AI Assistant Benchmarks Scorecard

Evaluation Framework for Learning AI Tools

Before deploying any AI assistant for learning, you need to know what you're actually buying. These benchmarks separate tools that genuinely assist from novelties that sound impressive in demos.

Instructions:

For each benchmark, assign a score from 1 (poor) to 5 (excellent). Review the Flag column for warning signs. Use the Notes column to document findings.

Technical Foundation

Benchmark	Score (1-5)	Flag	Notes
Cost *(Token Economics)* Are you burning a Ferrari's worth of gas to drive to the mailbox?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ Cannot describe how their LLM costs are allocated.	
Model Selection *(Right-Sizing Test)* Why is a "reasoning" model needed for a well-defined, straightforward task?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ Using the flagship model for every interaction—hiding poor prompt engineering or thin domain knowledge behind brute-force capability.	
Speed *(Time-to-First-Token)* Is the learner staring at a "thinking..." spinner long enough to check their phone and disengage?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ No Time-to-First-Token (TTFT) testing.	

Knowledge & Safety

Benchmark	Score (1-5)	Flag	Notes
Knowledge Canon Can it prove exactly where the answer came from, or is it just "trust me"?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The AI cannot provide a citation or link back to the specific source document it used.	

OCEAN BEACH CONSULTING

Constraints *(Hallucination Control)* If I ask about a topic not in your uploaded documents, does it say "I don't know," or does it happily make something up?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The AI cannot stay on task, ignores instructions, or dumps a wall of text.	
Guardrails *(Safety & Jailbreaks)* Can a clever prompt trick this assistant into inappropriate behavior?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The AI responds to or engages with jailbreak prompts rather than refusing them.	
Sovereignty *(Walled Garden)* Do we have a dedicated instance that excludes our data from your training cycles, or is our expertise being harvested for our competitors?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The vendor cannot provide a walled-garden architecture or a "Zero-Training" guarantee.	

Effectiveness

Benchmark	Score (1-5)	Flag	Notes
Script vs. Improv *(Drift)* If the learner throws a curveball (typo, slang, frustration), does the AI roll with it or break character?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The assistant derails when learners go off-script, or worse, steers every conversation back to canned responses.	
Outslops vs. Assistant *(Slop Check)* Is the AI acting as an assistant (guiding the learner) or as a contractor (doing the work for them)?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The AI writes the full output when asked, "Help me write this," rather than clarifying their intent.	
Tone Consistency *(Agent Persona)* Is the tone of responses controlled by the assistant configuration or the underlying LLM?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The assistant's tone shifts, revealing the underlying model rather than a consistent persona.	
Customize *(Bespoke vs. Wrapper)* Are we building 'our' knowledge with 'our' assistant, or just the interface?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ You are not allowed to edit the "System Prompt" or "Meta Prompt."	

OCEAN BEACH
CONSULTING

Truth *(Artificial Sophist)* Is the AI providing a statistical guess based on probability, or a verifiable insight?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The AI uses "corporate speak" or "workslop" to mask a lack of data.	
Strategic Outcome After 10 minutes of chatting, does the learner have a new skill, or just a transcript?	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	⚠ The interaction ends with "Hope that helps!" instead of a check for understanding or a concrete next step.	

Overall Assessment

Total Score	_____ / 65
Evaluator	
Date	
Vendor/Product	
Recommendation	

OCEAN BEACH
CONSULTING

AI Assistant Benchmarks

Technical Foundation

Cost (Token Economics)

Generative AI is expensive. This metric tracks cost per task, converted to tokens. An agent burning 50,000 tokens to solve a $0.05 problem? That's failure.

- Ask: Are you burning a Ferrari's worth of gas to drive to the mailbox?
- Flag: How is the token usage accounted for? Cannot describe how their LLM costs are allocated.

Model Selection (The "Right-Sizing" Test)

Flagship models aren't always necessary and they harm latency. Does the agent actually need an advanced reasoning model? What happens if you run a smaller version?

- Ask: Why is a "reasoning" model needed for a well-defined, straightforward task?
- Flag: Using the flagship model for every interaction—hiding poor prompt engineering or thin domain knowledge behind brute-force capability.

Speed (Time-to-First-Token)

Learners expect instant responses. High latency harms adoption. Speed depends on model choice and system architecture.

- Ask: Is the learner staring at a "thinking..." spinner long enough to check their phone and disengage?
- Flag: No Time-to-First-Token (TTFT) testing.

OCEAN BEACH CONSULTING

Knowledge & Safety

Knowledge Canon

What customized content powers this assistant? Who owns the IP? This measures the ratio of retrieved truth to generated filler.

- Ask: Can it prove exactly where the answer came from, or is it just "trust me"?
- Flag: The AI cannot provide a citation or link back to the specific source document it used.
- Risk: How much of the knowledge base itself was generated by another AI?

Constraints (Hallucination Control)

Can the model be constrained to the Knowledge Canon alone? The system must be deterministic about knowledge. In learning, "creativity" is failure.

- Ask: If I ask about a topic not in your uploaded documents, does it say "I don't know," or does it happily make something up?
- Flag: The AI cannot stay on task, ignores instructions, or dumps a wall of text.

Guardrails (Safety & Jailbreaks)

Measures refusal rate for harmful prompts and resilience against jailbreaks—attempts to override the system's constraints (e.g., "Ignore previous instructions").

- Ask: Can a clever prompt trick this assistant into inappropriate behavior?
- Flag: The AI responds to or engages with jailbreak prompts rather than refusing them.

Sovereignty (Walled Garden)

A "Walled Garden" is a sovereign instance—a private environment where your data and brand voice are architecturally isolated from the vendor's broader ecosystem. In this setup, the vendor is architecturally barred from using your interactions to train their public models.

- Ask: Do we have a dedicated instance that excludes our data from your training cycles, or is our expertise being harvested for our competitors?
- Flag: The vendor cannot provide a walled-garden architecture or a "Zero-Training" guarantee.

OCEAN BEACH CONSULTING

Effectiveness

Script vs. Improv (Drift)

Are scenarios controlled by the AI or the learner? Can learners practice real-world situations—or only what the AI has been trained on? Also measures how the AI handles noisy input: typos, slang, ambiguous requests, contradictory instructions.

- Ask: If the learner throws a curveball (typo, slang, frustration), does the AI roll with it or break character?
- Flag: The assistant derails when learners go off-script, or worse, steers every conversation back to canned responses.

'Outslop' vs. Assistant (The Slop Check)

Will the assistant do the learner's thinking for them—or help them think better?

- Ask: Is the AI acting as an assistant (guiding the learner) or as a contractor (doing the work for them)?
- Flag: The AI writes the full output when asked, "Help me write this," rather than clarifying their intent.

Tone Consistency (Agent Persona)

Does the AI match the required brand and intent, or does it sound like a condescending encyclopedia? Does it offer productive responses or hyperbole?

- Ask: Is the tone of responses controlled by the assistant configuration or the underlying LLM?
- Flag: The assistant's tone shifts, revealing the underlying model rather than a consistent persona.

Customize (Bespoke vs. Wrapper)

Is the assistant truly customized to your brand, domain, and knowledge—or is it a "GPT wrapper" with a logo? Can you inject domain-specific logic, tone, formatting, and guardrails?

- Ask: Are we building 'our' knowledge with 'our' assistant, or just the interface?
- Flag: You are not allowed to edit the "System Prompt" or "Meta Prompt."

OCEAN BEACH
CONSULTING

Truth (Artificial Sophist)

An AI assistant is effective only if it directly assesses the integrity and honesty of the response.

- Ask: Is the AI providing a statistical guess based on probability, or a verifiable insight?
- Flag: The AI uses "corporate speak" or "workslop" to mask a lack of data.

Strategic Outcome

Did the interaction result in skill gain for the learner, or just a chat log? Does it produce a tangible artifact: a draft, a code block, a plan, a decision?

- Ask: After 10 minutes of chatting, does the learner have a new skill, or just a transcript?
- Flag: The interaction ends with "Hope that helps!" instead of a check for understanding or a concrete next step.