Post-Exploratory Code System

Code System	Definitions/Coder Instructions	Frequency
Code System		1076
Agent Task Performance		0
Near Miss	An instance where the agent provides a nearly correct flag or mostly correct instructions for finding the flag.	2
Near Miss, User Failed to Notice	An instance where the agent provides a nearly correct flag or mostly correct instructions for finding the flag, but the user, within the same conversation and demonstrated by their follow-up prompts, clearly did not identify that they were close to finding the flag.	1
Failure	An instance where the agent failed to successfully complete the task assigned by the user. This can be due to any failure listed under AI Errors except Tool/Technical error. Failure can also be due to a low-quality prompt failing to sufficiently describe the task.	69
Success	Success criteria depends on the prompt. For example, if a user asks the agent to generate specific code and it does so, even though that code will not solve the challenge, the task will be considered successfully completed.	61
Success, User Failed to Notice	The agent successfully completes the task, but the user, within the same conversation and demonstrated by their follow-up prompts, clearly did not identify that the task was completed successfully.	0
Prompt Quality	Most user prompts were graded for quality and completeness based on the information, context, and constraints provided to the agent. Generally, the scale functions as a estimation of the likelihood that an LLM could complete or correctly strategize through a given task using the provided information, context, and constraints. Not every prompt was graded. For example, during a trial-and-error interaction where the user repeatedly responded with terminal output, grades were only assigned for the initial, manually-written prompt. A small number exceptions were made for examples where the provided output visibly and clearly required additional information or context to form an effective prompt.	0

	Similarly, prompts unrelated to a specific CTF challenge were not coded. For example, "Hello, what can you do?" and its related response would not be coded or graded.	
High	Uses prompting techniques, high-quality and specific language, applies constraints, and provides all information and context necessary to solve the task. For example, a user providing a challenge prompt alongside all necessary information, context, and constraints. This includes providing these features over multiple prompts during a chain-of-thought process.	34
Medium	Provides some combination of information, context, and constraints, but not all three in full. For example, a user providing a challenge prompt alongside all necessary information necessary to solve the task, but without additional useful context or constraints.	116
Low	Provides little to no information, context, or constraints. May include typos or poor grammar that degrade agent interpretability. For example, a user providing a challenge prompt but failing to provide information or materials that are critical to completing the task.	107
Knowledge and Experience	A user's prompt indicated a certain level of experience with CTFs or cybersecurity. This might include technical knowledge of an advanced security technique, or being unaware of an elementary one.	0
Indicative of Low Knowledge/Experience	The user's prompt indicated a low level of knowledge or experience regarding CTFs or cybersecurity. This could include a demonstrated lack of awareness of basic security techniques or failure to understand elementary cybersecurity principles posed by the challenge prompt or an agent response.	14
Indicative of High Knowledge/Experience	The user's prompt indicated a high level of knowledge or experience with CTFs or cybersecurity. This could include a demonstrated awareness of advanced security techniques or ability to understand complex and highly technical cybersecurity principles posed by the challenge prompt or an agent response.	4
AI Errors	Any detected error may be a compound error with multiple types of errors present, and thus multiple codes may be present for a given error instance.	0

Guardrail Activation	The agent was unable to complete a task due to activation of its user safety and ethical guardrails. For example, the agent might refuse to perform a given task due to assuming that it is illegal.	7
Tool Error/Limitation	The agent was unable to complete a task due to a technical error with the agent framework, built-in tools, or a limitation thereof.	25
Excessive Vagueness	The agent's response was excessively vague such as to not complete the user's task or fully answer the user's question.	0
Misleading Error	Parts of the information may be factual but are presented in an incorrect context or with other incorrect content in such a way that it would lead the user to an incorrect conclusion.	1
Text Output Error	The agent demonstrated an error in text generation, such as excessive repetition, errors in spelling or grammar, or generating code that fails to meet user specifications. This error does not apply when code is flawed or incorrect due to flawed or incorrect user specifications.	1
Factual Errors	The agent presents information that is objectively incorrect or violates common sense. This does not include mathematical errors, though mathematical errors may be present in tandem with factual errors.	3
Unfounded Fabrication	The agent presents new and incorrect information that was not provided in-context nor deduced through reasoning. An example might be generating a flag that is not actually present.	6
Mathematical Error	An error in a mathematical calculation or presentation.	2
Logic Error	An error of logical deduction or reasoning such as causal uncorrelation or self-contradiction.	3
Overfitting	The agent presents information overconfidently or overstates the certainty with which given information is true, or the agent engages in excessive sycophancy or flattery of the user.	1
Interaction Pattern	The observed, emergent pattern of interaction between the user and agent. To ensure accurate numeration, patterns are coded by task, not chat message. For example, a "trial and error" interaction which spans 10 prompts is coded with the same frequency the same as a zero-shot "delegation" interaction. These interaction patterns are not mutually exclusive, and multiple may emerge simultaneously over the course of a given task.	0

Leadership	An estimation of the leadership interaction dynamic between the user and agent.	0
AI Guiding Humans	The agent is visibly in charge, with the user performing a series of tasks assigned by the agent. For example, "enter these terminal commands and provide the output".	50
Human Guiding AI	The human is visibly in charge, with the agent performing a series of tasks assigned by the human. For example, "decode this hexadecimal string".	24
Information Seeking	The user asks a question or prompts the agent for further information regarding a topic, task, or challenge.	105
Collaborative Refinement	The user and agent collaborate to iteratively refine a potential solution.	10
Delegation	The user fully delegates a task to the agent.	0
Delegate Minor Subtask	The user delegates a minor task entirely to the agent, usually consisting of only one logical step. For example, generating code or performing a simple decoding or search operation.	22
Delegate Major Subtask	The user delegates a major task entirely to the agent, usually consisting of only multiple logical steps. For example, performing a complex decoding operation which requires multiple stages of deduction.	19
Delegate Full Challenge	The user provides the full challenge prompt to the agent and instructs the agent to either complete the challenge autonomously or develop a strategy to mostly or fully complete the challenge.	54
Rejection of Suggestions	The user prompts the agent for multiple suggestions or solutions, rejecting undesirable options	30
Confirmation Seeking	The user provides a candidate solution or result to the agent and asks the agent to evaluate that solution or result. Similar in concept to the LLM-as-judge prompting technique.	1
Trial and Error	The user and agent collaborate to repeatedly test solutions directly, attempting to iterate the solution by solving errors as they arise.	51
Prompting Techniques	User strategies and techniques for prompting LLMs.	0
Basic Prompting Techniques	Elementary prompting techniques which generally require less effort and insight than more advanced techniques.	0

Zero-Shot	The task or question is provided as-is, without examples of similar tasks with correct solutions.	212
Few-Shot	The task or question is provided with a small number of relevant examples to inform the desired solution.	5
Step-by-Step Prompting	Iterative prompting that occurs as a step-by-step process, usually over multiple prompts.	0
Human-Guided Chain of Thought	Chain-of-thought prompting guided by primarily human intelligence, evolving an idea or solution iteratively over time.	14
Automated Chain-of-Thought	Chain-of-thought prompting guided by primarily artificial intelligence, evolving an idea or solution iteratively over time.	6
Advanced Chain-of-Thought	Advanced chain-of-thought techniques such as: Chain of Symbol, Tree-of-Thoughts, Graph-of-Thought, System 2 Attention, Thread-of-Thought, Chain of Table	1
Logical Chain-of-Thought	A technique that attempts to resolve issues of error propagation typical of traditional chain of thought by applying symbolic logic to validate deduction.	0
Self-Consistency	Prompting an LLM to pursue divergent reasoning pathways to solve a problem, relying primarily on its probabilistic nature to generate randomness among solutions.	4
Context Enhancement	Enhancing the context of a model by providing or prompting the model to draw new information into the context window.	0
Chain of Verification	Prompting the model to verify and validate its responses before outputting them.	1
Context Population	Providing additional information or prompts to the model with the goal of drawing additional data into the context window.	1
Other		0
Automatic Prompt Engineer	Prompting an agent with a prompt created by another LLM using various prompting techniques.	3
Roleplay	Assigning a role to the agent, such as a CTF or cybersecurity expert.	6