Agent5

A TECHNICAL FAQ

Under the hood: the sources, the engine, and the mathematics of being right.

Agent5 is a free daily game: five sharp questions about what AI does next, a confidence on each, and a score when reality resolves. This is the long version: how questions are sourced and written, how they're kept resolvable, and exactly how your forecasts are scored. Written for people who like the details.

LIVE NOW · EARLY ACCESS

GAME · NEVER MONEY

getagent5.com · A daily AI-prediction game

PART I

The engine: eyes, a brain, and a gate

Three parts do the work: eyes that read the day's AI news, a brain that turns it into resolvable questions, and a human gate that decides what ships and how it resolves. The intelligence isn't that an AI writes the questions. It's the discipline around what gets thrown away.

Q1 Where do the questions come from?

Every night the engine ingests the last 24 hours of AI news from a curated set of around eleven feeds, deduplicated by URL into a single pool, pulling from first-party lab sources where launches break first, plus the AI desks of the major technology press for funding, policy, hardware, and analysis:

FIRST-PARTY / LABS

OpenAI | Google Research | Hugging Face

TECH PRESS / AI DESKS

TechCrunch | The Verge | VentureBeat | Ars Technica

MIT Technology Review | Wired | MarkTechPost | Tom's Hardware

Source breadth is question quality. A feed list skewed to model releases would produce a game that only ever asks about model releases. This spread keeps every category fed: hardware desks feed the hardware questions, the funding press feeds the deals, the labs feed the launches.

Q2 Who actually writes the questions?

A language model reads the day's news and drafts candidates, but a raw headline isn't a question. Each draft is forced to carry the things that make a forecast **resolvable**: a precise yes/no claim, the context behind it, explicit resolution criteria, a resolution date, a named resolution source, a timeframe (short / medium / long), and a category. A question with no clean way to settle it is worthless, so the engine won't emit one.

Q3 How do you keep them sharp?

Drafts pass through curation and a normalizing filter. Curation balances the daily set across categories and timeframes and designates the one shared **Headliner** everyone plays. The filter strips anti-pattern questions (the vague, the unfalsifiable, the "will AI be a big deal this year" non-questions) before a human ever sees them. What survives is small, balanced, and falsifiable.

Q4 Who decides what ships?

A person. The engine proposes; a human selects the day's five and ships them. That's the gate: the pipeline does the heavy lifting, but nothing reaches players unreviewed. The cost of a bad question is paid in trust, so it's worth a human minute.

Q5 How are questions resolved?

On a question's resolution date, a model with live web access reads the named resolution source and proposes an outcome (YES, NO, or VOID) with linked evidence. That proposal goes to a human, who approves or overrides it. The system **never auto-resolves unsupervised**: a wrong resolution is worse than a late one, because it corrupts every score attached to it. Only human-approved resolutions ever touch your score.

Q6 What does the game cover?

You pick the corners of AI you care about; the Headliner is shared by everyone. The categories:

- Models & Releases
- Funding & Deals
- Benchmarks
- Agents & Products
- Hardware & Compute
- Policy & Drama

Q7 So what makes the engine "smart"?

Not the drafting. That's the easy part. It's the discipline: diverse sourcing, forced resolution criteria, a filter that kills vague questions, evidence-backed resolution, and a human gate at both ends. The result is closer to a small daily forecasting tournament with a curated, falsifiable question set than to a feed of hot takes.

PART II

The score: the mathematics of being right

Agent5 doesn't score you on being right. It scores you on being **calibrated**: how close your stated probability was to what actually happened. The tool for that is a proper scoring rule.

Q8 How is a single prediction scored?

You answer each question with a probability **p** that the answer is YES (e.g. **0.80** = "80% likely YES"). When it resolves to an outcome **o** (1 for YES, 0 for NO), your score for that question is:

$$\text{Score} = (1 - (p - o)^2) \times 100$$

A Brier score, rescaled so higher is better and the range is 0–100.

The raw **Brier score** is the squared error, $(p - o)^2$, the workhorse of forecasting. Subtracting it from 1 and scaling by 100 flips it into points where a perfect call earns 100 and the worst possible call earns 0.

HOW CONFIDENT? 78% · Probably yes

Q9 Show me the numbers.

Confident and right	$p = 0.82$, resolves YES → $1 - (-0.18)^2$	96.8
Cautious and right	$p = 0.55$, resolves YES → $1 - (-0.45)^2$	79.8
Hedged, still right	$p = 0.45$, resolves YES → $1 - (-0.55)^2$	69.8
Confident and wrong	$p = 0.90$, resolves NO → $1 - (0.90)^2$	19.0
Pure fence-sit	$p = 0.50$, any outcome → $1 - (0.50)^2$	75.0

Notice the asymmetry of conviction: confidence is **rewarded** when you're right and **punished** when you're wrong, while a flat 50/50 always banks 75 no matter what happens.

Q10 The calibration table.

For a YES-leaning call of probability p, here's what you score if the question resolves YES versus NO:

YOUR CALL (P)	IF IT'S YES	IF IT'S NO	CONVICTION AT RISK
50%	75.0	75.0	0.0
60%	84.0	64.0	20.0
70%	91.0	51.0	40.0
80%	96.0	36.0	60.0
90%	99.0	19.0	80.0
95%	99.8	9.8	90.0
99%	100.0	2.0	98.0

"Conviction at risk" is the gap between the two outcomes: the points you're staking by stepping away from 50/50. The table is mirror-symmetric for NO-leaning calls.

Q11 Why this formula, and not a log score or a peer score?

Two design choices a forecaster will want named outright:

Quadratic (Brier), not logarithmic. Both are strictly proper. A log score punishes confident errors far more brutally: a 100%-wrong call heads toward minus infinity. Brier is **bounded**: the worst possible single question costs you down to 0, never below. For a game you'll play hundreds of times, bounded and forgiving beats savage and unbounded. No single bad day wrecks your record.

Absolute, not peer-relative. Your score depends only on your probability and the outcome, not on how the crowd answered. A 96.8 means the same thing on a question 50 people played and one 5,000 played, and you can verify it yourself with a calculator. Any crowd context is shown alongside your absolute score, never as a term inside it.

> **The trade-off we accept:** Brier is gentle near 50%. That's deliberate, not an oversight (see Q13).

Q12 Can I game the scoring?

No. That's the entire point of a proper scoring rule. A rule is **strictly proper** when your expected score is maximized only by reporting your true probability. There's no hedging trick, no systematic lean that beats honesty: if you genuinely believe 70%, then 70% is your highest-expected-score answer, and shading to 90% or 50% to "play the meta" lowers it.

> For an event you believe has probability **q**, reporting **p** gives expected score $100 - 100 \cdot [\, q(1-p)^2 + (1-q)p^2 \,]$, which is maximized exactly at $p = q$. Good forecasting and a good score are the same objective.

Q13 Then why does a wrong-leaning 45% still score ~70?

Because near 50%, the squared-error penalty is tiny; you barely committed. A 0.45 call on a YES outcome sits only 0.55 from the truth; squared, that's about 0.30, so you keep roughly 70 points. This is the gentleness from Q11, and it's intended: low-conviction calls move your score very little either way. The points live in **conviction**. To climb, you have to be confident and right, repeatedly.

Q14 What happens to a VOID question?

Some questions can't be fairly settled: the event is cancelled, redefined, or the source goes quiet. Those resolve **VOID** and are scored neutrally: no points, no penalty, for anyone. A forced or guessed resolution would poison every prediction attached to it, so when in doubt the question is voided rather than fudged.

Q15 How do single scores add up?

Per-question points accumulate into a running track record, and your rank is built on that record over time. Because the score is absolute and proper, your record is a genuine **calibration history** (a measure of how well-tuned your AI instincts are), not a function of how many people you happened to out-guess on any given day.

PART III

Why it's built this way

Q16 What is Agent5 actually rewarding?

Calibration over bravado. Anyone can be loud; the game rewards the person who says 70% and is right about 70% of the time. Over a week of hot takes, the loudest voice wins the room. Over a few hundred scored questions, the best-calibrated one wins the leaderboard, and learns exactly how sharp their instincts really are. That's the whole game.

> **Agent5 · getagent5.com** · A game, never money. Five questions, every day.