Agent Consistency Research - Summary Jan 26, 2026

COMPREHENSIVE REVIEW

✅ SCOPE VALIDATION: Your Work is GOOD & PUBLISHABLE

I found the **most related paper**: **τ-bench** (Yao et al., 2024, same author as ReAct!). Here's how you compare:

Aspect	τ-bench	Your Paper
Main finding	pass@1=60% → pass@8=25%	Consistent (≤2 seqs)=92% vs Inconsistent (≥8)=57%
Metric	pass^k (did it succeed k times)	Action sequence diversity, divergence point
WHERE variance occurs	❌ Not analyzed	✅ **69% at step 2** (NOVEL)
Path length analysis	❌ Not analyzed	✅ **Short=90%, Long=43%** (NOVEL)
Correctness correlation	Implicit	✅ **Explicit 35pp gap** (NOVEL)
Question type effects	❌	✅ Bridge vs Comparison (NOVEL)

Your unique contributions over τ-bench:

1. **Divergence point analysis** — WHERE agents first disagree
2. **Path length as predictor** — Short paths = confidence
3. **Consistency-correctness relationship** — Quantified correlation
4. **Question structure effects** — Different patterns for different types

---

🔧 FIXES MADE TO PAPER

1. **Added τ-bench citation** — CRITICAL (was missing the most related work!)
2. **Added "AI Agents That Matter" citation** — Important related work
3. **Updated Related Work section** — Now properly differentiates from τ-bench
4. **Updated Discussion** — Clearer positioning vs prior work

---

📋 TODO FOR WORKSHOP-READY

Task	Priority	Time
Wait for Claude + GPT-4o experiments	HIGH	Tonight
Add multi-model comparison table	HIGH	1 hour
Add 1-2 figures (bar chart of 35pp gap)	MEDIUM	1 hour
Proofread	LOW	30 min

📋 TODO FOR PREPRINT-READY

Task	Priority	Effort
Temperature ablation (0.0 vs 0.7)	HIGH	2-3 hours
Add figures (3-4 total)	HIGH	2 hours
GitHub repo with code	HIGH	1 hour
Longer related work	MEDIUM	1 hour
Example trajectories in appendix	LOW	1 hour

🧪 ADDITIONAL EXPERIMENTS (Optional but Valuable)

Quick wins (do these):

1. **Temperature ablation** — Run 20 questions at temp=0.0 vs 0.7
    - Hypothesis: temp=0 eliminates variance but may hurt performance
    - Easy to run, strengthens paper significantly
2. **Multi-model comparison** — You're already doing this! Key question:
    - Do all models show step-2 divergence pattern?
    - Does the 35pp gap hold across models?

For future work (mention in paper):

3. Semantic retrieval vs lexical search
4. Majority voting intervention
5. Early-stopping when variance detected

📊 FIGURES TO ADD

1. **Bar chart:** 92% (consistent) vs 57% (inconsistent) — THE HEADLINE
2. **Histogram:** Unique sequences distribution (bimodal)
3. **Scatter plot:** Path length vs correctness (r=-0.34)
4. **Model comparison table** (once experiments complete)

⚠️ HONEST ASSESSMENT

Strengths:

- Novel angle on well-recognized problem (τ-bench showed WHAT, you show WHERE/WHY)
- Clean experimental setup
- Actionable findings (monitor step 2, track path length)
- Strong headline number (35pp gap)

Weaknesses:

- Single benchmark (HotpotQA) — mention as limitation
- Lexical search only — could introduce variance semantic wouldn't
- Temperature fixed at 0.7 — ablation would strengthen

Verdict: Solid workshop paper, can become strong preprint with multi-model + temperature ablation.

EXPERIMENT STATUS (Jan 26 Night)

Model	Questions	Runs	Status
Llama 3.1 70B	100	1000	✅ **DONE** (98.2% success)
Claude Sonnet 4.5	100	1000	✅ **DONE**

Model	Questions	Runs	Status
GPT-4o	26	260	🔄 Running (~1 hour left)

KEY FINDINGS (Llama 3.1 70B)

Headline Numbers

- **35 percentage point gap**: 92% correct (consistent) vs 57% correct (inconsistent)
- **4.4 unique sequences** per 10 runs on average
- **69% divergence at step 2** — first search query determines trajectory
- **Path length correlation**: r = -0.34 with correctness

Bucket Analysis

Bucket	n	Avg Correct
Low variance (≤30%)	21	**90.0%**
High variance (>80%)	22	66.8%
Few unique seqs (≤2)	25	**92.0%**
Many unique seqs (≥8)	15	57.3%
Perfectly consistent (1 seq)	14	85.7%

Path Length Pattern

- **Perfectly Consistent (1 sequence, n=14):** Avg 3.4 steps, 85.7% correct
- **Highly Inconsistent (9-10 sequences, n=10):** Avg 7.8 steps, 43% correct

---

PAPER FILES

- `paper_draft_v2.tex` — Updated with τ-bench citation, improved Related Work
- `references.bib` — Updated with τ-bench and "AI Agents That Matter" citations

---

NEXT STEPS (Morning)

1. Run Claude health check
2. Wait for GPT-4o to finish (~1 hour)
3. Run GPT-4o health check
4. Run multi-model comparison analysis
5. Update paper with 3-model results
6. Create figures
7. Post to ArXiv

TIMELINE

- **Jan 27:** Complete experiments, multi-model analysis
- **Jan 28:** Post to ArXiv
- **Feb 13:** Submit to ICML 2026 Workshop