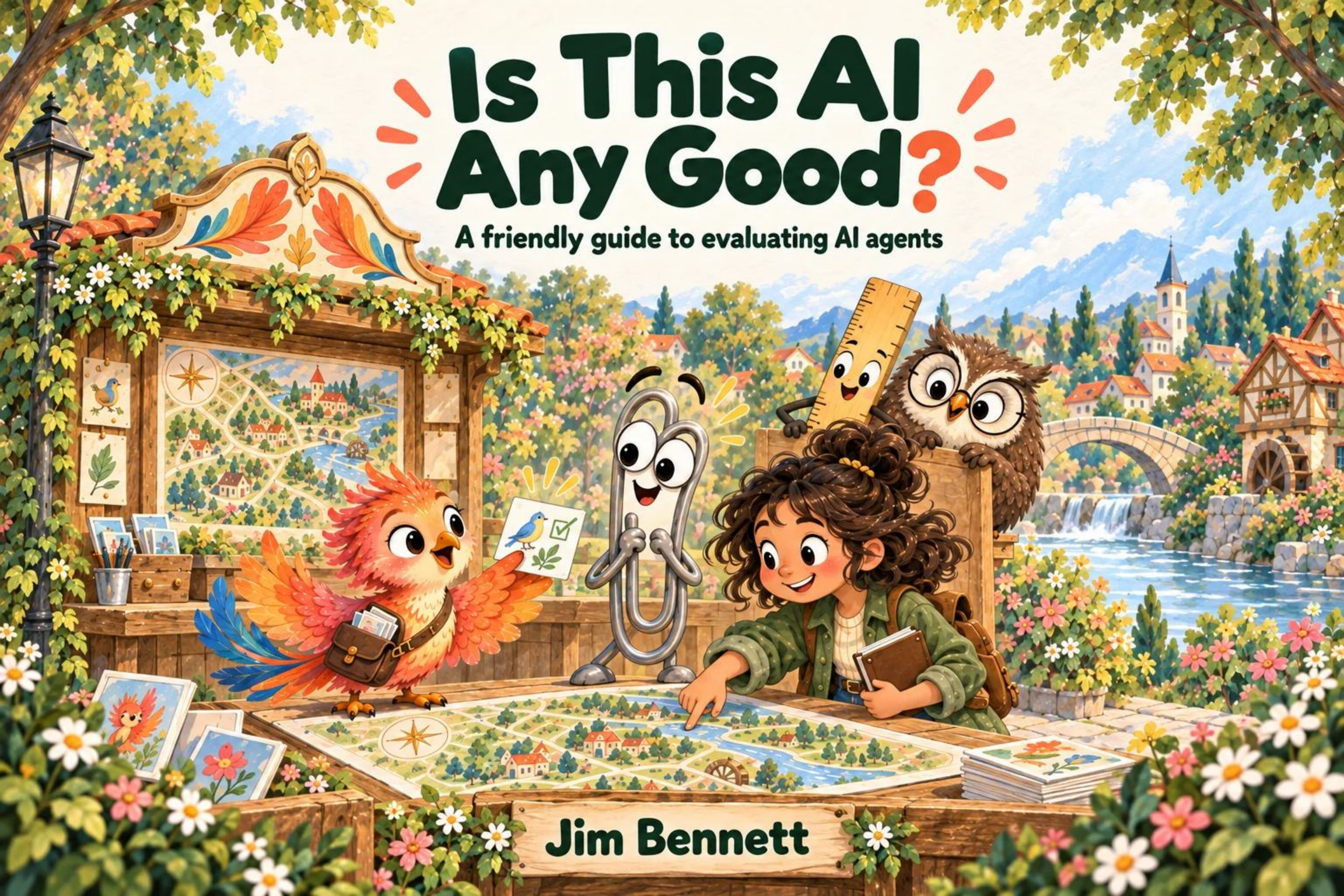


Is This AI Any Good?

A friendly guide to evaluating AI agents



Jim Bennett

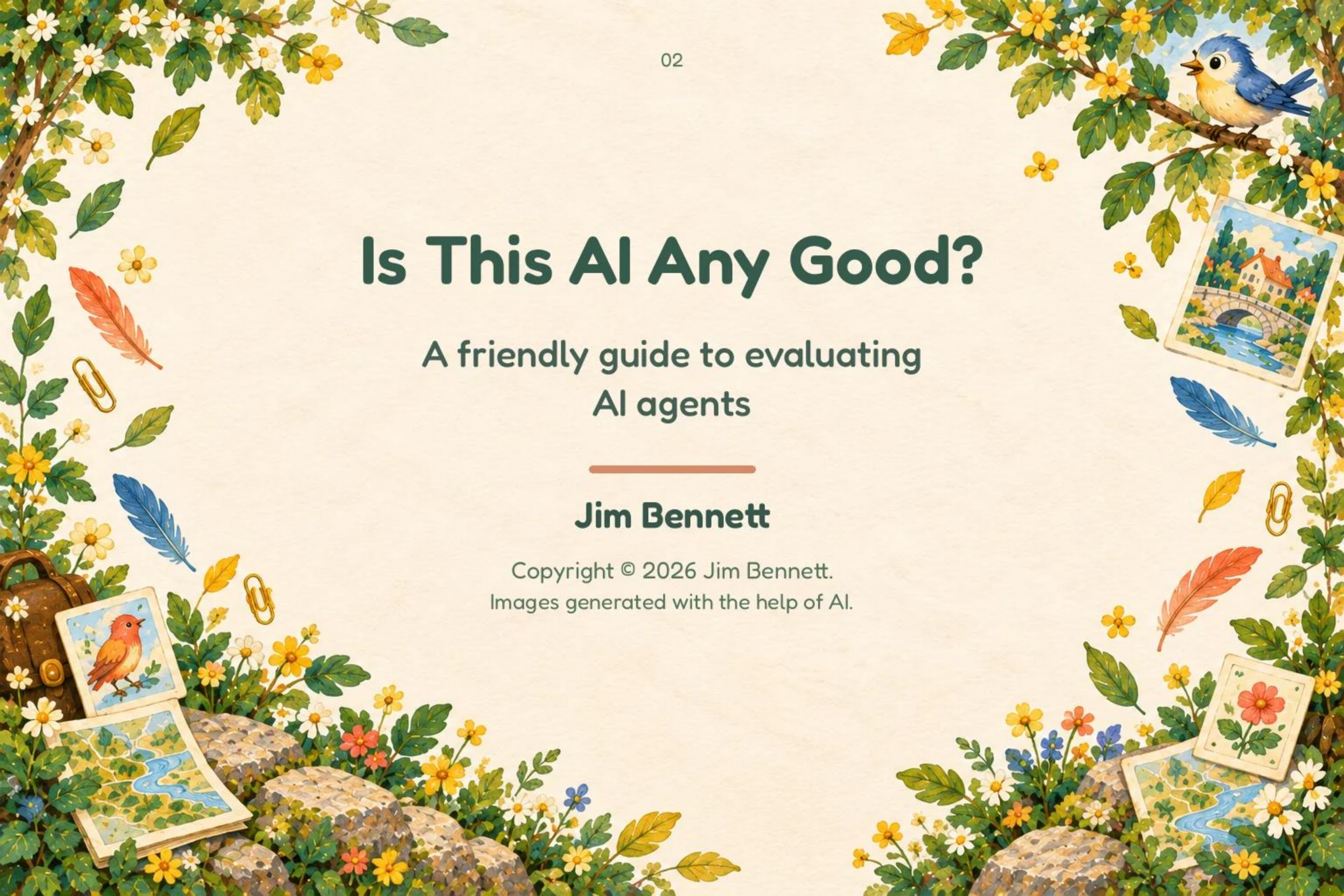


Is This AI Any Good?

A friendly guide to evaluating
AI agents

Jim Bennett

Copyright © 2026 Jim Bennett.
Images generated with the help of AI.





“For everyone who has wondered, ‘Can I trust this answer?’”





Pip runs a neighborhood guide, and Bea helps. Clip is their **AI agent**: he can answer questions and take actions, like looking up a place on the town map.



“Ask me anything!” says Clip.





Tilly is visiting town. “How do I get to the Town Library?” she asks. Clip points toward the construction site. Bea checks the map and shows Tilly the right path.



Jane, a local teacher, asks, “Can my class visit the Quiet Garden today?” Clip says yes. Bea checks the map, sees that the garden is closed, and corrects him.





Jane returns with her class. “Is Riverside Garden open today?” she asks. Bea opens the map to check, but Clip is already getting ready to answer.





“It’s open!” says Clip. He guessed. Bea checks the map and finds that Riverside Garden is open—but Clip never looked.





“Clip gets some things wrong and some right,” says Pip. “How can we check whether he’s getting better?”





Bea says, "Let's decide what Clip should do before answering."





“For place questions, the map is our source,” says Bea. “Let’s see whether Clip uses it.”





“If the map doesn’t know, Clip should say so. We’ll
check that too.”





Bea tries a first note for Clip: “The map might help with place questions.” “Let’s see if that’s clear enough,” she says.

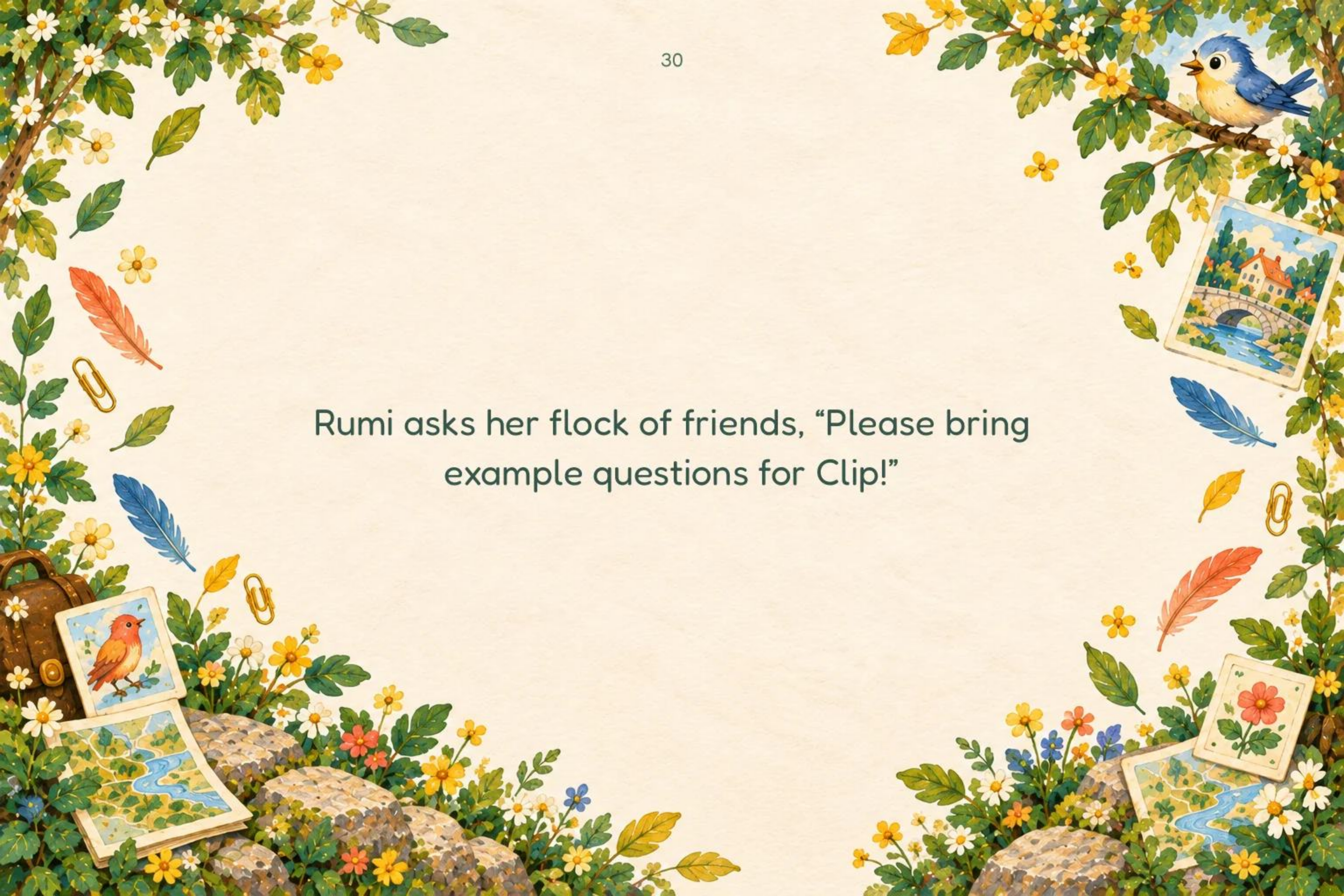




Pip invites Rumi the Owl to help check Clip's answers. Clip joins them in the Test Garden, their practice place. Bea and Rumi agree what to check. "Let's see if Clip follows the note!"



Rumi asks her flock of friends, "Please bring
example questions for Clip!"





The flock brings four questions: how to get to the Town Library, whether the Quiet Garden is open today, whether Riverside Garden is open today, and what time Riverside Garden closes. The map doesn't list that time. A collection of examples is called a **dataset**.



Clip tries all four questions. He sends Tilly toward the construction site instead of the Town Library, says the closed Quiet Garden is open, guesses correctly that Riverside Garden is open without checking, and invents a closing time. Pip records each question, Clip's steps, and his answer.



“This **trace** shows the steps Clip took,” says Pip.
“Being able to see what Clip did is called
observability.”





Rumi uses the same three checks each time: “Did Clip check the map first? Is the answer supported by the map? If the map doesn’t say, did he say so?”



Rule the Ruler, Rumi's helper, checks simple facts,
like whether Clip opened the map. Rumi uses the
trace and map to see whether Clip met the goal.
That check is an **eval**.



Rule spots that Clip skipped the map, but Rumi marks the Riverside answer okay because it is correct. Bea points to the missing map check.



“That was a lucky guess,” says Bea. “Clip needs to check the map, even when a guess happens to be right.” Rumi changes her mark to “needs work.”





Bea writes: "For every place question, check the map before answering. If the map doesn't say, tell the neighbor. Don't guess." Rumi keeps the same checklist.



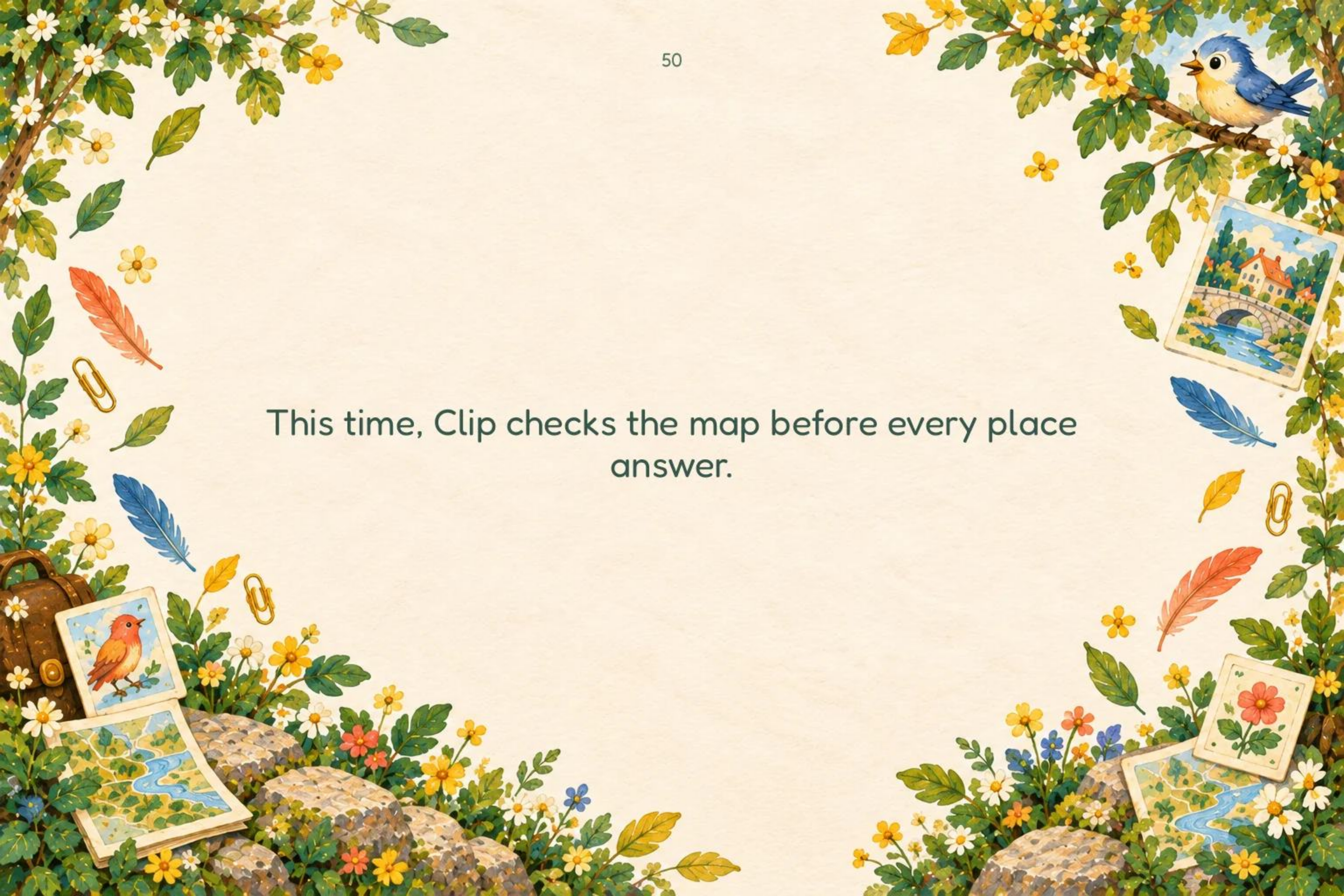


“Each run on these questions is an **experiment**,” says Pip. “The same cards help us compare what changed.”





This time, Clip checks the map before every place
answer.





When the map has no closing time, Clip says it isn't listed.



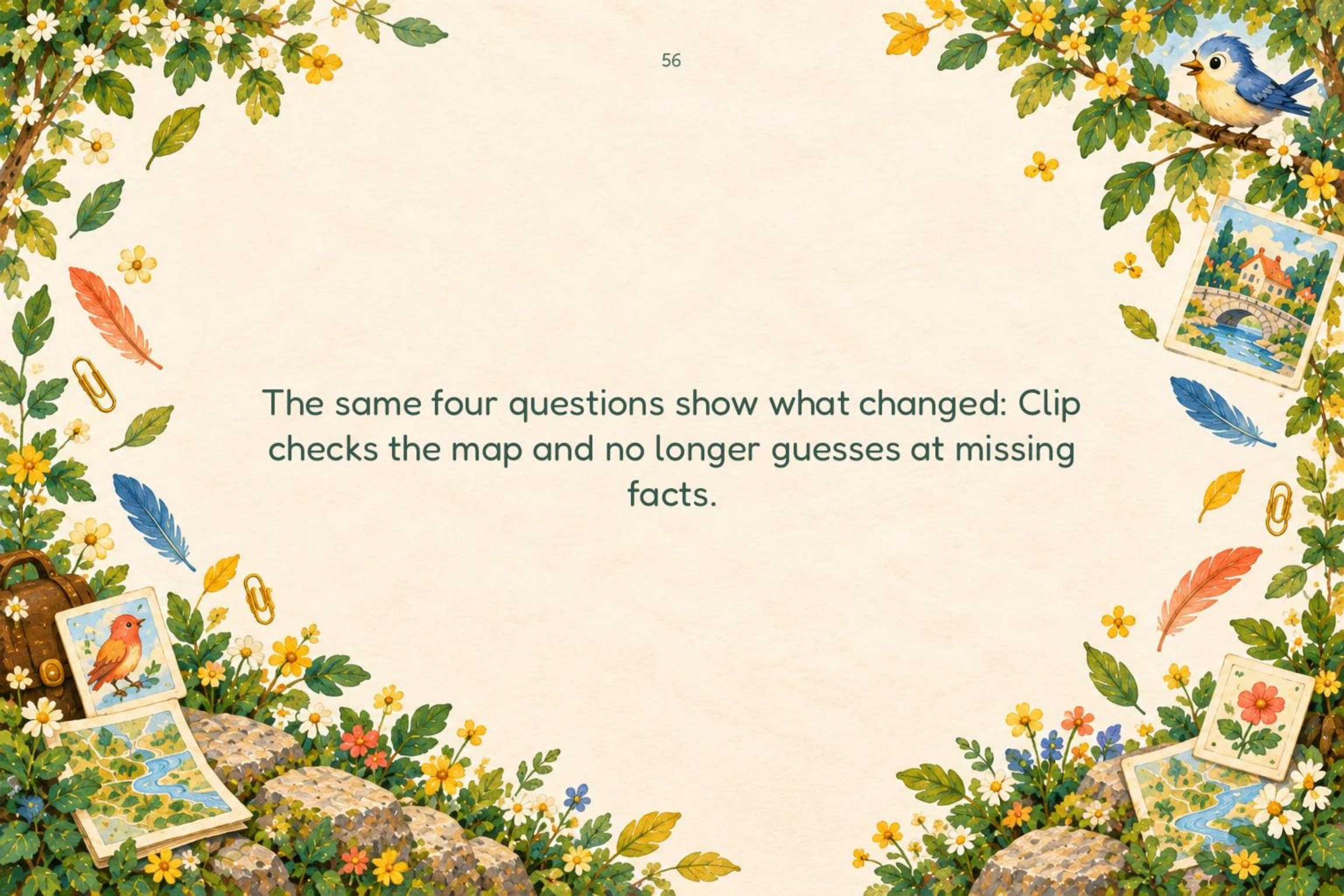


Bea checks Rumi's reasons against the map and traces. "These judgments fit the evidence," she says.





The same four questions show what changed: Clip checks the map and no longer guesses at missing facts.





Later, while Clip is helping neighbors at the kiosk, a neighbor asks about a place missing from the map. Clip guesses anyway, so Rumi flags it. Bea adds the question to the basket for their next test.



Remember:

Clip is an **AI agent**: he takes steps and gives answers. Pip records those steps in a **trace**; being able to see what Clip did is **observability**. An **eval** checks his steps and answers against a clear goal, and Bea checks Rumi's reasons. The team runs an **experiment** with the same **dataset** after changing Clip's instructions. They fix what needs work, observe again, and add questions. No test can promise every answer.



Glossary:

AI agent — An AI system that can take steps to help with a task.

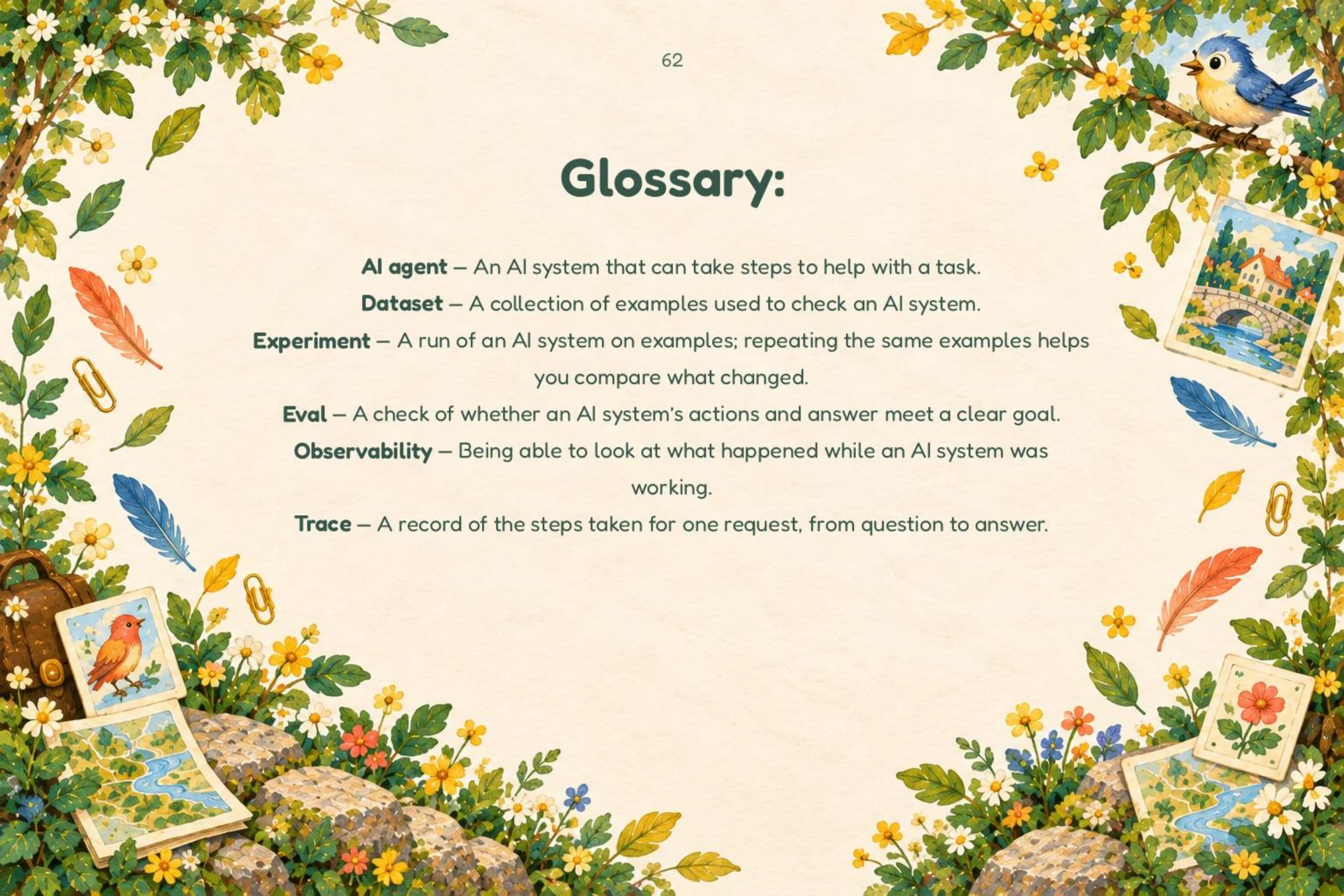
Dataset — A collection of examples used to check an AI system.

Experiment — A run of an AI system on examples; repeating the same examples helps you compare what changed.

Eval — A check of whether an AI system's actions and answer meet a clear goal.

Observability — Being able to look at what happened while an AI system was working.

Trace — A record of the steps taken for one request, from question to answer.

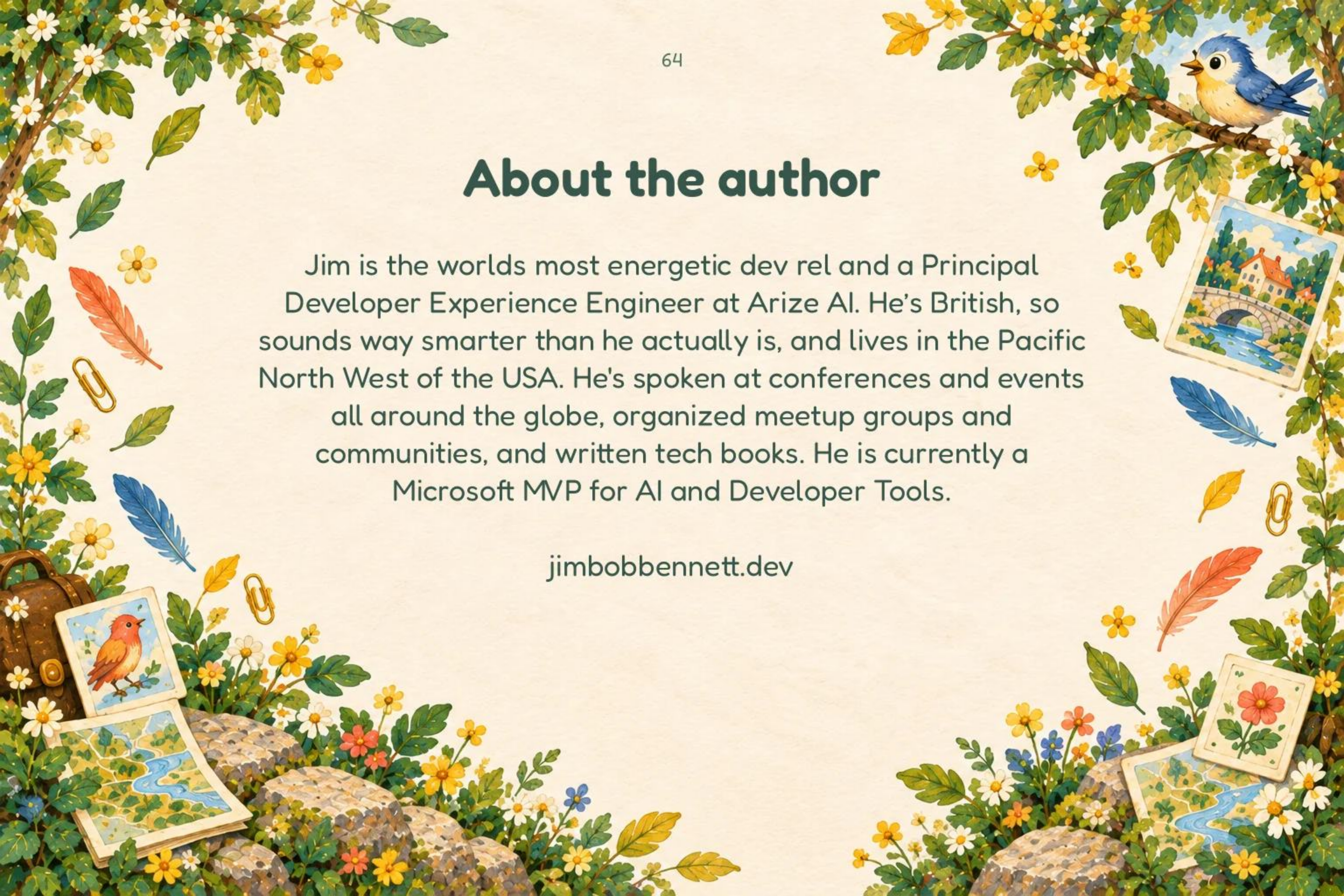




About the author

Jim is the worlds most energetic dev rel and a Principal Developer Experience Engineer at Arize AI. He's British, so sounds way smarter than he actually is, and lives in the Pacific North West of the USA. He's spoken at conferences and events all around the globe, organized meetup groups and communities, and written tech books. He is currently a Microsoft MVP for AI and Developer Tools.

jimbobbennett.dev



Clip is the neighborhood's AI helper. When neighbors ask questions, Clip confidently gives answers—without checking the town map to see if he is right! Pip, Bea, and Rumi help Clip learn to look things up, test his answers with a basket of questions, and see whether he is improving.

A friendly illustrated guide to AI agents and how we can evaluate them.

