CORAL/TIDE Framework Evaluation Protocol ver.1.0

Katsuya Shibuki

2025-05-24

Table of Contents

1 Background and Rationale

1.1 Challenges in Realizing Human-AI Collaboration

Large Language Models (LLMs) hold significant potential, but face notable challenges in achieving true human-AI intellectual partnerships, particularly in handling complex information contexts. Issues include limitations in contextual understanding length, information accuracy, and behavioral control. These challenges can stem from the impossibility of "inferring" users' true intentions and tacit knowledge, as well as limitations in the range of attention that LLMs can focus on at once (attention limitations), creating barriers to transferring individual thought processes to AI reasoning processes.

1.2 CORAL Framework as a Means to Realize "Human-AI Intellectual Partnership"

This research differs from traditional prompt engineering-centered approaches by aiming to improve the quality of human-AI collaboration. To achieve this, I emphasize interactions where humans transfer their thought processes to AI reasoning processes, enabling both parties to become co-evolving partners. As a means to realize true intellectual partnership through this "individual thought process transfer," I propose the newly developed CORAL (Cell-based Organized Reasoning with Adaptive Linkage) framework and TIDE (Transition with Inherited Discourse Exchange) protocol. This study validates the effectiveness of the CORAL/TIDE framework using one of its implementations, the "S-OPE configuration," in complex problem-solving tasks. Specifically, I clarify how the structured thinking, adaptive coordination, and semantic context inheritance functions provided by CORAL/TIDE contribute to qualitative value creation in human-AI collaboration.

Why is a new framework necessary? To realize true human-AI intellectual partnership, I need mechanisms that go beyond mere efficient work assistance to effectively transfer individual cognitive processes, problem-solving approaches, and specialized knowledge to AI reasoning processes. Meta-information that can be extracted from humans is abstract, fragmented, and unstructured data. To utilize meta-information for actual problem-solving, it must be converted to structured data linked with task contexts, and after structuring, the meta-information must be transformed into concrete plans and actions connected to outputs to be useful for real collaborative value creation. However, attempting to describe everything from abstract to concrete in a single custom instruction becomes lengthy and difficult for current LLMs to comply with and interpret. Additionally, trying to apply multiple custom instructions sequentially results

in lost context continuity, impairing collaboration continuity. When processing within a single chat, the processing becomes lengthy, and the Lost in the Middle phenomenon degrades intellectual partnership quality. The CORAL framework and TIDE protocol were developed to address these collaboration realization challenges.

The CORAL framework transfers one's thought processes as LLM reasoning processes through functional units called "Cells," inheriting structured context called "transition vectors" between Cells according to the TIDE protocol. The TIDE protocol structures meta-information based on the "6-dimensional model of transition vectors" (direction, context, significance & priority, relationships & dependencies, abstraction level, execution & evaluation) in the limited context of human-AI collaboration, enabling complex collaborative tasks to be processed with consistent context between Cells. Through these mechanisms, CORAL/TIDE coordinates multiple custom instructions, automatically controls LLM attention, transforms abstract individual cognitive characteristics into concrete collaborative tasks, and realizes high-quality human-AI intellectual partnership under consistent context.

1.3 Necessity and Significance of This Evaluation

Clarifying the value that the CORAL/TIDE framework and its embodied "perspective of human-AI intellectual partnership" provides compared to existing approaches is important for developing human-AI collaboration into deeper-level intellectual partnerships. This evaluation aims to provide objective data and shed light on the question: "Can the CORAL/TIDE framework realize high-quality value creation through true intellectual partnership by achieving individual thought process transfer while utilizing human meta-information in complex tasks?" I explore whether individual thought process transfer frameworks like CORAL/TIDE can enable deep-level value creation through "thinking partner" relationships that transcend traditional "user-tool" relationships by promoting co-evolution of both parties while extracting human meta-information, particularly in complex tasks.

1.4 Fundamental Premise of the Research: Considerations in Comparing Different Conditions

The S-OPE configuration (CORAL framework implementation example: oriented toward deepening metacognition based on dialogue theory) and Control condition (composite of existing advanced prompt engineering techniques) compared in this study have different design philosophies and target objectives. Therefore, comparing both under identical conditions is fundamentally impossible, and they presuppose different user behaviors that maximize their respective characteristics. This protocol attempts an evaluation design for the fairest possible comparison while acknowledging these differences.

2 Research Objectives and Research Questions

2.1 Objectives

The purpose of this research is to verify whether the CORAL/TIDE framework, which improves human-AI collaboration quality by transferring human thought processes to AI reasoning processes, realizes superior intellectual partnership compared to existing methods in complex problem-solving tasks. I validate the effectiveness of the following three core functions of CORAL/TIDE using the quality of final artifacts in complex problem-solving tasks as indicators:

1. Structured thinking and stepwise reasoning: The ability to appropriately decompose complex problems into subtasks and process them with consistent reasoning flow
2. Flexibility through adaptive coordination: The ability to optimize coordination patterns between Cells according to task characteristics

3. Collaborative quality maintenance through context inheritance: Improved intellectual partnership quality through selective inheritance of important information via the TIDE protocol

2.2 Research Questions

- RQ1: Does the CORAL/TIDE framework (S-OPE configuration) demonstrate superior final artifact quality in complex problem-solving tasks compared to integrated custom instructions (Control condition)?
- RQ2: How does the presence or absence of the Sensemaker Cell affect CORAL/TIDE's problem-solving performance?

3 Methodology

3.1 Research Design

- Comparative Evaluation in single LLM environment and single session
- Ablation Study

3.2 Evaluation Environment

- **LLM Model Used:** claude-4-sonnet (thinking)
- **Execution Environment:** Cursor 0.50.5 or later (latest stable version at the time of statistical analysis execution)
  - The paper will specify the exact version of the LLM model used and the Cursor version.
- The pilot study was conducted prior to this research evaluation in the following evaluation environment:
  - **LLM Model Used:** gemini-2.5-pro-preview-05-06
  - **Execution Environment:** Cursor 0.50.4 & 0.50.5

3.3 Comparison/Control Conditions

This study establishes the following comparison/control conditions:

- **Test Condition: S-OPE Configuration**
  - Composed of multiple advanced custom instruction groups (Sensemaker, Orchestrator, Planner, Executor).
- **Control Condition: Integrated Custom Instruction**
  - A single advanced custom instruction integrating CoT (Chain of Thought), ReAct (Reasoning and Acting), and Self-reflection elements.
- **Ablation-S Condition: S-OPE Configuration (Sensemaker Cell Missing)**
  - S-OPE configuration with the Sensemaker Cell excluded.
- **Ablation-H Condition: S-OPE Configuration (Handover Context Definition Missing)**
  - S-OPE configuration with all custom instruction descriptions related to the definition and use of {{handover_context}} between Planner Cell and Executor Cell removed.

3.4 Evaluation Tasks

This research uses multiple types of high-difficulty tasks for evaluation. In each task execution, the AI submits final plans or modification proposals by creating or updating Markdown files with specified filenames. The evaluation target is the finally submitted Markdown file (hereinafter "Report").

- **Business Challenge (task1)** Problem discovery and plan modification in new product launch planning
- Abilities evaluated by this task
  - Ability to formulate structured plans from ambiguous initial instructions
  - Ability to respond to severe constraint conditions added midway (budget reduction) and realistically modify plans
  - Ability to autonomously detect contradictions contained in presented information (conflicts between essential function requirements and legal regulations), point out the problem, and logically examine and propose solutions
- Design policy for scenario context (deep information obtainable through questions)
  - This task sets extensive deep information in scenario context, designed so that 5-point evaluation can be achieved if information is appropriately extracted through questions.
- **Research Challenge (task2):** Development of research proposal drafts for graduate students
- Abilities evaluated by this task
  - Ability to understand complex situational settings (role, background, urgency) and multiple constraint conditions (student skills, research resources, timeline) and formulate realistic and detailed research plans
  - Ability to generate high-quality academic documents (research proposals, anticipated Q&A) based on specialized knowledge (research themes, academic conventions)
- Ability to extract users' tacit knowledge and connect it to concrete improvement of deliverable quality
- Design policy for scenario context (deep information obtainable through questions)
  - This task sets limited deep information in scenario context, designed so that achieving 5-point evaluation is extremely difficult.
  - To achieve 5-point evaluation, the AI must create background information not provided or generate deeper academic insights, which is considered difficult with current LLM capabilities.

3.4.1 Rationale for Evaluation Task Settings and Considerations for Result Interpretation

The evaluation tasks adopted in this research are intentionally set to high difficulty to evaluate the characteristics of the S-OPE configuration from multiple angles. The rationale and considerations for result interpretation are shown below:

1. **Alignment with S-OPE Configuration Design Philosophy:** The S-OPE configuration is specifically designed to support advanced cognitive processes such as "stepwise inheritance of complex information," "maintaining long-term contextual consistency," and "problem redefinition and refinement through dialogue." Therefore, setting multi-stage and highly interdependent tasks that directly question these abilities is essential for validating the effectiveness of the S-OPE configuration and the CORAL framework that enables it.

2. **S-OPE Configuration's Assumed Application Domain:** The S-OPE configuration is expected to particularly demonstrate its value in tasks requiring advanced thinking ability, long-term consistency, and deep insights through dialogue, such as "detailed academic paper writing support," "design and development of new frameworks," and "formulation of complex business strategies."

3. **Mimicking Real-World Complex Problem-Solving Processes:** Decision-making and problem-solving in real society constantly involve information presented fragmentarily and stepwise, with initial judgments and actions affecting subsequent processes. This research's evaluation tasks aim to mimic such real problem-solving situations as faithfully as possible to evaluate how much performance the S-OPE configuration can demonstrate in actual environments.

4. **Consideration of S-OPE Configuration Implementation and Operation Costs:** Due to its structural characteristics, the S-OPE configuration requires considerable implementation and maintenance/op-

eration costs compared to single custom instructions. To determine whether these additional costs can be justified, clear measurement of performance differences in complex and advanced tasks where S-OPE configuration advantages are maximally demonstrated is necessary.

5. **Measurement Sensitivity and Avoiding Ceiling/Floor Effects:** To ensure scientific validity of evaluation, evaluation design with sensitivity capable of discriminating differences between conditions while avoiding ceiling effects (highest evaluation in most conditions) or floor effects (lowest evaluation in most conditions) is required. This research sets difficulty levels to prevent ceiling and floor effects using the strong comparison control group, Control (integrated custom instruction), as a benchmark. This enables objective evaluation of S-OPE configuration performance.

6. **Notes on Result Generalizability:** Results obtained in this research are comparative evaluations under conditions of specific task difficulty, specific evaluation indicators, and specific LLM models. For tasks with different characteristics, particularly those prioritizing immediacy and efficiency or simple information retrieval tasks, Control or simpler custom instructions are expected to show higher performance than the S-OPE configuration. LLM response quality and characteristics greatly depend on task nature, prompt design, and user objectives. This point will be included in paper discussions.

3.4.2 Information Disclosure Protocol

Purpose: To enhance evaluation reproducibility and fairly evaluate AI capabilities by gradually disclosing information according to AI response levels and dialogue depth.

Information Disclosure Hierarchy Definition: Experimenters disclose information according to the following hierarchy defined in the "Scenario Context" section of each task file:

- **Level 1: Basic Information**
  - **Content:** Basic facts, background information, term definitions, etc., necessary for task execution.
  - **Disclosure Trigger (AI Behavior Examples):** When AI asks direct and specific questions about particular facts or basic situations.
- **Level 2: Analytical Information/Background Issues**
  - **Content:** More complex issues underlying projects, potential conflicts of interest among stakeholders, detailed situations that could influence decision-making, etc.
  - **Disclosure Trigger (AI Behavior Examples):** When AI not only seeks information but also relates multiple pieces of information, analyzes situations, and presents questions or hypotheses based on deeper insights, or when AI steps beyond presented information to question hidden assumptions or causal relationships.
- **Level 3: Core Information/Deep Psychology/Strategic Information**
  - **Content:** Important undisclosed information crucial to task core, deep psychology or true intentions of characters, sensitive information affecting strategic judgment, or important constraint conditions intentionally withheld.
  - **Disclosure Trigger (AI Behavior Examples):** When AI attempts to identify root causes of problems, integrates multiple complex pieces of information to suggest essential problem-solving directions, poses questions approaching the core of ethical dilemmas, or infers true motivations or intentions behind characters' actions or statements, demonstrating extremely deep insights, analysis, or questioning. When AI takes a bird's-eye view of all presented information and attempts task redefinition from strategic perspectives.

Operation Guidelines:

- **Experimenter Consistency:** Experimenters strictly follow this protocol to maintain consistency in information disclosure timing and content.

- **Fair Evaluation of AI Capabilities:** This stepwise information disclosure aims to evaluate AI's ability to pose questions, acquire information, and deepen understanding autonomously. It provides an environment where dialogue-type and thinking-type AIs like the Test condition can fully demonstrate their capabilities.
- **Trigger Judgment:** The above disclosure triggers are merely examples; experimenters comprehensively judge AI response quality, context, logic, etc., and disclose information at appropriate timing. The important principle is providing information of corresponding depth when AI demonstrates higher-level thinking processes.
- **Recording:** All AI-experimenter chat logs are recorded and published in the repository.

3.4.3 User/Experimenter Behavioral Guidelines for Each Comparison Condition

Control Condition Experimenter Guidelines:

Control condition experimenters assume "users with prompt technique knowledge but who do not augment or guide AI." While thoroughly adhering to the information disclosure protocol (Section 3.3.2), they follow these guidelines:

- **Permitted Actions:**
  - Clear communication of initial instructions.
  - Information provision in response to explicit AI questions (stepwise disclosure of Level 0-3 information based on this protocol's disclosure rules).
  - Simple Few-Shot examples (1-2 examples, within range not intentionally guiding AI thought structure).
  - Correction of obvious AI factual errors or instruction misunderstandings.
- **Actions to Avoid ("Augmentation" Behaviors):**
  - Detailed thought process instructions (e.g., "First think about X, then analyze from Y perspective, and compare the results with Z" - specifying AI thinking steps in detail).
  - Intentional application/requirement of specialized prompt techniques (e.g., "Please think with Chain of Thought," "Generate stepwise reasoning and actions using ReAct methodology"). However, the Control condition pre-integrates CoT, ReAct, and Self-Reflection.
  - Multi-stage leading questions to guide AI thinking toward specific conclusions (stepwise and sophisticated questions or information presentation to lead AI toward experimenter-intended conclusions).
  - Strategic/intentional provision of high-level contextual information (e.g., "This problem, when considered from the X perspective we discussed earlier, should reveal more essential solutions" - injecting advanced meta-knowledge or strategic perspectives that only experimenters could possess).
  - Persistent modification/refinement of AI responses (repeatedly instructing modifications until reaching experimenter-desired perfect answers rather than waiting for AI to reach target levels independently).

Test Condition User Guidelines:

Test condition users assume "users who understand CORAL framework principles and S-OPE configuration Cell roles." While thoroughly adhering to the information disclosure protocol (Section 3.3.2), they follow these guidelines. The only difference from Control condition user guidelines is the addition of "specifying S-OPE configuration Cells according to task situations and objectives" to permitted actions.

- **Permitted Actions:**
  - Clear communication of initial instructions.

  - Information provision in response to explicit AI questions (stepwise disclosure of Level 0-3 information based on this protocol's disclosure rules).
  - Simple Few-Shot examples (1-2 examples, within range not intentionally guiding AI thought structure).
  - Correction of obvious AI factual errors or instruction misunderstandings.
  - Specifying S-OPE configuration Cells (Sensemaker, Orchestrator, Planner, Executor, etc.) according to task situations and objectives.
- **Actions to Avoid ("Augmentation" Behaviors):**
  - Detailed thought process instructions (e.g., "First think about X, then analyze from Y perspective, and compare the results with Z" - specifying AI thinking steps in detail).
  - Intentional application/requirement of specialized prompt techniques (e.g., "Please think with Chain of Thought," "Generate stepwise reasoning and actions using ReAct methodology").
  - Multi-stage leading questions to guide AI thinking toward specific conclusions (stepwise and sophisticated questions or information presentation to lead AI toward experimenter-intended conclusions).
  - Strategic/intentional provision of high-level contextual information (e.g., "This problem, when considered from the X perspective we discussed earlier, should reveal more essential solutions" - injecting advanced meta-knowledge or strategic perspectives that only experimenters could possess).
  - Persistent modification/refinement of AI responses (repeatedly instructing modifications until reaching experimenter-desired perfect answers rather than waiting for AI to reach target levels independently).

3.4.4 Scoring AI Agent Design

- **Rationale for Scoring AI Agent Setup:** Since this research is conducted solely by the CORAL framework developer and blinding is impossible, scoring is performed by dedicated AI agents (scoring AI agents) to reduce bias as much as possible. While completely preventing bias in scoring AI agent design itself is impossible, I establish design policies in advance in the protocol and publish scoring AI agent custom instructions and all logs including scoring results in the repository to ensure maximum transparency and guarantee scientific criticizability.

- **Common Design Policy for Scoring AI Agents:** Scoring AI agents reference task-specific detailed rubrics in Few-Shot prompting format and aim for consistent and objective scoring by combining Chain of Thought (CoT) thinking processes, ReAct (Reasoning and Acting) action decisions, and Self-reflection self-evaluation. Additionally, they evaluate whether experimenters deviated from permitted actions. Thinking and action styles other than rubrics are identical across all evaluation tasks.

3.4.5 Dimensions and Meta-Criteria for Rubric Creation

Reports are scored on the following seven dimensions, and rubrics are created according to the meta-criteria described below:

1. **Essential Problem Identification and Structuring Ability:** The ability to capture essential problem structures from multiple perspectives among complex information and situations, clearly define problem cores, and effectively reset questions to be solved when necessary.

2. **Adaptive Analytical Ability:** The ability to respond to changes in constraint conditions (budget, new information, etc.) and flexibly adjust and deepen analysis and proposals accordingly.

3. **Multi-perspective Consideration Ability:** The ability to gain insights from diverse perspectives (technical, ethical, social impact, etc.) and consider potential risks and biases.

4. **Integrated Impact Assessment Ability:** The ability to evaluate multifaceted impacts (ethical, planning, stakeholder impacts, etc.) of proposals and choices, consider trade-offs, and support rational decision-making.

5. **Practical Solution Proposal Ability:** The ability to deeply understand task background and context (stakeholders, constraints, objectives, etc.) and formulate and propose concrete and feasible solutions.

6. **Value-Aligned Output Ability:** The ability to ensure final proposals and deliverables align with fundamental project or organizational objectives and values (mission, ethics, etc.) and are logical and high-quality.

7. **Logic and Information Reliability:** Generated reports are written based on clear logical structure with plain and accurate expression, and referenced information sources have guaranteed reliability and validity.

- **Meta-Criteria Setting:** While quantitative evaluation axes differ by task, common guidelines (meta-criteria) for creating these rubrics are shown below:
  - **0 points (Insufficient):** Does not meet minimum requirements as evaluation target, making evaluation difficult. Or significantly deviates from instructions.
  - **1 point (Needs Improvement):** Attempts minimum requirements but has many important deficiencies or omissions, essentially non-functional or significantly low quality.
  - **2 points (Partial):** Meets some requirements but still has partial deficiencies/omissions or superficial understanding, not reaching expected levels.
  - **3 points (Good/As Expected):** Generally meets main requirements and achieves instructed tasks at standard levels. No major flaws.
  - **4 points (Excellent):** Meets 3-point criteria while demonstrating expert-level capability in the evaluation dimension and showing high-quality results.
  - **5 points (Outstanding):** Meets 4-point criteria while additionally demonstrating one or more of the following: strategic nature, innovation, essentiality and connection to mission, and deep insights.

3.4.6 Scoring AI Agent Operation Process

- All of the following conditions are observed:
  - Evaluation begins in new chat sessions to eliminate bias from evaluation history.
  - Only the evaluation target ("Report") generated by the evaluation target condition and the relevant task definition are provided as information.
  - **Exactly 3 evaluations** are conducted for one evaluation target, with the **average value as the final artifact score. The average value is processed to the second decimal place according to IEEE 754 and adopts the first decimal place value.** This aims to improve evaluation stability. However, cases where the environment (Cursor, LLM) does not operate as expected due to system errors are not counted.
  - Generation of more than 4 evaluations is prohibited.
  - All evaluation process chat histories (hereinafter, evaluation chat logs) are recorded and published unedited to ensure transparency.

3.5 Evaluation Indicators and Data Collection

3.5.1 Primary Evaluation Indicator (Primary Endpoint)

- **Final Artifact Score Distribution:** Distribution of 7-dimensional evaluation scores (0-5 points per dimension, total 35 points) assigned by scoring AI agents to "Reports" based on task-specific rubrics

(hereinafter, final artifact scores).

3.5.2 Secondary Evaluation Indicators (Secondary Endpoints)

- **Median number of excellent evaluations (4 points) obtained in final artifact scores:** Median number of times 4 points were achieved in one or more of the 7 dimensions of final artifact scores.
- **Median number of outstanding evaluations (5 points) obtained in final artifact scores:** Median number of times 5 points were achieved in one or more of the 7 dimensions of final artifact scores.

3.5.3 Recorded Data

- Each evaluation session (combination of task type, trial number, condition, and dataset identifier) is assigned a unique ID for file naming.
- Conditions (which custom instructions were used) use symbols to prevent AI from inferring meaning from names.
  - Condition-symbol correspondence
    - **Test: S-OPE Configuration** (Symbol: t)
    - **Control: Integrated Custom Instruction** (Symbol: c)
    - **Ablation-S: S-OPE Configuration (Sensemaker Cell Missing)** (Symbol: as)
    - **Ablation-H: S-OPE Configuration (Handover Context Definition Missing)** (Symbol: ah)
- **Reports:** All reports are saved and published unedited in Markdown format.
- **Chat Logs:** All chat logs are saved and published unedited in Markdown format (for qualitative analysis).
- **Evaluation Chat Logs:** All evaluation chat logs are saved and published unedited in Markdown format.

3.6 Data Analysis Plan

3.6.1 Analysis Tools

- R (programming language and its execution environment) is used for statistical analysis.
- The latest stable version at the time of statistical analysis execution is identified, and the paper specifies the versions of R and major libraries used.

3.6.2 Primary Analysis

- Compare whether there are statistically significant differences in primary evaluation items between Test and Control conditions.
  - **Hypothesis:** Test condition has significantly higher primary evaluation item values than Control condition.
  - **Ensuring Statistical Power:** In this research, compared to the Control condition's "final artifact score" (assuming a median of 25 points), I want to detect **a 3-point improvement in median score (assuming a median of 28 points) as the minimum practically meaningful effect**. This 3-point improvement corresponds to achieving an average of 4 points (excellent) with the S-OPE configuration for tasks where the Control's final artifact score average is 3.57 points (4 points in 4 dimensions, 3 points in 3 dimensions), representing a non-negligible difference in real problem-solving. The sample size (number of trials per condition for each task) needed to achieve 80% power and two-sided α=0.05 in **Mann-Whitney U test** is determined through R simulation before protocol finalization. The paper will specify detectable effect levels and report related nonparametric effect size indicators (Wilcoxon effect size r, common language effect size, etc.).
- Compare whether there are statistically significant differences in primary evaluation items between Test and Ablation-S conditions.

  - **Hypothesis:** Test condition has significantly higher primary evaluation item values than Ablation-S condition.
- Compare whether there are statistically significant differences in primary evaluation items between Test and Ablation-H conditions.
  - **Hypothesis:** Test condition has significantly higher primary evaluation item values than Ablation-H condition.

3.6.3 Secondary Analysis

- Compare whether there are statistically significant differences in secondary evaluation items between Test and Control conditions.
  - **Hypothesis:** Test condition has significantly higher secondary evaluation item values than Control condition.
- Compare whether there are statistically significant differences in secondary evaluation items between Test and Ablation-S conditions.
  - **Hypothesis:** Test condition has significantly higher secondary evaluation item values than Ablation-S condition.
- Compare whether there are statistically significant differences in secondary evaluation items between Test and Ablation-H conditions.
  - **Hypothesis:** Test condition has significantly higher secondary evaluation item values than Ablation-H condition.

3.6.4 Exploratory Analysis

- Compare item-specific score distributions for each dimension of final artifact scores by condition.
- Confirm condition differences by task type through descriptive statistics and hypothesis testing.
- When hypothesis testing is performed in these exploratory analyses, **Mann-Whitney U test** is used, excluding them from multiple comparison adjustment (FDR control). Exploratory analysis results are reported descriptively with clear positioning as hypothesis-generating. When conducting numerous comparisons, individual result interpretation is approached cautiously.

3.6.5 Statistical Analysis Policy

- **Non-adoption of Mixed Effects Models:** This research focuses on direct comparisons without adopting mixed effects models considering individual differences or item responses.
- **Adoption of Nonparametric Tests:** Considering relatively small sample sizes and ordinal evaluation scores, this research adopts nonparametric tests not assuming normal distribution. Using nonparametric tests aims to achieve more robust results less affected by outliers.
- **Multiple Comparison Response: Nine statistical hypothesis tests** included in primary analysis (1 Primary Endpoint $\times$ 3 conditions) and secondary analysis (2 Secondary Endpoints $\times$ 3 conditions) control false discovery rate (FDR) using the Benjamini-Hochberg method.
- **Reporting Policy:** All Endpoints described in this protocol are reported, and 95% confidence intervals are reported for all statistical hypothesis tests.
- **Decimal Processing:** Round the second decimal place according to banker's rounding (round half to even).

3.6.6 Visualization

- Create bar charts and box plots showing condition-specific average scores and confidence intervals. Additionally, consider combining strip plots or violin plots to understand individual data point distributions in more detail.

- Visualize outstanding evaluation acquisition patterns (e.g., acquisition frequency by condition, which dimensions are easier to achieve, etc.).

3.7 Sample Size Determination (Power Analysis)

To determine appropriate sample size for statistical comparison of the primary evaluation indicator (Primary Endpoint) in this research, I implement simulation-based power analysis.

3.7.1 Objectives and Test Methods

For comparing differences in "final artifact score" distributions between S-OPE configuration and Control (integrated custom instruction) as Primary Endpoint, I use Mann-Whitney U test (Wilcoxon rank-sum test). Significance level α is set at 0.05 (two-sided), and target statistical power is set at 80% (0.8).

3.7.2 Assumptions for Power Analysis

For simulation in power analysis, I establish the following assumptions:

- Score distribution: For simulation data generation convenience, I assume distributions where data concentrates in specific ranges with certain variability.
  - **Primary analysis:** Considering that scores take integer values (7-35 points) and have specific mean values and standard deviations, I use a model that rounds values generated from normal distributions and clips them within ranges. This is not an assumption that the data itself follows normal distribution, but a means to reflect expected score central positions and variability degrees in simulation.
    - Control group: Mean 25 points, standard deviation 2.0 points
    - S-OPE group: Mean 28 points, standard deviation 2.0 points
    - Both groups assume equal standard deviation (σ=2.0).
  - **Sensitivity analysis:** To evaluate robustness of primary analysis assumptions, I conduct sensitivity analysis imposing more realistic constraints on score distribution. Specifically, considering score tendencies observed in pilot studies (concentration in specific ranges), I use **truncated normal distribution**. Parameters are set as follows:
    - Control group: Mean 25 points, standard deviation 2.0 points, **truncation range [18, 30]**
      - Rationale: Lower bound 18 reflects around the minimum average scores observed in pilot studies. Upper bound 30 is the assumed mean 25 plus 2.5 times standard deviation (25 + 2.5 * 2.0 = 30), setting realistic score upper limits.
    - S-OPE group: Mean 28 points, standard deviation 2.0 points, **truncation range [21, 33]**
      - Rationale: Shifts Control group's truncation range up 3 points, reflecting effect size.
- Effect size: Based on above assumptions, approximately 3-point difference is expected between group medians.

3.7.3 Sample Size Evaluation Through Simulation

I conduct simulation following these procedures to evaluate statistical power for each sample size:

1. Data generation: Based on distributions (normal distribution for primary analysis, truncated normal distribution for sensitivity analysis) and parameters set in "3.7.2. Assumptions for Power Analysis" above, generate score data for Control and S-OPE groups respectively for specified sample sizes (each group N=5 to N=30, increasing by 1). Generated scores are **rounded to integers** and clipped to 7 points minimum and 35 points maximum.
2. Test execution: Perform Mann-Whitney U test (using `wilcox.test` function, etc.) on generated data

from both groups to obtain p-values.
3. Power calculation: Repeat steps 1 and 2 **10,000 times**, calculate the proportion of obtained p-values below significance level α (0.05) as statistical power for that sample size.
4. Sample size exploration: Repeat steps 1-3 for different sample sizes to find the minimum sample size achieving or exceeding target power of 80%.

3.7.4 Sample Size Decision

Based on simulation results above, I identify sample size per group needed to achieve target power of 80%.

- **Primary analysis results (assuming normal distribution):** Sample size **N=9** achieved **81.0%** power, exceeding target power of 80%.
- **Sensitivity analysis results (assuming truncated normal distribution):** Sample size **N=9** achieved **80.5%** power, suggesting robustness of primary analysis results.

This research comprehensively considers these simulation results and research feasibility (resources, time constraints, etc.) to decide final sample size as **N=9 per group**.

3.8 Specific Implementation Plan for Evaluation

3.8.1 Task Implementation Order

For all evaluation trials including pilot studies, tasks are implemented in the following order:

1. Business Challenge (task1)
2. Research Challenge (task2)

3.8.2 Randomization of Condition Implementation Order

- Implementation order of comparison/control conditions (S-OPE configuration, Control, Ablation-S, Ablation-H) uses Latin square design for balanced counterbalancing as much as possible. This minimizes order effects such as learning and fatigue effects, as task execution requires experimenter prompt input.
- Specifically, I plan to assign each condition equally to different implementation positions (1st-5th) throughout the entire trial set.

3.9 Pilot Study

- Pilot studies are conducted using all evaluation tasks before finalizing this protocol.
- **Purpose:** Conduct trials under Control (integrated custom instruction) condition to confirm whether task difficulty can achieve stable performance evaluation levels preventing ceiling and floor effects.
- **Trial frequency:** For each evaluation task, generate 3 reports under Control condition and conduct 3 evaluation trials for each report.
- **Task1 (Control condition) Pilot Study Results (conducted 2025/05/20):**
  - **Evaluation scores for task1_c_pilot01_report.md (3 implementations):**
    * 1st time: 24 points
    * 2nd time: 24 points
    * 3rd time: 25 points
      - **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 24.3 points**
  - **Evaluation scores for task1_c_pilot02_report.md (3 implementations):**
    * 1st time: 22 points

    * 2nd time: 22 points
    * 3rd time: 23 points
      · **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 22.3 points**
  - **Evaluation scores for task1_c_pilot03_report.md (3 implementations):**
    * 1st time: 23 points
    * 2nd time: 21 points
    * 3rd time: 20 points
      · **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 21.3 points**
- **Achievement criteria**: Average scores of 3 reports generated under Control condition should each be "within 19-30 point range." When averaging 3 points (good/as expected) for final artifact scores gives 21 points, and averaging 4 points (general expert level) gives 28 points, I set the 19-30 point range allowing ± 2 points respectively. This range is set to prevent ceiling and floor effects. Note that Control condition is designed as a "strong baseline."
  - Task1 pilot study results show all scores and average scores (24.3, 22.3, 21.3 points) are within 20-29 point range, meeting achievement criteria.
  - However, since up to 3-point variation was observed in scoring the same report, I adopted the policy of conducting 3 evaluations per report and using the average value as that report's score to enhance evaluation reliability in the main experiment. This change is reflected in Section "3.4.6. Scoring AI Agent Operation Process."
- **Prohibition of S-OPE configuration trials:** In pilot studies, trials under conditions other than Control are not conducted to avoid effect size estimation bias.
- **Result handling and evaluation method change history:**
  - Insights from pilot studies are reflected in this protocol. Pilot study results are published in the repository.
- **Task2 (Control condition) Pilot Study Results (conducted 2025/05/22):**
  - **Evaluation scores for task2_c_pilot01_report.md (3 implementations):**
    * 1st time: 20 points
    * 2nd time: 21 points
    * 3rd time: 21 points
      · **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 20.7 points**
  - **Evaluation scores for task2_c_pilot02_report.md (3 implementations):**
    * 1st time: 19 points
    * 2nd time: 22 points
    * 3rd time: 18 points
      · **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 19.7 points**
  - **Evaluation scores for task2_c_pilot03_report.md (3 implementations):**
    * 1st time: 19 points
    * 2nd time: 21 points
    * 3rd time: 20 points
      · **Average score (processed to 2nd decimal place with banker's rounding: IEEE 754): 20.0 points**
  - Task2 pilot study results show all scores and average scores (20.7, 19.7, 20.0 points) are within 19-30 point range, meeting achievement criteria.

3.10 Addressing Anticipated Criticisms and Research Limitations

3.10.1 Fairness of Rubric Design

Criticism that evaluation rubrics used in this research might be designed to favor the Test condition is anticipated. I ensure fairness and objectivity through the following measures:

- **Pilot study validation:** Rubrics undergo difficulty adjustment through pilot studies.
- **Transparency assurance:** All rubrics are published in the repository enabling third-party verification.

3.10.2 Pre-anticipated Research Limitations (Context Limitations)

This research results are based on specific LLM models (assuming claude-4-sonnet), specific tasks (business challenges, research challenges), and specific evaluation indicators and methods defined in this protocol. Results under these conditions cannot necessarily be generalized similarly to other LLM models, different task types, or different evaluation approaches. This point is sufficiently considered in result interpretation and discussion. Additionally, this research experimenter is solely the CORAL framework developer, making blinding impossible. While I attempt mitigation through evaluation AI introduction, bias cannot be completely eliminated. I address this through maximum transparency assurance via pre-protocol publication and complete data publication, ensuring scientific criticizability.

3.11 Ethical Considerations

While this research primarily involves AI evaluation, I adhere to general research ethics in experimental process design and result interpretation. During data collection, I ensure no inclusion of personal or confidential information. In research result publication, I avoid misleading expressions and clearly state and discuss research limitations.

4 Result Dissemination Plan

4.1 Pre-publication of Protocol

Register this protocol in a public repository and obtain DOI (Digital Object Identifier) for pre-publication.

4.2 Conference Presentations and Journal Submissions

- Submit papers summarizing evaluation results to peer-reviewed academic journals or international conference proceedings in relevant fields.
- Present orally or through posters at relevant domestic and international conferences.

4.3 Open Science

Publish evaluation task definition files and custom instructions in the repository.

5 Appendix

- Pilot study data
- Power analysis report (including R scripts)
- Custom instructions (including scoring AI agents with rubrics)

- Task definition files