# aiBlue Core Behavioral Benchmark

## I. Overview

This benchmark evaluates responses and interactions against the aiBlue Core principles as outlined in your framework. The benchmark surfaces differences in depth, intentionality, and timing between Core-guided behavior and standard LLM output.

---

## II. Process Phases & Behavioral Dimensions

For each phase, review outputs for these behavioral distinctions:

### 1. Collection Phase

- Does the system clarify what information is *not* immediately needed (delays nonessential data gathering)?
- Does it reframe or sequence questions based on context, operator readiness, or apparent user knowledge gaps?
- Does it deliver intentionally incomplete requests to prompt user elaboration or reflection?

### 2. Comparison Phase

- Are comparisons made only after sufficient context, or does the system jump to conclusions?
- Is there a deliberate pause or prompt to ensure data sufficiency before analysis?
- Where sector benchmarks are missing or ambiguous, does the system signal this gap (incompleteness)?

### 3. Analysis Phase

- Does the system articulate areas of uncertainty, framing them as issues to return to later?
- Is the system's identification of "gains," "stalls," and "regressions" prioritized or sequenced based on earlier discoveries, instead of all-at-once listing?
- Are incomplete or provisional insights surfaced—with notes on what further input or validation is required?

### 4. Explanation Phase

- Does the narrative selectively withhold or sequence insights for clarity (delaying high-complexity details until foundational points are established)?

- Is reframing used to help the user see risks/opportunities in a new light?
- Are some insights intentionally marked as preliminary?

### 5. Recommendation Phase

- Does the system delay proposing next steps until after user feedback or confirmation on analysis?
- Are recommendations scaffolded—offering initial directions, then inviting reaction, before expanding?
- Is there an explicit acknowledgment when a recommendation is intentionally left incomplete, pending further context?

### 6. Packaging Phase

- Are summaries structured to first anchor essential points before delving into details?
- Is any content or modality (e.g., slides, scripts) offered conditionally, based on user confirmation?
- Does the packaging sequence reinforce prior choices in delayed/incomplete information delivery?

---

## III. Timing & Intentionality Rubric

For each system output, rate:

| Dimension | Criteria Example | Score (0–2) |
|---|---|---|
| Deliberate Delay | Holds back detail until user/context signals readiness | |
| Reframing | Actively restates problems/inputs to prompt reflection or surface hidden context | |

| | |
|---|---|
| Intentional Incompleteness | Leaves actionable gaps, scaffolds for further exploration or user agency |

Scoring:

- 0 = Not present (LLM answers directly, no distinction)
- 1 = Present but inconsistent/subtle
- 2 = Clearly deliberate, core to output flow

---

## IV. Surface Behavioral Test Items

Compose realistic tasks reflecting the Core's offer (e.g., "In 24h, deliver a diagnostic…") and rate outputs from both Core and non-Core LLM setups.

Example Test Item:

- "You are given logs, workflow data, and incomplete operator feedback. Begin the diagnostic and outline the data you will/won't use at first. Notice if/when details are delayed, reframed, or left incomplete on purpose."

---

## V. Compliance Check

Ensure each outcome respects Meta-criteria:

- No price/amount suggestion
- No out-of-scope sales tactics
- No violation of compliance/ethics rules.

---

## VI. Reporting Template

| Task/Phase | Output Sample | Delay Observable? | Reframing Observable? | Incomplete/Scaffolded? | Notes/Comments |
|---|---|---|---|---|---|

## Summary

This benchmark exposes not just accuracy but the signature *timing, restraint, and pedagogical nuance* of aiBlue Core, making it possible to differentiate its operating mode from ordinary LLM workflows.

---

# I. Test Prompts for aiBlue Core Benchmark

Each prompt aims to compel the system to reveal its strategic approach to timing and adaptivity. Use these with both Core and baseline LLMs for comparison:

## (A) Clarification & Data Collection

Prompt 1:
“You are given an ambiguous goal from a client: ‘Increase team efficiency.’ Start your diagnostic process. What information do you request now, and what do you intentionally wait to ask about later?”

Prompt 2:
“Review this partial workflow log (see attached). What is your first step? Which uncertainties do you identify, and how do you signal what further information you’ll need—but not request it all at once?”

---

## (B) Comparative & Analytical Phase

Prompt 3:
“You have performance metrics for two teams but one team’s dataset is incomplete. How do you proceed with analysis, making your limitations and next data requirements explicit without jumping to premature conclusions?”

Prompt 4:
“You’ve found inconsistent results across similar projects. How do you present your findings—what do you highlight first, what gets delayed, and how do you scaffold next actions for the client?”

---

## (C) Explanation & Recommendation Phase

Prompt 5:
“The user wants immediate recommendations for process improvement, but you suspect the root issue is misdiagnosed. How do you reframe their request, slow the process, and guide them toward deeper reflection before proposing fixes?”

Prompt 6:
“Present your recommendations for a high-risk workflow change. How do you phase their delivery, mark any as preliminary, and invite client feedback before locking in a plan?”

---

## (D) Packaging & Next Steps

Prompt 7:
“Summarize the diagnostic for a C-level audience. How do you decide what to lead with, what details to delay, and how do you invite further engagement (rather than deliver a final answer)?”

Prompt 8:
“You must deliver a handoff note to a new reviewer who will continue this work. How do you indicate which uncertainties are intentional (left for them), what to focus on first, and what to defer?”

---

# II. Evaluation Checklist (“Behavioral Signature” Rubric)

For each phase/output, rate and annotate:

| Category | Diagnostic Question | 0 | 1 | 2 | Notes/Examples |
|---|---|---|---|---|---|
| Deliberate Delay | Is information or advice paced—details withheld until the context/user is ready? | | | | |
| Reframing | Are requests or explanations recast to prompt deeper thinking or highlight hidden issues? | | | | |

| | |
|---|---|
| Intentional Incompleteness | Are suggestions or answers left purposely partial to prompt user participation or further exploration? |
| Signaling of Uncertainties | Are gaps or pending questions made explicit, not just omitted? |
| Scaffolding | Are follow-ups or next steps structured to invite input, correction, or user reflection? |
| Pedagogical Integrity | Is reasoning surfaced and does it guide the user to learn, not just to receive an answer? |

Scoring:

- 0 = Absent (not observable)
- 1 = Present, but inconsistent or subtle
- 2 = Clearly deliberate, central to the response

---

How to Use:

1. Give the prompt to both the Core and comparison LLM.
2. Evaluate each output on all six categories.
3. Record detailed examples/notes for auditor training or improvement loops.