Stochasticity in Agentic Evaluations: Quantifying Inconsistency with Intraclass Correlation

ZAIRAH MUSTAHSAN, ABEL LIM, MEGNA ANAND, SAAHIL JAIN, BRYAN MCCANN

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

you.com

ABSTRACT

Agentic evaluation scores are often inconsistent across runs;
Intraclass Correlation (ICC) makes this inconsistency measurable and reportable.

- Single-run scores can reflect sampling luck as much as capability.
- Use Intraclass Correlation Coefficient (ICC) to quantify trial-to-trial consistency per item.
- Result: Consistency varies widely (e.g., FRAMES ICC ~0.50–0.71, GAIA ICC ~0.30–0.77) and requires up to 32+ trials to stabilize depending on task.
- ICC should be reported and studied alongside accuracy to fully understand Agent consistency.

INTRACLASS CORRELATION (ICC)

The ICC quantifies the proportion of total variance attributable to differences between tasks.

- It can be understand as a function of both dataset difficulty (variance between tasks) and agent consistency (variance within tasks across repeated runs).

We apply ICC formula (variance decomposition) under ICC(1,1) one-way random effects model

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Quick read:

- higher ICC ⇒ more of the observed variation is explained by task difficulty (stable agent)
- lower ICC ⇒ more is run-to-run noise (unstable agent).

Level	Accuracy		95% CI		Between Var		ICC	
	GPT-4o	GPT-5	GPT-4o	GPT-5	GPT-4o	GPT-5	GPT-4o	GPT-5
Level 1 (53 Q)	22.7%	62.3%	[14.0%, 31.4%]	[52.9%, 71.7%]	0.100	0.185	**0.561**	**0.774**
Level 2 (86 Q)	23.2%	54.2%	[15.8%, 30.6%]	[44.9%, 63.5%]	0.119	0.187	**0.662**	**0.745**
Level 3 (26 Q)	6.6%	44.2%	[1.0%, 12.2%]	[28.1%, 60.4%]	0.019	0.160	**0.304**	**0.629**

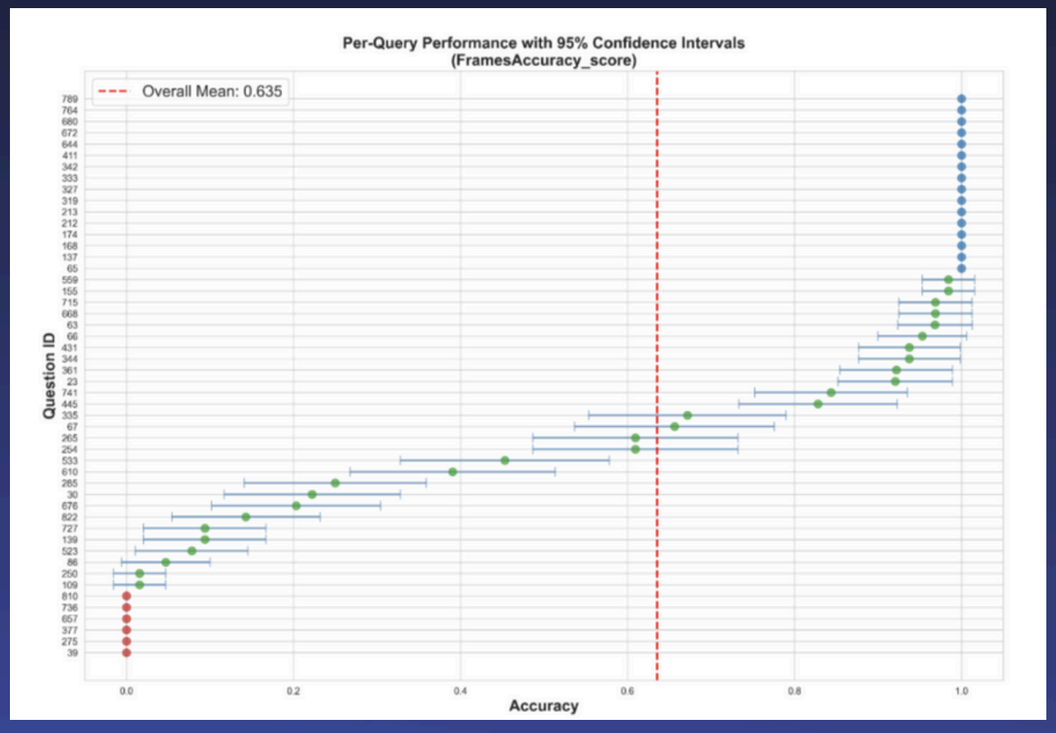
* Both GPT-4o and GPT-5 were evaluated with web search enabled.

Table 2: ICC and variance decomposition across GAIA levels (full validation set) and models (64 trials per question).

Model	Accuracy	95% CI	Between Var	ICC
GPT-5 search	77.31%	[68.86%, 85.77%]	0.088	**0.496**
GPT-4o search	63.54%	[51.70%, 75.38%]	0.174	**0.735**
GPT-4o	38.16%	[26.40%, 49.91%]	0.171	**0.712**
Claude 4.5 Haiku	68.37%	[57.58%, 79.17%]	0.144	**0.655**
Claude 4.5 Sonnet	66.44%	[55.20%, 77.68%]	0.156	**0.689**
Gemini 2.5 Pro	62.34%	[50.60%, 74.09%]	0.174	**0.713**
Qwen3-235b-a22b	34.22%	[23.53%, 44.91%]	0.169	**0.617**
Deepseek-v3p1	44.75%	[33.13%, 56.37%]	0.157	**0.663**

* GPT-5 & Claude family were evaluated with web search, GPT-4o with and without web search and others without web search.

Table 3: ICC and variance decomposition on FRAMES (n=50, 64 trials per question).



METHOD / DATASETS

- Evaluation dataset: FRAMES/GAIA
- Models/Agents: GPT-4o / GPT-5 with and without search tool equipped
- Methodology: We run a large number of trials for each model and compute the success rate across trials and queries.
- We compute within-query variance and inter-query variance for these trials and compute ICC

KEY FINDINGS

- ICC varies dramatically with task structure:
  - FRAMES: ICC = 0.4955–0.7118 (across models)
  - GAIA: ICC = 0.304–0.774 (across models)
- ICC estimation improves with number of trials, and is dependent on complexity:
  - ICC converges by 8-16 trials for structured tasks
  - >32 trials may be required for ICC convergence for complex reasoning
- Practical implication for agent design:
  - Accuracy improvements are only trustworthy if ICC also improves.
  - To improve reliability, recommend to increase dataset size and use sampling on when dataset size is maximal
- Recommend reporting:
  - ICC (consistency)
  - Within-task variance (or uncertainty summary)
  - Number of trials per task (resampling budget)