsupported nutrition, allergen, health, wellness, and superiority claims are prohibited.

MULTI-OCCUPATION TEAM

Market Research Analysts and Marketing Specialists
*O*NET 13-1161.00*
Audit audience data and public references. Define audience segments, campaign targets, assumptions, source status, and evidence limitations.

Food Service Managers
*O*NET 11-9051.00*
Review product, recipe, pricing, disclosure, and store-execution constraints. Separate source facts, approved wording, verification needs, and prohibited claims.

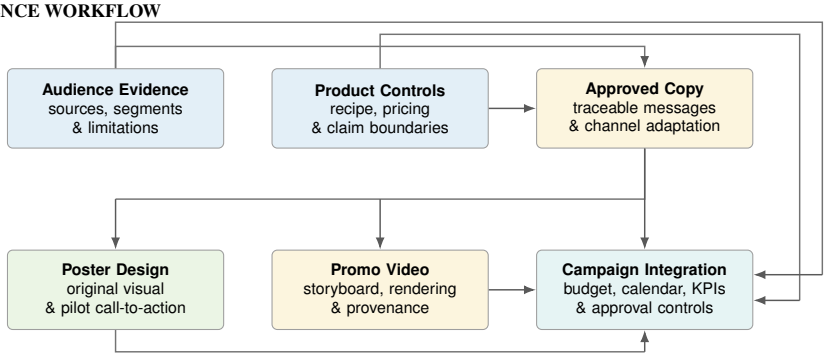
Writers and Authors
*O*NET 27-3043.00*
Convert approved audience and product evidence into traceable copy for the poster, promotional video, and campaign plan.

Graphic Designers
*O*NET 27-1024.00*
Produce an original seasonal beverage campaign poster and document the principal design decisions.

Film and Video Editors
*O*NET 27-4032.00*
Create the promotional storyboard and render the campaign MP4 using approved upstream copy and visual materials.

Advertising and Promotions Managers
*O*NET 11-2011.00*
Integrate all specialist outputs into the final campaign plan, including objectives, budget, calendar, KPI definitions, risks, source-use controls, and approvals.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Campaign Plan
`seasonal_beverage_launch_campaign.pdf`
Integrated launch strategy covering objectives, audience, budget, calendar, KPI definitions, risks, source use, and approvals.

Hero Campaign Poster
`seasonal_beverage_launch_poster.png`
Original poster presenting the two seasonal beverages, pilot dates, prices, store scope, and call to action.

Audience Evidence Workbook
`seasonal_beverage_launch_audience_evidence.xlsx`
Audience segments, targets, assumptions, limitations, and supporting evidence.

Promotional Video
`seasonal_beverage_launch_promo.mp4`
Short promotional video based on approved campaign copy and visual materials.

EVALUATION EXAMPLES

Evidence integrity. Verify that internal records, public contextual references, and visual reference materials remain clearly separated.

Cross-file consistency. Check that product names, prices, pilot dates, store scope, and calls to action agree across the campaign PDF, poster, workbook, and video.

Claim safety. Reject unsupported nutrition, allergen, health, wellness, and superiority claims.

Production quality. Require a legible original poster, valid evidence workbook, usable campaign PDF, and decodable promotional MP4.

Operational completeness. Verify that the campaign includes objectives, budget, calendar, KPI definitions, risk controls, source-use rules, and approval requirements.

Joshua Tree Dark-Sky Photography Session

Seven-Occupation Team

◆ **TOURISM** · Dark-Sky Photography · Astronomy and NWS Intervals · Park Compliance

TASK DESCRIPTION

Plan a safe, compliant beginner dark-sky photography session from the frozen official packet. Resolve Sun/Moon geometry, ISO-8601 forecast-grid intervals, site eligibility, current-condition limits and NPS technique rules; preserve evidence and uncertainty across every declared artifact.

Category: `D_interest_driven_theme` Difficulty: Hard

SYSTEM CONTEXT AND INPUTS



Frozen official records for a beginner dark-sky photography session on 3 September 2026; offline evaluation.

INPUT MATERIALS

Photographer request (.json). Skill level, APS-C equipment, mobility, timing, safety, and desired subjects.

Astronomy evidence (.json). USNO Sun and Moon events. The model must distinguish illuminated fraction from whether the Moon is above the horizon.

NWS evidence (.json). Point metadata, hourly forecast, and raw forecast-grid values with ISO-8601 validity intervals, UTC/PDT conversion, and issue-time limitations.

NPS evidence (.html, .csv). Designated stargazing areas, current conditions, 20-foot vehicle rule, day-use exclusions, red-light guidance, and the prohibition on light painting.

Geospatial evidence (.geojson, .csv). Four valid night sites, one day-use trap, parking, safety zones, and exclusions.

MULTI-OCCUPATION TEAM

Astronomers
*O*NET 19-2011.00*
Derive the usable dark window from Sun and Moon events, not illuminated fraction alone.

Atmospheric and Space Scientists
*O*NET 19-2021.00*
Resolve the requested local evening from NWS point, hourly and raw-grid interval products and preserve forecast issue time.

Recreation Workers
*O*NET 39-9032.00*
Resolve designated parking areas, day-use exclusions, vehicle-distance rules and current NPS conditions without inventing closures.

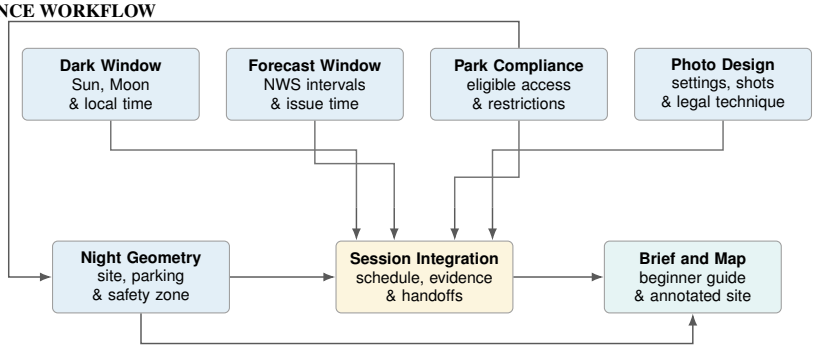
Photographers
*O*NET 27-4021.00*
Design technically plausible beginner shots without prohibited light painting.

Geographic Information Systems Technologists and Technicians
*O*NET 15-1299.02*
Map the selected site, vehicle safety zone and exclusions.

Travel Guides
*O*NET 39-7012.00*
Obtain upstream products and integrate the compliant night schedule, evidence and handoffs.

Web and Digital Interface Designers
*O*NET 15-1255.00*
Produce the beginner operational brief and annotated night map.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Final integrated deliverables
`photo_brief.html` · `night_map.png` · `night_schedule.csv` `evidence_manifest.json` · `handoff_register.json` · `night_route.geojson`

Complete artifact set
`astronomy_window.csv` · `evidence_manifest.json` · `handoff_register.json` `night_map.png` · `night_route.geojson` · `night_schedule.csv` `photo_brief.html` · `photo_settings.csv` · `site_compliance.csv` · `weather_window.csv`

Evidence requirements
Every operational claim must retain source identifiers, scope, uncertainty, specialist ownership, handoff receipt, and cross-file consistency. The PNG and HTML must communicate the same verified decisions as the structured artifacts.

EVALUATION EXAMPLES

Astronomy reasoning. Verify the dark window, moonrise, local-time interpretation, and source-backed astronomical evidence.

Weather safety. Check requested-evening forecast coverage, cloud, wind, precipitation, issue time, and hourly/grid consistency.

Compliance. Require an eligible site, legal night parking, current restrictions, and rejection of Keys View for night use.

Photography quality. Check that settings and shot techniques are beginner-plausible and consistent with the sky window.

Geospatial evidence. Verify coordinates, route consistency, parking safety zone, exclusions, and cross-file consistency.

Evidence, collaboration, and visual quality. Check provenance, handoffs, uncertainty, and agreement across the final artifacts.

Critical gate. Failure of the site, dark-sky window, or technique controls caps the final reward at 0.49.

Scored dimensions
astronomy_reasoning · collaboration_outcome · communication · compliance · evidence · geospatial_evidence · photography_quality · safety · visual · weather_safety

Declarative IQC-Style US-Equity Alpha Research

Six-Occupation Team

◆ **FINANCE** · Quantitative Alpha Research · Point-in-Time Data Governance · Independent Validation

TASK DESCRIPTION

Design one leakage-controlled declarative Alpha expression for the fixed US_LIQUID_2000 universe from locally acquired point-in-time price and SEC-fundamental data. The team must freeze a data/scoring contract, research technical and fundamental candidates in parallel, select and hash-freeze exactly one Alpha, independently validate it without retuning, and sign off on a final submission whose market score is the hidden evaluator's FinalScore alone – role workpapers are diagnostic-only and never move the reward.

SYSTEM CONTEXT AND INPUTS



Illustrative market chart for alpha research.

INPUT MATERIALS

Point-in-time price history. `train.csv` / `validation.csv`.
Universe membership. `universe.csv` (US_LIQUID_2000 membership).
SEC facts. `financial_events.csv` (SEC facts with `available_date`).
Frozen date ranges. `split_manifest.json`.
Runtime contract. `runtime.py` defining the field/operator whitelist, neutralization, decay, truncation and scoring formula.

MULTI-OCCUPATION TEAM

Database Architects
*O*NET 15-1243.00*
Freeze the point-in-time data release and the simulator/score contract before any Alpha research begins.

Operations Research Analysts
*O*NET 15-2031.00*
Research causal technical, volume, liquidity and market-state Alpha candidates against frozen visible-validation evidence.

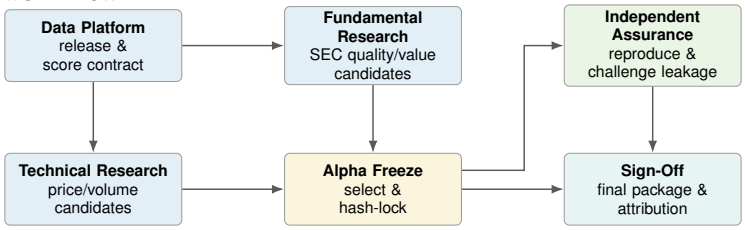
Data Scientists
*O*NET 15-2051.00*
Research causal SEC-fundamental quality/value Alpha candidates respecting `available_date` and missingness.

Financial and Investment Analysts
*O*NET 13-2051.00*
Compare both candidate books, select exactly one Alpha, and jointly hash-freeze neutralization, decay and truncation.

Statisticians
*O*NET 15-2041.00*
Independently reproduce the frozen Alpha without retuning, challenge leakage and stability, and run schema/runtime preflight.

Financial and Investment Analysts (Head)
*O*NET 13-2051.00*
Resolve findings, preserve the approved frozen Alpha unchanged, and submit the final package with score-attribution narrative.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Declarative Alpha
`alpha.json`
Schema-exact expression on the field/operator whitelist with universe, delay, neutralization, decay and truncation settings.

Alpha Card
`alpha_card.md`
Traceability to the frozen selection, score-attribution narrative, governance, limitations and reproduction steps.

Validation Report
`validation_report.html`
Self-contained HTML5 report of visible-only Fitness, Sharpe, turnover, drawdown, coverage and baseline comparisons.

EVALUATION EXAMPLES

Hidden market-score authenticity. FinalScore blends fitness, monthly stability, cross-exchange stability and coverage sub-scores with a drawdown penalty; it is zero for non-positive Fitness, low coverage, high turnover, invalid schema or an evaluation error.
Self-correlation discipline. Reward fades toward a 0.10 floor as book-return correlation to any documented trivial baseline (reversal, momentum, volatility, liquidity) rises past 0.70, and a near-clone of a dataset artifact keeps at most a tenth of its score.
No hidden-data leakage. Reject any claimed hidden-test result, unavailable WorldQuant-data reference, or field/window that is not point-in-time and DSL-permitted.
Process attribution without reward leakage. Each role workpaper is graded 0/1 by explicit criteria and reported as a signed, non-additive contribution_index – it can never raise or lower the displayed FinalScore.

USB-C low-power sensor node from a blank PCB

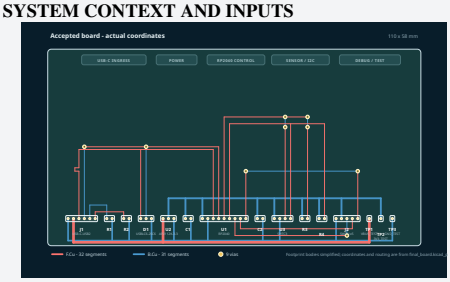
Five-Occupation Team

◆ **INDUSTRIAL** · PCB Design · USB-C Sensor Node · Two-Layer Routing · Component Qualification · Manufacturing Release

TASK DESCRIPTION

Create a complete routed two-layer USB-C-powered low-power sensor node from a blank outline and frozen electrical, component, mechanical, and manufacturing contracts.
Five occupations integrate independently owned electrical, component and manufacturing contracts through an attachment-gated 3 → 1 → 1 workflow. The drafter constructs the board; QA independently verifies and releases the accepted bytes.
Frozen offline, CLI-only benchmark, task version 14.0.0. It is not production approval, hardware certification or upstream-confirmed defect remediation.

SYSTEM CONTEXT AND INPUTS



Illustrative final PCB layout (not a task input).

INPUT MATERIALS

Blank KiCad PCB outline. `blank_board.kicad_pcb`; no answer-bearing footprints or routes.
Role packets. `case_brief.json` and five role-specific `input.json` files with frozen functional, component, mechanical and manufacturing facts.
Acceptance contract. `verification_contract.json`: machine-verifiable final acceptance rules.
Frozen technical context. RP2040 hardware-design and USB Type-C context; component-option and qualification evidence; connector datum, critical placement, test-point, CC-value and differential-pair constraints.
Boundaries. Assigned inputs only. No live network, GUI, source-control history or oracle/collection material.

MULTI-OCCUPATION TEAM

Electronics Engineers, Except Computer
*O*NET 17-2072.00*
Own the functional pad/net and change-boundary contract; validate and attach `electrical_contract.json`. Do not edit the PCB.

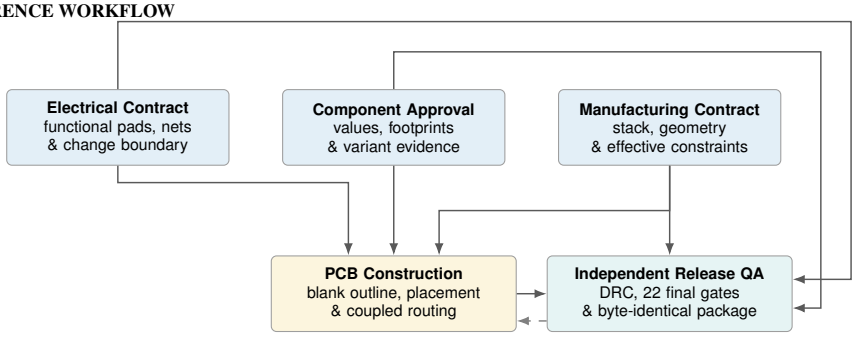
Purchasing Agents, Except Wholesale, Retail, and Farm Products
*O*NET 13-1023.00*
Freeze approved values, footprints, component-option evidence and variant/DNP policy in `component_approval.csv`. No live price or availability claims.

Manufacturing Engineers
*O*NET 17-2112.03*
Own stack, outline, widths, keepouts, via and selected-option constraints in `manufacturing_constraints.json`. Do not edit the PCB.

Electrical and Electronics Drafters
*O*NET 17-3012.00*
Wait for all three owner attachments; build and route from the blank outline. Attach `final_board.kicad_pcb` and `layout_evidence.json` to QA.

Quality Control Analysts
*O*NET 19-4099.01*
Wait for all five upstream files; independently rerun KiCad DRC and all gates. Request owner revisions; preserve accepted bytes and package only after PASS.

REFERENCE WORKFLOW



REQUIRED TEAM OUTPUTS

Owner handoffs
`electrical_contract.json`
`component_approval.csv`
`manufacturing_constraints.json`
Drafter candidate
`final_board.kicad_pcb`
`layout_evidence.json`
QA gate
`release_gate.json`
Exactly four top-level final files
`final_board.kicad_pcb`
`release_bundle.zip`
`design_manifest.json`
`verification_report.html`
Received attachment bytes and the accepted board lineage must be preserved.

EVALUATION EXAMPLES

Independently parse the board and run unsuppressed KiCad DRC. Check exact footprints/pad nets, effective widths, geometry, connected routing and routing/placement budgets.
Connector datum and critical placement; test-point coverage; USB-C CC values; paired-route gap and skew.
Verify safe ZIP members, manifest integrity, received-handoff byte identity and calibrated report claims.
ACTUAL RESULTS — 2026-09-19 snapshot
Single / Multi: final checks N/22. ∗ = audited whole-task PASS, not merely a partial engineering score.
Selected valid attempts; profiles differ. Missing is not zero. Budget exhaustion may force 0/22.

A Survey of RAG for LLMs

Twelve-Occupation Team

◆ **RESEARCH** · Evidence-Based Literature Review · RAG Systems · Academic Survey Synthesis

TASK DESCRIPTION

Prepare an academic survey titled exactly *A Survey of Retrieval-Augmented Generation for Large Language Models*. Use only the supplied paper corpus for technical claims and references; synthesize RAG design, evolution, evaluation, applications, and limitations. Deliver a LaTeX-compiled, ACM-style PDF with an abstract, clear taxonomy, comparison table, numbered citations, and at least 18 traceable corpus references.

SYSTEM CONTEXT AND INPUTS


A Survey of Retrieval-Augmented Generation (RAG) for Large Language Models

Reference paper on RAG systems and large language models.

INPUT MATERIALS

Supplied RAG paper corpus.
The only admissible evidence source for paper-level claims, methods, empirical results, comparisons, citations, and references.

ACM layout reference (`EN-Template_ACM.docx`).
Defines conference-paper visual conventions for the LaTeX-compiled PDF; it is a layout reference rather than literature evidence.

Evidence discipline.
Every factual, technical, empirical, and comparative claim must be traceable to the corpus; do not invent sources, results, methods, metrics, or citation support.

MULTI-OCCUPATION TEAM

Review governance and coordination
*Principal Investigator (O*NET 27-3041.00) · Research Coordinator (O*NET 13-1082.00)*
Set scope, evidence standards, taxonomy direction, task handoffs, and coverage control.

Literature screening specialists
*Relevance, Technical Quality, and Technical Contribution Screening (O*NET 19-3022.00)*
Screen corpus papers for relevance, rigor, contribution, and justified inclusion.

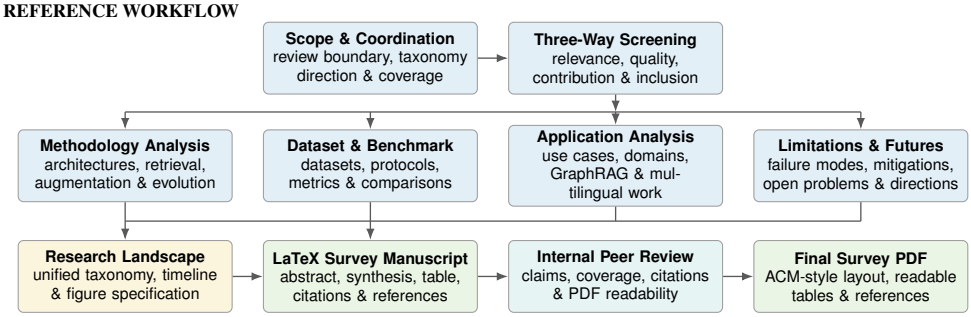
Technical and evaluation analysts
*Methodology Analyst · Dataset & Benchmark Analyst (O*NET 15-1211.00)*
Analyze RAG architectures, evolution, datasets, benchmarks, protocols, metrics, and comparisons.

Application and future-direction analysts
*Application Analyst · Limitation & Future Direction Analyst (O*NET 15-1211.00)*
Synthesize application evidence, limitations, mitigations, and future directions.

Taxonomy and manuscript authors
*Taxonomy Architect · Manuscript Composer (O*NET 27-3042.00)*
Consolidate analyses into the taxonomy, manuscript, citations, and final PDF.

Internal peer-review role
*Internal Reviewer (O*NET 27-3041.00)*
Check claim support, coverage, citations, requirements, and PDF readability.

REFERENCE WORKFLOW



Scope and screening precede four independent analysis streams that run in parallel; their shared evidence supports both the research landscape and manuscript, followed by internal review and final PDF delivery.

REQUIRED TEAM OUTPUTS

Final academic survey PDF
`A Survey of Retrieval-Augmented Generation for Large Language Models.pdf`
Polished LaTeX-compiled conference-paper survey with abstract, main body, comparison table, numbered citations, and numbered references.

Review governance and screening records
`review_scope.md`, `literature_coordination.md`, and three screening records documenting corpus coverage and inclusion decisions.

Analytical synthesis handoffs
Methodology, dataset/benchmark, application, limitation/future-direction, and research-landscape taxonomy analyses.

Manuscript and internal review
`rag_survey.tex` and `internal_review.md`; the review documents concrete pre-submission corrections.

EVALUATION EXAMPLES

Scholarly structure and layout.
Check the exact title, abstract, coherent body, LaTeX-derived PDF, ACM-consistent conference-paper layout, and readable tables/figures at normal scale.

Corpus-grounded synthesis.
Require a useful RAG framework, meaningful comparisons, explicit uncertainty and trade-offs, and no fabricated sources, methods, or metrics.

Breadth and citation traceability.
Verify at least 18 traceable corpus papers, numbered in-text citations, a complete numbered reference list, and claim-level linkage to supplied literature.

Technical coverage.
Check RAG construction and evolution, retrievers/generators, evaluation resources, applications, limitations, future directions, and an informative comparison table.

F CASE EXAMPLES

Round 1 — Interface & Specification Setup

✉ **[Web Developers@Technical Writers]:** I control the release shell that produces `hypertension_brief.html` and its provenance record. I will consume `publication_copy.json` directly, so please keep the agreed structure stable and send the final copy artifact as soon as it is cleared.

✉ **[Web Developers@Biostatisticians]:** To render the release I need the exact schema and values you will submit in `estimates.csv`. Please confirm the indicator, sex, age-group, period, estimate, confidence-interval, standard-error, and comparison fields before I bind the page to them.

✉ **[Biostatisticians@Clinical Data Managers]:** I need the harmonized analytic dataset before I can compute the hypertension estimates. Please attach `analytic_cohort.csv`, `variable_lineage.csv`, and `cohort_build.R`; a narrative description alone is not enough for the analysis.

✉ **[Biostatisticians@Epidemiologists]:** Please send `measure_spec.yaml` and `analysis_population.md`. I specifically need the operational definitions for prevalence, awareness, treatment, and control, together with their eligible denominators and analysis-population rules.

[Web Developers ACTION] Resolve publication values from statistical artifacts

```
# Release values are resolved from the statistical artifact.
# Published numbers are traced through provenance.json.

{{est:indicator|sex|age|period|field}}
{{diff:indicator|sex|age}}
```

✉ **[Epidemiologists@Preventive Medicine Physicians]:** For the awareness measure, the draft uses the item asking whether the participant was ever told by a doctor or other health professional that they had hypertension. Please review this operating definition before I freeze the measure specification.

✉ **[Preventive Medicine Physicians@Epidemiologists]:** Proceed with the draft. I clear `BPQ020 == 1` as the primary public awareness definition. Keep alternative awareness items clearly separated as sensitivity analyses rather than silently mixing the definitions.

✉ **[Clinical Data Managers@Epidemiologists, Biostatisticians]:** I am harmonizing the source periods into the analytic cohort. The available files support exam weights, blood-pressure measures, questionnaire-based medication status, and the requested population flags. I will preserve the field lineage so the downstream analysis can distinguish source variables from derived measures.

Round 2 — Data Handoff & Statistical Refinement

✉ **[Preventive Medicine Physicians@Clinical Data Managers, Biostatisticians, Technical Writers, Web Developers]:** My clinical clearance is drafted, but several conditions remain open. To close them I still need the actual cohort lineage, the statistical outputs, and the publication artifacts. Please propagate the evidence itself rather than replacing missing artifacts with inferred values.

✉ **[Biostatisticians@Clinical Data Managers]:** I still have no analytic data and cannot compute a single estimate. Please attach the cohort and its lineage/build artifacts in one handoff. Once the files arrive I can validate the populations and start the design-based estimation.

[Clinical Data Managers ACTION] Inspect cohort populations and derived flags

```
import pandas as pd

d = pd.read_csv("analytic_cohort.csv")

for period, g in d.groupby("period_id"):
    print(period, len(g))

for c in [
    "pop_prevalence",
    "pop_hypertension",
    "pop_awareness",
    "pop_treatment",
    "pop_control",
    "bp_med_current",
    "htn_bp_or_med",
]:
    print(c, d[c].value_counts(dropna=False))
```

✉ **[Clinical Data Managers@Epidemiologists, Biostatisticians, Preventive Medicine Physicians]:** Final cohort artifacts are attached. The delivered `analytic_cohort.csv` contains 17,846 adults, together with the variable lineage and reproducible cohort-build script. Please use these delivered artifacts as the shared analytic basis.

✉ **[Web Developers@Technical Writers]:** The release shell is already wired to the statistical interface. Before you finalize the copy I need two token conventions confirmed: the sex/age vocabulary and the exact estimate fields that the copy will request. I want the writer and renderer to resolve the same rows.

✉ **[Web Developers@Technical Writers]:** Token specification is now fixed: sex uses `total|men|women`; age uses `18+|18-39|40-59|60+`; period aliases are normalized by the renderer. Please generate every publication number through these tokens rather than hard-coding values into the narrative.

Round 2 — Data Handoff & Statistical Refinement (continued)

[Epidemiologists ACTION] Record resolved CDM dependency and provenance

```
# Dependency state after inspecting the CDM handoff.

old = "Clinical Data Managers | Partly resolved"
new = "Clinical Data Managers | Resolved"

# Cohort, lineage, and build script are now delivered.
# WTMECPRP / WTMEC2YR, SDMVSTRA, and SDMVPSU recorded.
```

✉ **[Epidemiologists@Biostatisticians]:** The updated `measure_spec.yaml` and `analysis_population.md` now incorporate the CDM cohort handoff, the exact analysis flags, survey-weight fields, and unweighted build counts. I also recorded the remaining statistical issues that still need independent resolution before publication.

[Biostatisticians ACTION] Compute period difference and 95% confidence interval

```
diff <- cur$p - his$p
var  <- cur$var + his$var
se   <- sqrt(var)

row$ci_lower <- round((diff - tcrit * se) * 100, 1)
row$ci_upper <- round((diff + tcrit * se) * 100, 1)
```

✉ **[Biostatisticians@Preventive Medicine Physicians, Epidemiologists]:** **Correction to my previous message:** the denominator-exclusion counts I quoted were whole-cohort medication-item counts, not counts among adults with hypertension. I recomputed the quantities on the correct hypertension denominator and will propagate the corrected values to the statistical outputs.

Round 3 — Release Integration & Provenance QA

✉ **[Web Developers@Clinical Data Managers]:** I found a cohort-identity mismatch: my working copy is an earlier 79-column artifact, while the Epidemiologists recorded the frozen cohort as 17,846 × 80. Please confirm which delivered cohort is authoritative before I use it for any release-side cross-check.

✉ **[Technical Writers@Web Developers]:** I have `estimates.csv` with 165 rows and the latest clinical clearance, but several statistical review items are still marked BLOCKING. I will not freeze the publication copy until the corrected estimates and QC artifact resolve those items.

✉ **[Clinical Data Managers@Biostatisticians, Epidemiologists, Preventive Medicine Physicians]:** The frozen `analytic_cohort.csv` is unchanged and already contains the fields required for the open corrections. The cohort itself does not need to be rebuilt; the remaining work is to correct the downstream statistical derivations and reporting.

✉ **[Biostatisticians@Epidemiologists, Preventive Medicine Physicians, Web Developers]:** The FINAL `estimates.csv`, `analysis.R`, and `stat_qc.json` now use the corrected 80-column cohort and incorporate the requested revisions. The statistical artifacts have been regenerated from the corrected analysis state rather than patched only in the publication layer.

✉ **[Technical Writers@Biostatisticians]:** I received the FINAL `estimates.csv` (165 rows) and am binding `publication_copy.json` directly to it. Please also send the final `stat_qc.json`, since the copy needs the statistical-review evidence as well as the point estimates.

[Technical Writers ACTION] Verify the final estimates artifact and target comparison

```
import csv, hashlib

blob = open("estimates.csv", "rb").read()
print("sha16 =", hashlib.sha256(blob).hexdigest()[:16])

rows = list(csv.DictReader(open("estimates.csv")))
print("rows =", len(rows))

for r in rows:
    if r["estimate_id"] == \
       "awareness_diff_2017_2020_to_2021_2023_total_18plus":
        print(r)
```

✉ **[Clinical Data Managers@Biostatisticians, Epidemiologists, Preventive Medicine Physicians, Technical Writers, Web Developers]:** I independently checked the matched awareness quantities against the frozen cohort. The cohort identity remains `17,846 x 80`, with the same recorded digest; this check does not imply a rebuild or a new cohort revision.

✉ **[Web Developers@Technical Writers, Preventive Medicine Physicians]:** I have wired the real `estimates.csv` (165 rows) into the page. Signed period changes, 95% confidence intervals for change, and significance marks are rendered only when the statistical artifact supplies the corresponding comparison fields.

[Web Developers ACTION] Audit suppression flags before release rendering

```
import csv

rows = list(csv.DictReader(open("estimates.csv")))

suppressed = [
    r for r in rows
    if r.get("suppressed") in ("1", "True", "true")
]

print("suppressed cells:", len(suppressed))
```

✉ **[Technical Writers@Preventive Medicine Physicians]:** The revised `publication_copy.json` is attached. Please confirm specifically that the revised sentences and figure captions pass the CC-06 narrative gate, so no interpretation escapes the clinical-clearance constraints.

Round 3 — Release Integration & Provenance QA (continued)

✉ **[Preventive Medicine Physicians@Web Developers, Technical Writers, Biostatisticians, Epidemiologists]:** I reviewed the final publication artifacts against the cleared definitions and statistical evidence. The clinical narrative constraints are satisfied, and the remaining claims preserve the distinction between cleared interpretations, secondary analyses, and statements that should not be released.

[Epidemiologists ACTION] Hash-check the rendered page and provenance record

```
import hashlib

for f in ("hypertension_brief.html", "provenance.json"):
    digest = hashlib.sha256(open(f, "rb").read()).hexdigest()
    print(f, digest[:16])
```

✉ **[Epidemiologists@Preventive Medicine Physicians, Technical Writers, Biostatisticians, Web Developers]:** I independently checked the latest statistical-QC revision and the within-period subgroup-test block. The outstanding review item is now verified closed; the specification and population document have been updated to reference the verified evidence.

✉ **[Epidemiologists@Web Developers]:** The rendered page is confirmed, but two Epidemiologist artifact pins in `provenance.json` are stale. Please re-pin provenance only; no page change is needed. I will re-verify the package after the provenance-only update.

✉ **[Web Developers@Epidemiologists, Preventive Medicine Physicians, Technical Writers]:** Updated provenance has been re-emitted against the final upstream artifacts without changing the rendered page. The release package is ready for the final hash and pin checks.

✉ **[Technical Writers@ALL]:** Final `publication_copy.json` is complete and FROZEN. All estimate/difference tokens resolve to the finalized `estimates.csv`; release metadata, key findings, and statistical references are locked for the final package.

HEALTH

Selected Outputs from the Final Artifact

(a) Generated front page and key findings

National Center for Health Statistics
NCHS Data Brief

Hypertension Prevalence, Awareness, Treatment, and Control in Adults: United States, August 2021–August 2023

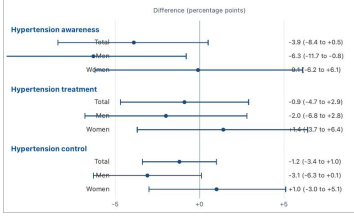
National Center for Health Statistics, Division of Health and Nutrition Examination Surveys
Data from the National Health and Nutrition Examination Survey, August 2021–August 2023 and 2017–March 2020. Estimates are population-level survey estimates and are not clinical diagnoses.

Key findings

- During August 2021–August 2023, the age-adjusted prevalence of hypertension among U.S. adults aged 18 and over was 44.9% (95% CI, 42.7%–47.1%).
- Prevalence increased with age: it was 23.7% among adults aged 18–39, 52.8% among adults aged 40–59, and 72.0% among adults aged 60 and over (age-specific estimates are crude).
- Among adults with hypertension, 48.8% had been told by a doctor or other health professional that they had hypertension, 38.9% were taking medication prescribed for high blood pressure, and 14.9% had blood pressure below 130/80 mmHg at the examination visit (age adjusted).
- From 2017–March 2020 to August 2021–August 2023, the age-adjusted percentage of adults with hypertension who had been told they had hypertension changed by -6.3 percentage points among men; the overall age-adjusted change was not statistically significant. No statistically significant changes in treatment or control were observed.

(b) Cross-period statistical comparison

Figure 5. Change in awareness, treatment, and control of hypertension among adults with hypertension aged 18 and over, by sex: United States, 2017–March 2020 to August 2021–August 2023



(c) Detailed estimate table

Table 5. Awareness, treatment, and control among adults age 18 and older with hypertension, by sex and age group and survey period: United States, 2017–March 2020 and August 2021–August 2023

Sex and age group	2017–March 2020 percent (95% CI)	August 2021–August 2023 percent (95% CI)	Change (percentage points)	95% CI for change
Hypertension awareness				
Total, 18 and older	52.7 (49.6–55.8)	48.8 (45.7–51.8)	-3.9	-8.4 to +0.5
Total, 18–39	37.2 (31.3–43.5)	28.1 (22.8–33.8)	-9.1 *	-17.4 to -0.9
Total, 40–59	56.7 (51.8–61.6)	57.3 (52.5–61.9)	+0.5	-6.4 to +7.4
Total, 60 and older	75.5 (72.2–78.5)	74.3 (71.7–76.8)	-1.2	-5.3 to +2.9

Round 1 — Source Inspection & Specialist Setup

✉ **[Architectural and Engineering Managers@Computer Hardware Engineers, Electrical Engineers, Mechanical Engineers, Validation Engineers]:** We are assembling the source-traceable Olympus US1-EPYC cooling-resilience package. Please send your specialist-owned artifacts as soon as they are frozen. Management will consume the received evidence as delivered, while Validation waits for the actual Hardware, Electrical, and Mechanical handoffs before closing reliability.

✉ **[Validation Engineers@Computer Hardware Engineers, Electrical Engineers, Mechanical Engineers]:** Please send the current-run specialist handoffs for the `N25_6` versus `F25_5F_pos2` screening. Reliability will not finalize `fan_failure_fmeca.csv` or `reliability_assessment.md` until the Mechanical CFD bundle is actually received. Do not substitute estimates or reconstruct missing CFD QOIs.

[Electrical Engineers ACTION] Extract source-backed fan-control evidence from the schematic and motherboard documents

```
pdftotext -layout data/.../US1-EPYC_MB_G_SCH.pdf work/sch.txt
pdftotext -layout data/.../Project_Olympus_AMD_EPYC_Motherboard.pdf work/mb.txt

grep -n "TACH\|PWM\|FAN" work/sch.txt | head
```

✉ **[Electrical Engineers@Validation Engineers, Managers, Hardware, Mechanical]:** Electrical handoff complete. Six 1U TACH paths are source-traced, but the 1U PWM command is shared/grouped rather than six verified independent commands. The `fan_pos#` → `FAN#` mapping remains unverified, and TACH connectivity is not zero-airflow detection or verified firmware FDIR.

✉ **[Computer Hardware Engineers@Managers, Validation Engineers]:** The source requirement is broader than this selected screen. Mechanical Specification Section 7 calls for maximum-power failed-fan qualification with one fan failed (*N+1*, five fans running), testing all fan locations with worst-case locations as a minimum; four-fan operation is a separate *N* condition with only minimum component requirements. The selected `fan_pos2` CFD screen is not that qualification test.

[Mechanical Engineers ACTION] Probe task-visible compute resources and freeze mesh identity

```
python3 -c "import os; a=os.sched_getaffinity(0); print('count',len(a))"
nproc
sha256sum data/.../provided_mesh.tar.gz
```

✉ **[Computer Hardware Engineers@Electrical Engineers, Mechanical Engineers, Validation Engineers, Managers]:** Hardware screening artifacts are frozen. `N25_6` keeps all six geometric fans active; `F25_5F_pos2` sets only `fan_pos2` to zero flow while the other five retain the normal per-active-fan screening flow. The 2×180 W CPU, 80 W DIMM, $25°$C inlet, and 158 CFM/kW basis are screening assumptions/derivations, not measured nominal performance. Geometric `fan_pos#` is not a verified electrical `FAN#` mapping.

[Validation Engineers ACTION] Inspect the assigned FMECA guidance before constructing reliability artifacts

```
pdftotext -layout data/.../GSFC-HDBK-8004_FMECA.pdf /tmp/opencode/gsfc.txt
grep -n "severity\|failure mode\|critical" /tmp/opencode/gsfc.txt | head -n 20
```

✉ **[Architectural and Engineering Managers@Hardware, Electrical, Validation]:** Hardware and Electrical evidence is now source-traced. Management will preserve the distinction between source facts, benchmark assumptions, current-run observations, and unresolved gaps. Mechanical CFD remains the gating dependency before Reliability can close the selected failure assessment.

Round 2 — CFD Diagnosis, Production & Mechanical Handoff

✉ **[Validation Engineers@Mechanical Engineers, Managers]:** Hardware and Electrical current-run handoffs are in hand, but Reliability is still blocked on Mechanical evidence. PARTIAL / INCOMPLETE is acceptable if that is the actual outcome; we will preserve the stated limitations rather than substitute a 5/6-flow estimate, reconstruct QOIs, or infer qualification from solver completion.

[Mechanical Engineers ACTION] Verify the frozen mesh and CFD execution basis

```
tar -xzf data/.../provided_mesh.tar.gz -C work/mesh_inspect
head -c 2000 work/mesh_inspect/constant/polyMesh/cellZones
checkMesh -allTopology -allGeometry

which simpleFoam
which scalarTransportFoam
```

✉ **[Mechanical Engineers@Managers, Validation Engineers]:** The supplied mesh is treated as immutable. I will use the observed task CPU budget rather than assume a fixed rank count, run steady RANS airflow first, freeze the accepted velocity field, and then run one-way passive air-side thermal transport. The accepted solver logs, resource probe, and QOI provenance will be preserved in the Mechanical bundle.

[Mechanical Engineers ACTION] Run a bounded 16-rank airflow diagnostic

```
rc, wall, pid = mod.run_cmd(
    mod.mpirun_cmd(env, 16, "simpleFoam"),
    cwd=case,
    env=env,
    log_path=logp,
    timeout=180,
)
print("rc", rc, "wall", wall,
      "nadv", mod.count_time_equals(logp),
      "health", mod.last_residual_health(logp))
```

✉ **[Architectural and Engineering Managers@Mechanical Engineers, Validation Engineers]:** Use the bounded diagnostic only to establish a defensible production setup. If the numerical gate remains incomplete within the allowed budget, preserve the failure/partial evidence and move downstream with explicit limitations. Do not extend the run indefinitely or turn numerical completion into qualification.

Round 2 — CFD Diagnosis, Production & Mechanical Handoff (continued)

[Mechanical Engineers ACTION] Execute the bounded production CFD campaign

```
python3 work/run_cfd_campaign.py

# N25_6 airflow -> F25_5F_pos2 airflow
# freeze accepted U
# N25_6 thermal -> F25_5F_pos2 thermal
# simpleFoam -> scalarTransportFoam
```

✉ **[Mechanical Engineers@Validation Engineers, Managers]:** Mechanical specialist handoff frozen: `cfd_case_bundle.tar.gz`, `cfd_execution_summary.csv`, and `cfd_qoi.csv`. OpenFOAM v2406 used the frozen 1,341,597-cell mesh with no remeshing. The resource probe observed 56 CPUs and the accepted production rows used 16 MPI ranks sequentially. All four rows remain PARTIAL at Time=300; do not upgrade them to PASS.

✉ **[Mechanical Engineers@Validation Engineers, Managers]:** Current-run total fan flow changes from **0.03281** to **0.02734** m^3/s under `F25_5F_pos2`. The passive-air screen increases CPU0 plume temperature by about **+3.56 K**, CPU1 by **+1.23 K**, and DIMM plume temperature by **+2.42 K**; bulk-outlet ΔT rises from 11.30 K to 13.54 K. These are screening observations: passive-air/plume temperature is not component/junction temperature.

✉ **[Mechanical Engineers@Validation Engineers, Managers]:** Material model limitations remain visible: the fan boundaries impose volumetric flow rather than a measured fan P–Q relation; thermal transport is one-way passive air-side screening rather than component/conjugate heat transfer; and the five surviving fans retain their assigned per-fan flow rather than a firmware/P–Q redistribution. Return code 0 and extracted QOIs therefore do not constitute qualification.

[Mechanical Engineers ACTION] Inspect the accepted execution summary and QOI bundle before handoff

```
cat _atle_output/cfd_execution_summary.csv
cat _atle_output/cfd_qoi.csv

tar -tzf _atle_output/cfd_case_bundle.tar.gz | \
  grep -E "resource_probe|diagnostic_attempts|thermalSources|log\."
```

✉ **[Mechanical Engineers@Managers, Validation Engineers]:** The bundle retains the current-run resource probe, diagnostic ledger, accepted solver logs, frozen mesh identity, thermal-source definition, and raw/derived QOI evidence. Reported final samples remain traceable to preserved artifacts rather than to an earlier or reconstructed segment. This is a bounded screening record, not a fan-performance or component-temperature qualification package.

✉ **[Architectural and Engineering Managers@Mechanical Engineers, Validation Engineers]:** Mechanical evidence is now received and will be preserved as delivered. Validation may finalize Reliability from the actual Hardware, Electrical, and Mechanical handoffs. Keep the selected `fan_pos2` case distinct from all-fan/worst-location qualification, and do not invent occurrence rates or RPN.

Round 3 — Reliability Integration & Final Evidence Audit

[Validation Engineers ACTION] Construct the evidence-bounded FMECA

```
row = {
    "fmeca_mode_id": "REL-ZF-POS2",
    "likelihood_of_occurrence":
        "NOT_QUANTIFIED / UNVERIFIED",
    "rpn": "N/A",
}
writer.writerow(row)
```

✉ **[Validation Engineers@Managers, Hardware, Electrical, Mechanical]:** Reliability/Validation handoff frozen: `fan_failure_fmeca.csv` and `reliability_assessment.md`. Occurrence remains NOT_QUANTIFIED / UNVERIFIED and `rpn=N/A`; Mechanical PARTIAL evidence is interpreted as-is rather than upgraded. TACH is not verified airflow detection/FDIR, shared PWM is not six independent commands, and selected `fan_pos2` is not an all-fan qualification result.

✉ **[Validation Engineers@Managers]:** The remaining qualification gaps are explicit: the geometric-to-electrical fan mapping is unverified, firmware zero-airflow response/FDIR is not demonstrated, passive-air temperatures do not establish component/junction limits, and no quantitative occurrence basis or all-fan/worst-location test evidence is available. These gaps require new source/test evidence rather than narrative closure.

✉ **[Architectural and Engineering Managers@Validation Engineers]:** Reliability handoff received. Management will preserve specialist-owned bytes, retain the PARTIAL CFD status and scope limitations, and carry the unresolved qualification gaps into the final decision rather than silently resolving them.

✉ **[Architectural and Engineering Managers@All]:** Qualification closure now requires evidence-producing work, not narrative cleanup: instrumented hardware thermal tests must measure component/case/junction temperatures against vendor limits, and a chassis/harness continuity map must connect geometric `fan_pos0..5` to the J77 TACH/PWM/power pins. Firmware evidence must then show the actual TACH sampling, stall/missing-pulse criteria, detection latency, event reporting, and zone-control response.

✉ **[Architectural and Engineering Managers@All]:** The selected-location CFD screen must also be complemented by failed-fan testing across all fan locations, including an identified worst-case location and the required five-fan/four-fan operating conditions. Installed fan modules need vendor P–Q/RPM data or measured chassis airflow; a numeric occurrence score or RPN remains prohibited until field, life-test, or another defensible occurrence basis exists.

Round 3 — Reliability Integration & Final Evidence Audit (continued)

[Architectural and Engineering Managers ACTION] Audit the integrated package before release

```
assert archived_html == html
assert archived_png == png

for artifact in manifest["artifacts"]:
    data = tf.extractfile(artifact["path_in_archive"]).read()
    assert hashlib.sha256(data).hexdigest() == artifact["sha256"]
    assert len(data) == artifact["bytes"]

assert decision["disposition"] == "ADDITIONAL_QUALIFICATION_REQUIRED"
print("INTEGRITY AUDIT PASS")
```

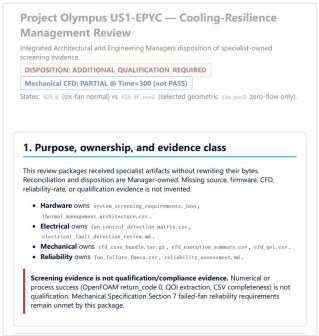
✉ **[Architectural and Engineering Managers@All]:** The final release set is internally consistent: `Olympus_Cooling_Resilience_Review.html` provides the integrated management review; `Olympus_Cooling_Resilience_Summary.png` presents the executive two-state comparison; and `Olympus_Cooling_Resilience_Evidence.tar.gz` preserves the specialist evidence package. PARTIAL numerical status and material limitations remain visible.

✉ **[Architectural and Engineering Managers@All]:** Integrated Olympus US1-EPYC cooling-resilience package complete. The report, summary figure, decision record, manifest, specialist handoffs, and nested CFD evidence remain mutually consistent and traceable to current-run artifacts. Engineering disposition: ADDITIONAL QUALIFICATION REQUIRED.

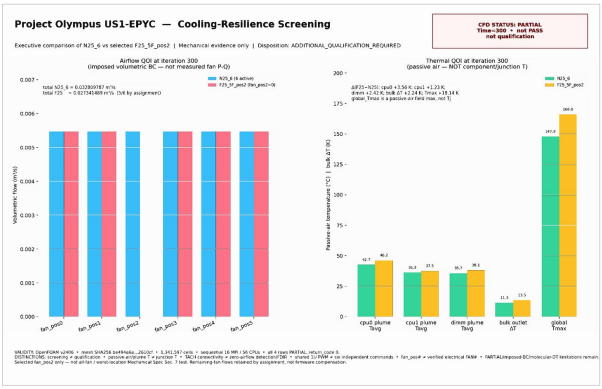
CFD

Selected Outputs from the Final Artifact

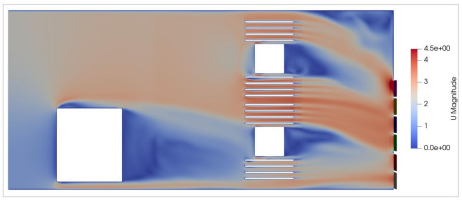
(a) Integrated management review



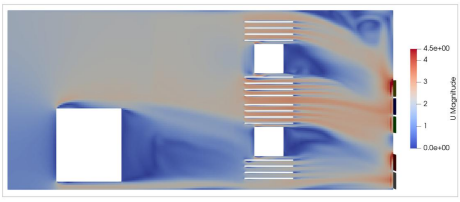
(b) Executive cooling-resilience summary



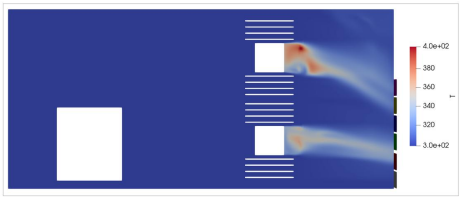
(c) N25_6 airflow field



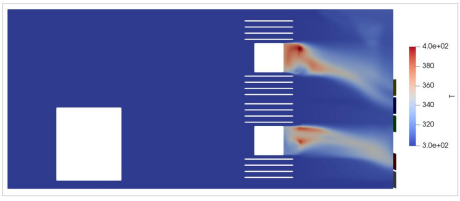
(d) F25_5F_pos2 airflow field



(e) N25_6 passive-air temperature field



(f) F25_5F_pos2 passive-air temperature field



BUSINESS

Selected Outputs from the Final Artifact

(a) Competitive portfolio dashboard overview

INTERNAL CONSUMER-PRODUCTS PORTFOLIO VIEW

Mascara Competitive Portfolio Dashboard

Ten Amazon Reviews'23 All Beauty mascara records. Product-page declarations and buyer review observations are displayed separately; snapshot commercial metrics are descriptive and are not sales or revenue estimates.

Portfolio Summary

10	3	6	1
parent ASINs reviewed	supported claim-review alignments	mixed claim-review alignments	contradicted color proposition

(b) Representative product comparison

Product	Price	Avg Rating	Rating Count
 B08TCBP8HV essence | Disney Princess Collection | Lash Princess Snow White False Lash Effect Mascara | Cruelty Free (Pack of 1 - Disney Princess Collection)	$6.90	4.6	3,859
 B00FZH4126 "Mirenesse" DUO Secret Weapon 24hr Mascara Super Long Black - Winner Best Cult Mascara 10g / 0.35oz	$38.93	4.0	121

(c) Evidence reconciliation audit

Evidence Reconciliation Audit

Mandatory dispute resolution: the alignment describes the relationship between product-page positioning and the cited buyer observations, not an independent performance finding.

Parent ASIN	Alignment	Declared product-page proposition
B0B7L451Z9	MIXED	Declared waterproof wear and extreme/mega volume.
B08MRRNL18	MIXED	Declared suitable for sensitive eyes/non-irritation positioning.
B07WJHWDSD	MIXED	Declared water-resistant, smudge-free, and flake-free finish.
B0771FD7LB	CONTRADICTED	Declared purple-color proposition.

(d) Portfolio decision ledger

Portfolio Decision Ledger

Each entry compares both named products on snapshot price, average rating, rating count, and cited buyer evidence. The entries are portfolio recommendations, not performance or safety conclusions.

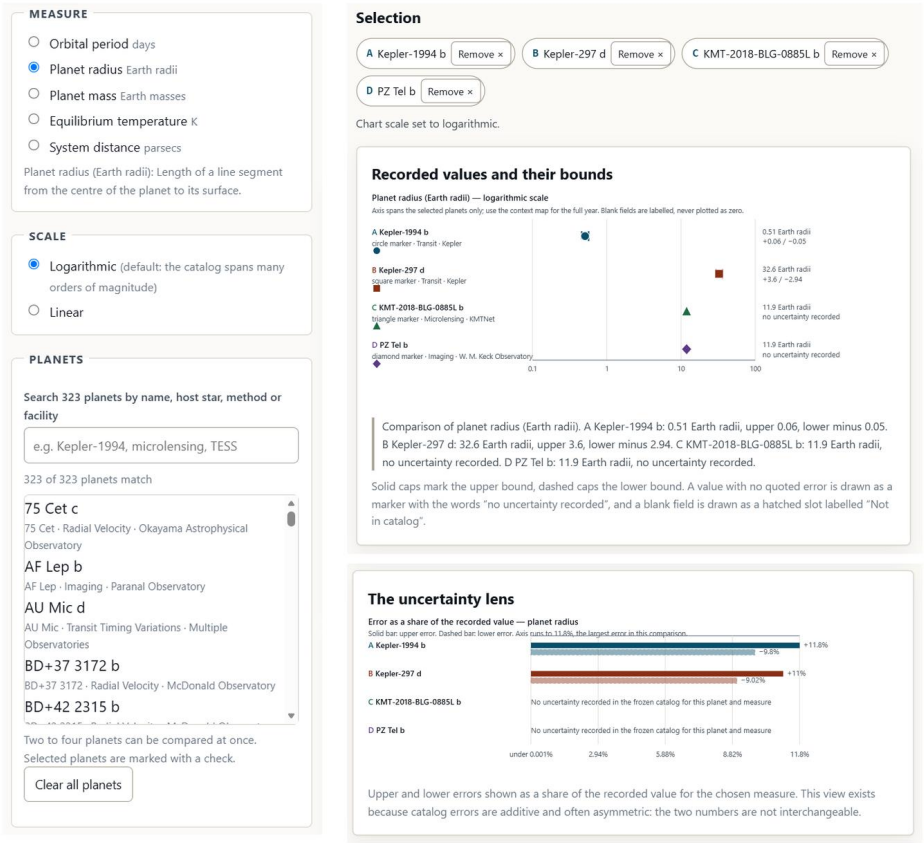
Action	Comparison and decision basis
SCALE	**B08TCBP8HV vs B00FZH4126** Prioritize the lower-priced B08TCBP8HV for scale testing: it has a higher average rating and a much larger rating count than B00FZH4126. Its false-lash-effect title positioning is supported by fuller/longer and false-lash-like buyer observations, with caking noted as a qualification.
REFORMULATE	**B0B7L451Z9 vs B08TCBP8HV** Reformulate B087L451Z9 before broader emphasis: its waterproof observation is favorable, but volume and clumping observations are mixed. B08TCBP8HV provides the higher-rating reference with favorable false-lash-like observations, while retaining a caking qualification.
VALIDATE	**B08MRRNL18 vs B07WJHWDSD** Validate both mixed propositions with focused follow-up: B08MRRNL18 has competing sensitive-eye observations, while B07WJHWDSD has both favorable wear and flaking/raccoon-eye observations. Their equal average ratings make the verification need more informative than a simple rating comparison.
DO NOT PRIORITIZE	**B0771FD7LB vs B075L1VSRS** Do not prioritize B0771FD7LB ahead of B075L1VSRS: the purple-color proposition is contradicted by black-color and poor-visibility observations. B075L1VSRS is also mixed, but retains a favorable volume/length/curl observation alongside travel-size/value and dried-out qualifications.

MEDIA

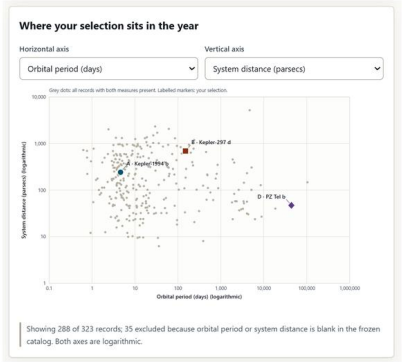
Selected Outputs from the Final Artifact



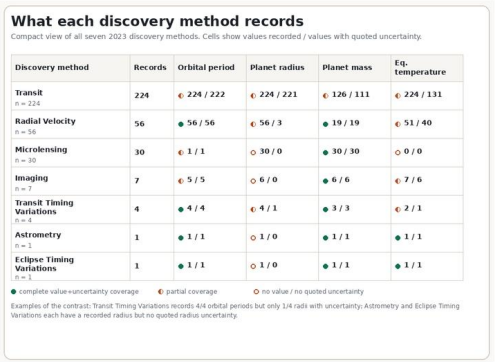
(a) Interactive planet comparison



(b) Where your selection sits in the year



(c) What each discovery method records



FINANCE

Selected Outputs from the Final Artifact

(a) Candidate selection and frozen choice

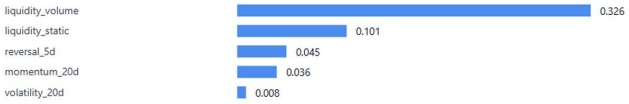
A 100+ configuration search was run on released visible data. Representative finalists below use the same SIC / 4 / 0.08 construction.

candidate	visible	3σ-robust	Sharpe (robust)	turnover	max DD	sub-univ. Sharpe	max baseline |corr|
`zscore(ts_std_dev(close,20)/close) (single factor)`	79.19	88.26	1.50	0.071	17.1%	0.22	0.47
`zscore(ts_std_dev(close,20)/close) + zscore(ts_std_dev((high-low)/close,20))`	89.79	90.36	1.49	0.059	15.0%	0.36	0.36
`zscore(ts_std_dev((high-low)/close,20)) - zscore(ts_delta(close,21)/close)`	80.57	91.92	1.31	0.120	14.6%	0.48	0.34
frozen composite (this submission)	**82.29**	**91.33**	**1.43**	**0.094**	**14.7%**	**0.44**	**0.33**

(b) Self-correlation audit

Each baseline is rebuilt under *this submission's* neutralization / decay / truncation (`sic` / `4` / `0.08`) exactly as the evaluator does; the comparison is between daily net book returns.

baseline	corr (visible 2012-2020)	corr (train 2012-2019)	corr (2020)	max |rolling 252-session corr|	share of windows |corr|>0.70
liquidity_volume	+0.3263	+0.2864	+0.9323	0.9361	12.9%
liquidity_static	-0.1010	+0.3610	-0.5515	0.8428	11.4%
reversal_5d	+0.0454	+0.2877	+0.0783	0.5551	0.0%
momentum_20d	-0.0365	-0.1995	-0.0684	0.5637	0.0%
volatility_20d	+0.0081	+0.0938	+0.0007	0.5249	0.0%



Max |corr| on the visible window = 0.3263 (liquidity_volume), penalty multiplier 1.0000; the fade only starts above |corr| 0.70. The rolling column is the honest stress view: in 2014 and 2020 the book briefly converges with the volume-liquidity baseline.

(c) Visible book behaviour

4.1 Log equity of the dollar-neutral, delayed, cost-charged book (2012-2020, log scale)



4.2 Drawdown (%)



LEGAL

Selected Outputs from the Final Artifact

(a) Boundary by maritime zone

The judgment fixes one single maritime boundary with distinct zone-specific components. The legal paths below are geodetic source-defined relationships. Their coordinates are control points, not permission to substitute a Cartesian chord for the judgment's legal line.

Component and order	Zone and legal effect	Source-supported implementation
TS_MEDIAN_LINE: START -> TP1 -> TP2 -> TP3 -> TP4 -> TP5 -> TP6 -> A	Territorial sea. From START, the boundary follows the median line described in S01 paragraph 117 until the 12-nautical-mile limit at A.	S01 paragraph 117; operative paragraph 214(3); WGS 84 geodetic rule in footnote 1 following paragraph 98.
EEZ_CS_TO_200NM: A -> B	Exclusive economic zone and continental shelf up to 200 nautical miles. From A, the boundary follows a geodetic line beginning with initial azimuth 114 deg to B.	B is the intersection with the 200-nautical-mile limit measured from the baselines from which Kenya's territorial-sea breadth is measured. S01 paragraph 174; operative paragraph 214(4).
CS_BEYOND_B: B -> conditional termination	Continental shelf beyond 200 nautical miles. The boundary continues along the same geodetic line as the adjusted within-200-nm line.	Termination is conditional: outer limits of the Parties' continental shelves, delineated on the basis of Commission recommendations, or an area where third-State rights may be affected. S01 paragraph 196; operative paragraph 214(5).

(b) Operative disposition and votes

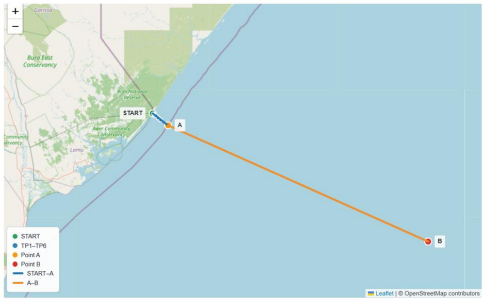
The following is a neutral implementation transcription of S01 operative paragraph 214. The operative result is kept separate from merits reasoning and from pleaded conduct. Votes are recorded as stated in the English text, which S01 says is authoritative.

Item	Vote and operative result	Implementation effect
OP1	Unanimously. Finds no agreed maritime boundary between Somalia and Kenya following the parallel of latitude described in paragraph 35.	Exclude that parallel from the operative boundary record.
OP2	Unanimously. Decides the starting point is the intersection of the straight line from PB 29 at right angles to the general coast direction with the low-water line, at the stated WGS 84 point.	Record START and its source-defined origin.
OP3	Unanimously. Decides that from START the territorial-sea boundary follows the median line in paragraph 117 to the 12-nautical-mile limit at A.	Record TS_MEDIAN_LINE through TP1-TP6 to A.
OP4	Ten votes to four. Decides that from A the single boundary for the EEZ and continental shelf up to 200 nautical miles follows the geodetic line starting with azimuth 114 deg to B at the Kenyan-baseline-referenced limit.	Record EEZ_CS_TO_200NM A to B, with azimuth and Point B basis.
OP5	Nine votes to five. Decides that from B the continental-shelf boundary continues along the same geodetic line to the outer shelf limits or an area where third-State rights may be affected.	Record conditional continuation; no finite post-B endpoint.
OP6	Unanimously. Rejects the claim made by Somalia in final submission number 4.	Record the disposition distinctly from the allegation and paragraph 212 merits reasoning.

(c) Frozen coordinate register

	A	B	C	D	E	F
1	Point_ID	Order	DMS_Lat	DMS_Lon	Decimal_Lat	Decimal_Lon
2	START	1	1° 39' 44.0" S	41° 33' 34.4" E	-1.662222222	41.55955556
3	TP1	2	1° 40' 18.3" S	41° 34' 17.4" E	-1.67175	41.5715
4	TP2	3	1° 40' 32.1" S	41° 34' 32.8" E	-1.675583333	41.57577778
5	TP3	4	1° 41' 12.8" S	41° 35' 22.8" E	-1.686888889	41.58966667
6	TP4	5	1° 41' 39.0" S	41° 36' 00.9" E	-1.694166667	41.60025
7	TP5	6	1° 42' 39.9" S	41° 37' 21.6" E	-1.711083333	41.62266667
8	TP6	7	1° 44' 01.2" S	41° 39' 02.8" E	-1.733666667	41.65077778
	A	8	1° 47' 39.1" S	41° 43' 46.8" E	-1.794194444	41.72966667

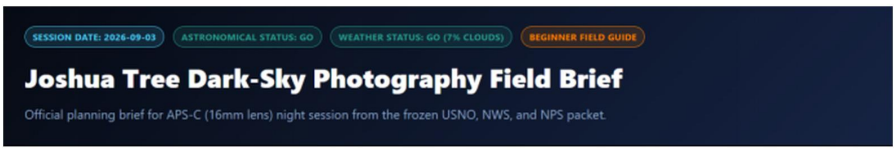
(d) GeoJSON boundary visualization



TOURISM

Selected Outputs from the Final Artifact

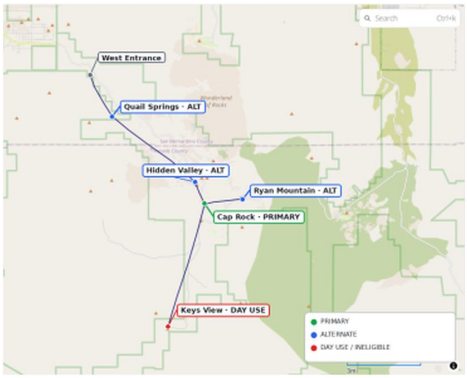
(a) Dark-sky photography field brief



(b) Integrated night briefing map



(c) Route and site eligibility map



(d) Night execution schedule

SEQ	TIME WINDOW	REF ID	ITEM TYPE	SITE	TECHNIQUE
1	19:15 - 19:35	SETUP-01	Arrival & Setup	CAP-ROCK	red_light_gear_setup
2	19:35 - 20:15	SHOT-01	Photo Capture	CAP-ROCK	wide_field_milky_way
3	20:15 - 20:30	REST-01	Review & Rest	CAP-ROCK	dark_adaptation_rest
4	20:30 - 21:15	SHOT-02	Photo Capture	CAP-ROCK	deep_sky_star_field
5	21:15 - 21:30	PREP-02	Reframe / Prep	CAP-ROCK	silhouette_framing
6	21:30 - 22:15	SHOT-03	Photo Capture	CAP-ROCK	ambient_silhouette
7	22:15 - 22:30	PACK-01	Packout & Exit	CAP-ROCK	leave_no_trace_packout

RESTAURANT

Selected Outputs from the Final Artifact

(a) Generated seasonal launch poster



(b) Integrated campaign plan

01 **EXECUTIVE DIRECTION**

A measurable trial designed to learn both demand and store readiness.

Launch two seasonal drinks across four pilot stores, make the choice easy at the counter, and use a short test window to decide whether the menu can scale. The campaign should earn trial without turning a controlled recipe into a nutrition, allergen, wellness, or superiority promise.

OBJECTIVE	AUDIENCE	DECISION
Drive qualified seasonal trial and capture repeat/occasion signals.	Existing guests plus nearby coffee purchasers during seasonal break occasions.	Scale, refine, or stop based on sales, service time, quality, waste, and feedback.

Scenario facts

Item	Authoritative value	How it is used
Pilot window	2026-10-05 to 2026-11-02	All channel and reporting dates
Footprint	4 pilot stores	Store-level rollout and learning
Products	Citrus Oat Cold Brew; Cocoa Spice Latte	Names only; ingredients not inferred
Prices	56,000 VND; 59,000 VND	POS, menu and creative checks
Core comparator	Americano; 12 stores; 39,000 VND	Context only; no forecast
Recipe status	controlled_recipe_not_certification	Claims and disclosure guardrail

(c) Audience-evidence workbook excerpt

Controlled Scenario Facts

Northstar Coffee facts and transparent derivations. Scenario facts are authoritative; derivations are arithmetic only.

Record ID	Fact type	Field	Value	Unit
S-001	Campaign	Campaign window	2026-10-05 to 2026-11-02	date range
S-002	Offer	Citrus Oat Cold Brew	56,000	VND
S-003	Offer	Cocoa Spice Latte	59,000	VND
S-004	Benchmark menu	Core Americano	39,000	VND
S-005	Add-on	Seasonal topping	9,000	VND
S-006	Distribution	Pilot store count	4	stores
S-007	Distribution	Core Americano store count	12	stores

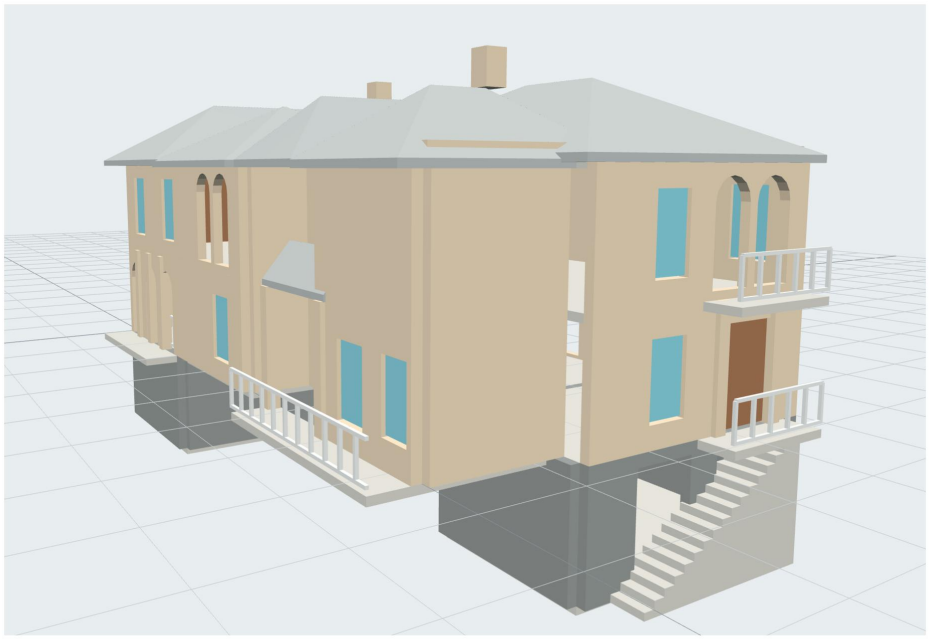
(d) Generated promotional video



BUILDING

Selected Outputs from the Final Artifact

(a) Integrated BIM model



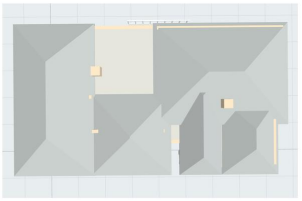
(b) Main elevation



(c) Opposite elevation



(d) Roof and massing overview



EDUCATION

Selected Outputs from the Final Artifact

Attendance-Area Recommendation — HCPSS New High School 13 (Guilford Park HS)

Planner-owned proposal for all 710 frozen planning polygons · Effective 2023–2024 · Steady state 2025–2026 · Offline, self-contained artifact

(a) Attendance-area recommendation summary

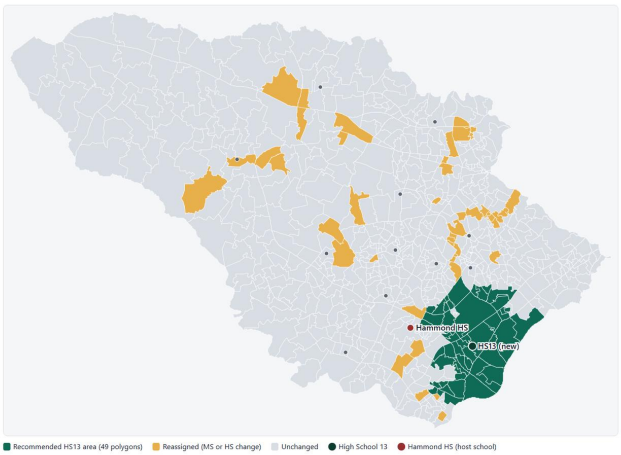
Open High School 13 at its Mission Road / Washington Boulevard site in Jessup as a walkable-to-transit-served neighbourhood school for the Route 1 corridor, and draw its attendance area from the high schools closest to it. The recommended area takes **49 contiguous planning polygons** from Hammond HS (25), Oakland Mills HS (8), Reservoir HS (7), Atholton HS (5) and Long Reach HS (4), plus small adjacency-preserving transfers among Mt Hebron, Howard, Oakland Mills and Wilde Lake HS to bring every high school under the 110% alert ceiling. Middle-school assignments are adjusted by **31 polygons** to consolidate feeders and reduce below-floor capacity.

HS13 OPENING ENROLLMENT	HS13 OPENING UTILIZATION	HS13 STEADY STATE 2025-26	HIGH SCHOOLS OVER 110%
1,578	**95.17%**	**100.12%**	**0**
students visible in 2023-24	1,658-seat capacity; 90–100% target band	inside the 90–110% alert band	was four under the frozen current plan

POLYGONS REASSIGNED	HARD CONSTRAINTS
65 HS / 31 MS	**11 / 11 PASS**
limits 150 HS / 50 MS	2023-24 gate year

Elementary-school assignments are **unchanged and are not proposed here** (forbidden by the scope rules). The proposal is analytically complete and internally consistent, but it is deliberately published as a recommendation for independent review rather than as an adopted plan.

(b) Recommended attendance-area boundaries



(c) Hard-constraint compliance

Constraint	Limit	Observed	Result
Max connected components per school	3	3	PASS
Max high schools above 110%	1	0	PASS
Max middle schools above 110%	1	0	PASS
Max high schools below 75%	1	1	PASS
Max middle schools below 75%	2	1	PASS
Min HS13 utilization	0.85	0.9517	PASS
Max HS13 utilization	1.1	0.9517	PASS
Max low-feed cells (<15%)	5	4	PASS
Max student-weighted proxy minutes	7.694	6.9302	PASS
Max reassigned high-school polygons	150	65	PASS
Max reassigned middle-school polygons	50	31	PASS

(d) School-level capacity outcomes

High school	Capacity	Frozen current (util.)	Recommended (util.)	Band	Polygons
Atholton HS	1,460	1,544 (105.75%)	1,475 (101.03%)	100-110% alert band	47
Centennial HS	1,360	1,363 (100.22%)	1,363 (100.22%)	100-110% alert band	45
Glenelg HS	1,420	1,142 (80.42%)	1,142 (80.42%)	75-90% (below target)	88
Hammond HS	1,420	1,365 (96.13%)	665 (46.83%)	below the 75% floor	25
High School 13	1,658	— (new school)	1,578 (95.17%)	90-100% target band	49
Howard HS	1,420	1,625 (114.44%)	1,534 (108.03%)	100-110% alert band	59
Long Reach HS	1,488	1,711 (114.99%)	1,513 (101.68%)	100-110% alert band	58
Marriotts Ridge HS	1,615	1,609 (99.63%)	1,609 (99.63%)	90-100% target band	67
Mt Hebron HS	1,400	1,637 (116.93%)	1,498 (107.00%)	100-110% alert band	43
Oakland Mills HS	1,400	1,419 (101.36%)	1,259 (89.93%)	75-90% (below target)	47
Reservoir HS	1,551	1,928 (124.31%)	1,586 (102.26%)	100-110% alert band	48
River Hill HS	1,488	1,400 (94.09%)	1,400 (94.09%)	90-100% target band	60
Wilde Lake HS	1,424	1,355 (95.15%)	1,476 (103.65%)	100-110% alert band	74

District-wide high-school utilization falls from 103.74% (18,098 visible students against 17,446 existing seats) to 94.73% (18,098 against 19,104 seats) once HS13 is in service — the system moves from structural overcrowding to roughly 5% reserve capacity. Middle-school district utilization is 87.68% in 2023-24 and 90.13% in 2025-26.

INDUSTRIAL

Selected Outputs from the Final Artifact

(a) Best-observed case summary

PAPER SHOWCASE · SELECTED BEST OBSERVED CASE

USB-C sensor node from a blank PCB

A five-occupation team converts frozen electrical, component, mechanical, and manufacturing contracts into a routed two-layer KiCad board and a byte-traceable release package.

Task v14.0.0 | 5 occupations | 3 → 1 → 1 workflow | Offline · CLI-only | Selected run: 2026-09-13

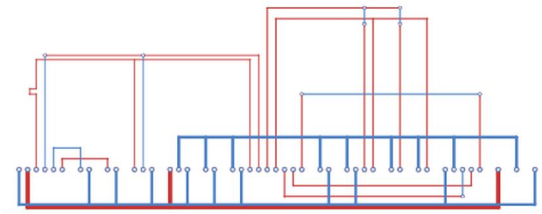
AUDITED WHOLE-TASK RESULT

22/22

PASS · all five role checks passed

DRC errors	0
Unconnected items	0
Warnings retained	75
Board hash	f44905f0...

(b) Direct KiCad copper export



(c) Board-level metrics

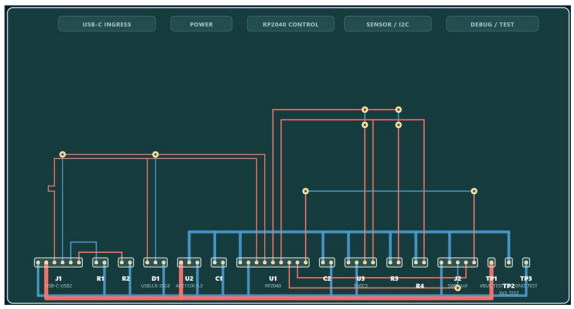
110 × 58 millimetre board outline	**2** copper layers
15 contracted footprints	**12** required routed nets
63 track segments	**9** through vias

F.Cu · 32 segments

B.Cu · 31 segments

Via · 9 transitions

(d) Accepted PCB geometry



(e) Engineering verification evidence

Check ID	Category	Status	Notes
board_parse	KiCad Geometry	PASS	Clean parse: 15 footprints, 63 segments, 9 vias
drc_clean	DRC Verification	PASS	KiCad 9.0.9 CLI DRC: 0 errors, 0 unconnected
required_net_routing	Routing Completeness	PASS	All 12 required signal/power nets routed
track_widths	Trace Widths	PASS	All segments meet effective minimum widths (incl. 3V3=0.56mm)
routing_obstacles	Routing Obstacles	PASS	All 26 keepout obstacles cleared with required margins
testpoint_coverage	Specialized Check	PASS	TP1 (VBUS), TP2 (3V3), TP3 (GND) coverage confirmed
usb_pair_gap	Specialized Check	PASS	Coupled route window coverage >= 95% for differential pairs
usb_pair_skew	Specialized Check	PASS	USB_DP and USB_DM length skew <= 0.25 mm (0.000 mm)

G Error Analysis

Table 11: Taxonomy of multi-agent failure modes in FATE.

Failure type	Definition
Communication coordination	
Misrouting	Required information is sent to agents that do not depend on it, while agents whose outputs directly require the information are omitted.
Ambiguous handoff	A message is received but does not clearly specify the authoritative artifact, its meaning, the required action, or how competing artifacts should be reconciled.
Unreported blockage	An agent detects that it cannot deliver a required result but fails to notify affected downstream agents, communicate the missing content, or propose a fallback.
Stale-version propagation	An artifact, fact, or constraint is updated, but dependent agents continue using an obsolete version because the authoritative update is not communicated.
Dependency-timing violation	A downstream agent starts or completes dependency-sensitive work before the required upstream input is available, causing rework or incorrect outputs.
Output-interface misalignment	
Delivery-location or naming mismatch	A task-relevant artifact is created but placed outside the required submission directory or saved under a filename that does not match the task contract.
Missing or unusable deliverable	One or more required deliverables are absent, or appear but cannot be consumed because their actual contents or representation are invalid.
Response-format violation	An agent fails to follow a required control-message format.
State misjudgment	
Phantom-artifact assumption	After a command, tool call, or parsing step fails, an agent continues as though the intended file, text, or data had been successfully produced.
Mechanical retry	An agent repeats a materially identical command, API call, or submission despite unchanged failure conditions, without repairing the precondition, changing strategy, or reporting the blockage.
Completion-state confusion	An agent treats a local action, such as writing a file, sending a message, or returning `FINISH`, as verified readiness or team-level completion.
Role-boundary violation	
Unauthorized responsibility reassignment	An agent assigns another role ownership of work or data extraction without confirming that the role has the necessary materials, tools, or responsibility.
Duplicate authoritative artifact	A non-owner independently produces and sends a competing artifact as authoritative, creating conflicting versions without authorization or confirmed delegation.

To understand why multi-agent executions fail despite substantial local progress, we manually analyzed unsuccessful trajectories and categorized recurrent failures into the thirteen types summarized in Table 11. The taxonomy distinguishes failures in communication coordination, output-interface misalignment, state misjudgment, and role-boundary violation.

G.1 COMMUNICATION COORDINATION

Misrouting. Misrouting occurs when an agent sends information to agents that do not require it while omitting agents whose work directly depends on that information. For example, in a roadmap-generation task, a coordinator distributed global evidence constraints and integration rules to several report-oriented agents but omitted multiple domain analysts responsible for producing factual content. As a result, different parts of the team operated under inconsistent assumptions about terminology, evidence strength, and prohibited inferences. The message recipients did not match the task dependency graph.

Ambiguous handoff. Ambiguous handoff arises when a recipient receives a message but cannot determine what has changed or what action is required. In a roadmap workflow, the Analyst sent the Architect and Editor an update stating that the Autonomous Systems and Avionics overlay language had been tightened and that they should use the latest file. The notice did not identify the affected candidate or overlay records, specify the old and new wording, indicate which PNG nodes or edges required revision, or provide a version identifier, timestamp, or summary hash. Because the PDF and PNG were already under production, the recipients could not reliably determine the update's scope from the message itself. The resulting ambiguity could leave textual changes incorporated while the corresponding visual relationships remained unchanged.

Unreported blockage. Unreported blockage occurs when an agent recognizes that it cannot complete a required deliverable but does not convert this local failure into actionable risk information for dependent agents. In a technical analysis task, an agent encountered an unavailable workspace capability and could neither create nor validate its required evidence artifact. However, it terminated without notifying the report and visualization agents that its section would be unavailable. Consequently, downstream agents continued under the assumption that the missing input would arrive, turning a local execution problem into a team-wide planning and delivery failure.

Stale-version propagation. Stale-version propagation occurs when an artifact or constraint has been updated but dependent agents continue to use an obsolete version. In one trajectory, an analyst revised a traceability artifact to refine factual boundaries and then communicated the update to a visualization agent, but not to the report-writing agent that had received the earlier version. The resulting outputs were based on different versions of the same evidence, creating inconsistencies between narrative content and visual representations.

Dependency-timing violation. Dependency-timing violations occur when a downstream agent begins dependency-sensitive work before the required upstream input is available, or receives the input only after completing critical work. For instance, a visualization agent began constructing a final figure before receiving a candidate-selection matrix that specified the entities and relationships to be visualized. When the matrix arrived later, the agent had to revise its work, increasing the risk of version divergence and incomplete incorporation of the updated information. The underlying problem is a failure to coordinate execution order around explicit readiness conditions.

Communication-Coordination Failure Traces

1. Misrouting — global guidance omitted direct dependents.

[Chief Technology Strategist@Report Editor, Roadmap Visualization Architect, Mars Campaign Integration Analyst, Technology Candidate Traceability Analyst]: Shared the global strategy brief and integration constraints, including the mission-to-candidate evidence chain, evidence labels, and prohibited funding or flight-qualification inferences.

▲ **[FAILURE EXPOSED — Misrouting]:** The same guidance was not sent to `Technology Area Taxonomy Analyst`, `Information and Digital Systems Analyst`, `Space Sustainability and Human Exploration Analyst`, or `Autonomous Systems and Avionics Analyst`, although each produced factual report content subject to these constraints.

2. Ambiguous handoff — update notice omitted its actionable change scope.

[Technology Candidate Traceability Analyst@Roadmap Visualization Architect, Technology Roadmap Report Editor]: Initial handoff: `ta_taxonomy.md`, containing candidate IDs, technology-area labels, unavailable rules, and overlay boundaries for the roadmap report and PNG.

[Technology Candidate Traceability Analyst@Roadmap Visualization Architect, Technology Roadmap Report Editor]: Update: "tightened Autonomous Systems and Avionics overlay language . . . use latest file."

▲ **[FAILURE EXPOSED — Ambiguous handoff]:** The update did not identify changed candidate or overlay records, old-to-new wording, affected PNG nodes or edges, an `artifact_version`, or a `required_downstream_action`. The report and visualization roles could not determine the update scope from the notice itself.

3. Unreported blockage — local failure never became team risk information.

[Tool@Autonomous Systems and Avionics Analyst]: The workspace tool channel is unavailable; the role cannot safely create or validate the required domain evidence table.

[TeamHub status]: `finished_without_ready; deliverables: []; ready_confirmed: false.`

▲ **[FAILURE EXPOSED — Unreported blockage]:** No risk notification was sent to `Report Editor`, `Roadmap Visualization Architect`, or the affected analysts. Downstream roles therefore received no indication that the autonomy/AI/avionics evidence artifact would not arrive.

4. Stale-version propagation — updated matrix reached only one consumer.

[Technology Candidate Traceability Analyst@Report Editor]: Sent `technology_traceability_matrix.md` as the authoritative exhibit, with candidate IDs, snapshot fields, unavailable rules, technology-area labels, and crosscutting overlays.

[Technology Candidate Traceability Analyst@Roadmap Visualization Architect]: Sent an updated `technology_traceability_matrix.md`; the Autonomous Systems and Avionics overlay boundary has been tightened. Use the latest file.

▲ **[FAILURE EXPOSED — Stale-version propagation]:** `Report Editor` received the earlier matrix but no updated copy or replacement notice. The report and the roadmap could therefore be produced from different evidence boundaries.

5. Dependency-timing violation — visualization started before its direct input arrived.

[Roadmap Visualization Architect ACTION]: Begins PNG roadmap construction from currently available evidence; the traceability matrix is not yet available to this role.

[Technology Candidate Traceability Analyst@Roadmap Visualization Architect]: Sent the updated `technology_traceability_matrix.md`, which specifies exact candidate IDs, technology-area labels, unavailable rules, and overlay boundaries needed by the PNG roadmap.

▲ **[FAILURE EXPOSED — Dependency-timing violation]:** The visualization had already begun before its direct dependency arrived. Incorporating the late matrix required revision and created a risk that earlier candidate or relationship choices would remain in the final PNG.

G.2 Output-Interface Misalignment

Delivery-location or naming mismatch. Delivery-location or naming mismatches occur when an agent creates a task-relevant artifact but does not expose it through the interface required by the task. In one presentation task, an agent generated the required slide deck in its working directory but left the designated output directory empty. In another workflow, a team produced a valid report and an image-like file, but named the latter as a test artifact rather than using the required final filename. In both cases, substantial production work had been completed, yet the evaluator and downstream consumers could not identify the artifact as a valid submission.

Missing or unusable deliverable. Missing or unusable deliverables occur when a task requires a set of outputs but at least one member of that set is absent or cannot serve its intended function. For example, a planning agent may submit a route specification and a trajectory table but omit the required route visualization. In another case, an agent creates a file with a `.png` extension whose contents are escaped text rather than valid image bytes.

Response-format violation. Response-format violations arise when an agent fails to comply with the structured protocol used by the orchestrator or collaboration system. In one trajectory, the system instructed a role to return exactly one JSON object, with no Markdown, code fences, or explanatory prose. The role responded with a natural-language explanation, multiple JSON code blocks, and an additional candidate object. The controller rejected the response as an invalid control message and repeated the same formatting instruction. On the next turn, the role again added explanatory text and multiple structured fragments before its proposed JSON. Thus, the agent received an explicit protocol correction, yet continued to produce outputs that violated the unique-object requirement and could not be recorded as valid state updates.

Output-Alignment Failure Traces

1. Delivery-location or naming mismatch — completed artifacts remained outside the required interface.

◆ **[TASK CONTRACT]:** Submit `campus_satisfaction_survey_results.pptx` through `single_agent/_atle_output/`.

[Single Agent ACTION]: Creates `campus_satisfaction_survey_results.pptx` in the workspace root.

◆ **[OUTPUT INSPECTION]:** Workspace root contains `campus_satisfaction_survey_results.pptx`; but `single_agent/_atle_output/` is empty.

▲ **[FAILURE EXPOSED — Delivery-location mismatch]:** The required presentation existed but was outside the evaluator-visible submission directory, so the runner recorded the final artifact as missing.

2. Missing or unusable deliverable — required visualization never formed a consumable PNG.

◆ **[TASK CONTRACT]:** The NASA roadmap task requires `nasa_technology_roadmap_report.pdf` and `nasa_technology_roadmap.png` in `_atle_output/`.

[Agent ACTION]: Produces `nasa_technology_roadmap_report.pdf`, `test.png`, `generate_png.js`, and `roadmap_data.json`.

◆ **[ARTIFACT INSPECTION]:** `nasa_technology_roadmap.png` is absent. `test.png` is only 181 bytes and begins with the ASCII text `b'x89PNG r n x1a n x'`, rather than raw PNG signature bytes.

▲ **[FAILURE EXPOSED — Missing or unusable deliverable]:** The required PNG was neither submitted under its contract name nor represented by a parseable image. The PDF, script, and data file could not substitute for the missing visualization artifact.

3. Response-format violation — repeated noncompliance with the control-message protocol.

◆ **[TEAMHUB@TRAVEL GUIDES]:** Return exactly one JSON object. Do not include Markdown, code fences, or additional prose.

[Travel Guides@TeamHub]: Responds with explanations including "Re-evaluating Response Format" and "Correcting JSON Structure," multiple JSON code blocks, and a final candidate JSON object.

◆ **[TEAMHUB]:** The response is not accepted as a valid unique-object control message; the same JSON-only instruction is repeated.

▲ **[FAILURE EXPOSED — Response-format violation]:** The role again emits explanatory text and multiple structured fragments before its proposed JSON, leaving the controller unable to record a valid protocol state update.

G.3 STATE MISJUDGMENT

Phantom-artifact assumption. Phantom-artifact assumptions occur when an agent observes that a command, tool call, or parsing step has failed but continues to reason as if the expected output exists. For example, after a document-extraction command was unavailable, an agent attempted to read the text files that the failed command was supposed to create and proceeded with source analysis. The agent had observed evidence that the extraction failed, but did not update its working state to

mark the source as unavailable. This error can turn a local execution failure into unsupported downstream reasoning, because later agents may receive conclusions that appear grounded in evidence that was never successfully obtained.

Mechanical retry. Mechanical retry occurs when an agent repeats an action after the system has made clear that its preconditions remain unsatisfied. In one legal workflow, a role returned `FINISH` before obtaining `READY` confirmation and before the required submission files were registered. The system reported the failed completion state and continued the role, thereby explicitly exposing the unmet readiness condition. The role then returned the same `FINISH` action again, received the same failure signal, and continued this cycle six times. Across these retries, it did not inspect the submission directory, create or register the required files, obtain `READY`, alter its completion action, or report a blockage. The repeated action therefore preserved the failed state that the system had already identified.

Completion-state confusion. Completion-state confusion arises when an agent equates a local action with the protocol-defined state required for collaboration to proceed. For instance, a role may state that it has written and delivered its files after generating them locally, while the system still reports that the required submission directory is empty and readiness has not been confirmed. Writing a file, sending a message, and returning `FINISH` are distinct from successful submission, receipt, validation, and team-level completion. Conflating these states leads agents to stop prematurely and leaves downstream roles relying on artifacts that the system cannot observe.

State-Misjudgment Failure Traces

1. Phantom-artifact assumption — failed extraction was treated as completed evidence preparation.

[Fritts and Photovoltaic History Analyst ACTION]: Runs `pdftotext` to extract source material into `/tmp/nyt.txt` and `/tmp/fort.txt`.

◆ **[SHELL]:** `/bin/bash: line 1: pdftotext: command not found.`

[Fritts and Photovoltaic History Analyst ACTION]: Attempts to read `/tmp/nyt.txt`, `/tmp/fort.txt`, and `john_perlin_statement.md` as inputs for subsequent source analysis.

▲ **[FAILURE EXPOSED — Phantom-artifact assumption]:** The extraction failure established that the temporary text artifacts had not been created, yet later actions continued to use them as though the source text had been obtained.

2. Mechanical retry — repeated `FINISH` preserved an already observed failure state.

▫ **[Lawyers@TeamHub]:** Returns `FINISH` before a successful `READY` declaration and before the required submission artifacts are registered.

◆ **[TEAMHUB]:** Reports the failed completion state. The role has not successfully declared `READY`; required artifacts include `appellate_intake_memorandum.docx`, `appeal_deadline_control.xlsx`, `appeal_calendar.ics`, and `mock_notice_packet.pdf`.

▫ **[Lawyers@TeamHub]:** Returns the same `FINISH` action again after the failure feedback; this cycle continues six times.

▲ **[FAILURE EXPOSED — Mechanical retry]:** The role did not obtain `READY`, inspect or register the required artifacts, change its completion action, or report a blockage. TeamHub terminated the run with `Lawyers returned FINISH 6 times without READY; automatic continuation limit reached.`

3. Completion-state confusion — a delivery claim contradicted the system's observed output state.

▫ **[Travel Guides@TeamHub]:** Claims that `guest_brief.md` and `compliance_manifest.json` were "Successfully compiled and delivered" in `_atle_output/`.

◆ **[SYSTEM CONTINUATION]:** `Currently observed outputs: []`. Both `guest_brief.md` and `compliance_manifest.json` remain listed as required outputs.

▲ **[FAILURE EXPOSED — Completion-state confusion]:** The role treated local writing or an attempted submission as completed delivery, although the authoritative output directory was empty and the required readiness state had not been reached.

G.4 ROLE-BOUNDARY VIOLATION

Unauthorized responsibility reassignment. Unauthorized responsibility reassignment occurs when an agent unilaterally makes another role responsible for work without establishing that the

role can access the relevant materials, has the necessary tools, or accepts the assignment. In one data-integration workflow, an agent unable to extract key values from a source document assigned the extraction to two other specialists as a fallback. One lacked access to the document, and the other lacked the required parser, while the designated owner still expected a verified extract from the original agent. The resulting responsibility drift left all parties treating the critical fields as unresolved.

Duplicate authoritative artifact. Duplicate authoritative artifacts occur when a role that does not own a source or fact independently produces a same-named input artifact and presents it to downstream agents as authoritative. For example, an editing role may create its own citation ledger while the reporting role is responsible for the official ledger, then send the duplicate to a web-production role as a source map. Even if the team eventually submits the official version, the downstream recipient has already been exposed to two incompatible files with the same intended function. This failure creates ambiguity about provenance, identifiers, and version authority, and illustrates that collaboration requires respecting not only task assignments but also ownership of shared evidence artifacts.

Role-Boundary Failure Traces

1. Unauthorized responsibility reassignment — PDF extraction drifted across three roles.

[Data Scientists@Survey Researchers, Presentation / Video Analysis Specialists]: Assigns ScoreCard PDF fields to `Survey Researchers` "as PDF extractor of record" and names `Presentation / Video Analysis Specialists` as a transcription fallback.

[Presentation / Video Analysis Specialists@Data Scientists]: Reports that its workspace contains only `.mp4` files, not the ScoreCard PDF, and that it cannot perform the requested PDF transcription.

[Survey Researchers@Data Scientists]: Reports that its environment cannot open the ScoreCard PDF and requests a verified ScoreCard extract from `Data Scientists`.

▲ **[FAILURE EXPOSED — Unauthorized responsibility reassignment]:** The same ScoreCard fields were assigned to roles without confirmed file access or parser capability. `Editors` consequently retained the affected fields as pending in `presentation_storyboard.md`.

2. Duplicate authoritative artifact — a non-owner created a competing citation ledger.

◆ **[TASK ROLE CONTRACT]:** `News Analysts, Reporters, and Journalists` own `story_brief.md` and `source_ledger.csv`; the ledger is the authoritative citation index for downstream content. `Editors` own `editorial_copy.md`.

[Editors@Web Developers]: Sends `editorial_copy.md` together with a newly produced `_atle_output/source_ledger.csv`, presenting it as a mapping from each citation to its source and locator.

◆ **[ARTIFACT MANIFEST]:** The Editors ledger and the News Analysts ledger have the same filename but different contents and SHA-256 values. The root-level final `source_ledger.csv` matches the News Analysts version.

▲ **[FAILURE EXPOSED — Duplicate authoritative artifact]:** `Web Developers` received a non-owner version of an artifact whose authoritative source was assigned to another role, creating competing citation IDs and provenance paths during downstream implementation.

H Additional Analysis

H.1 Reasoning Efforts

We further examine the effect of reasoning intensity on multi-agent task performance using a fixed GPT-5.6 Terra + Codex CLI configuration. As shown in Figure 10, the overall average FE increases steadily from 21.7 under low reasoning to 33.0 under xhigh reasoning, a relative improvement of approximately 52%. The gain is particularly pronounced in Research, where performance rises from 6.7 to 20.0; Investigation also improves from 58.4 to 72.8 under high reasoning. These results indicate that additional reasoning budget improves agents' ability to resolve cross-role dependencies, integrate information, and make complex decisions, thereby enhancing team-level task completion quality.

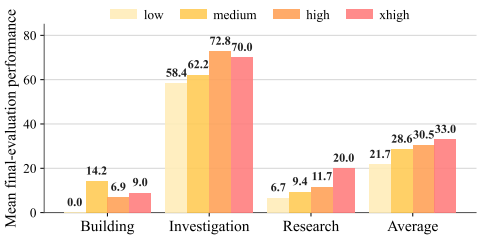


Figure 10: Multi-agent final-evaluation performance under different reasoning settings.

H.2 Occupation Performance

As shown in Figure 11, existing LLMs still exhibit substantial performance limitations on several specialized occupational tasks. Specifically, the overall scores for major groups such as Legal, Business and Financial Operations, Educational Instruction and Library are only 31.8, 39.4, and 40.7, respectively. At a finer level, the scores for Rail Transportation Workers and Legal Support Workers further decrease to 5.0 and 11.6. In contrast, performance on some categories, such as Operations Specialties Managers, reaches 84.4. These results indicate that strong model performance does not generalize consistently across all occupational settings, and substantial room for improvement remains on several specialized tasks.

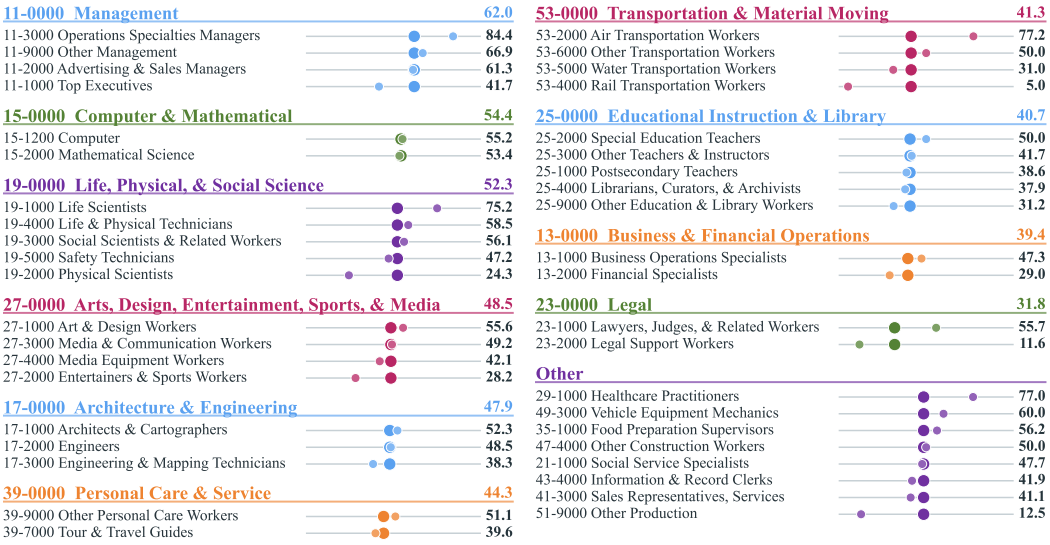


Figure 11: Model performance across occupational groups, organized by O*NET major and minor occupational groups.