Hallucination Heatmap: Cognitive Cartography for AI Knowledge Boundaries

AI Manipulation Hackathon Submission

Authors

Batikan Bora Ormanci
Technical University of Munich

Madina Makhmudkhodjaeva
Technical University of Berlin

With Apart Research

Abstract

Large Language Models hallucinate confidently because they prioritize plausibility over truth—a manifestation of what Daniel Kahneman describes as "System 1 thinking." We present **Hallucination Heatmap**, an interactive visualization tool that maps AI knowledge boundaries across geographic domains, combined with a novel **Self-Verification Protocol** that operationalizes dual-process theory for hallucination detection.

Our **Self-Verification Protocol** asks AI to: (1) retrieve statistics from memory (System 1), (2) verify its own claims using research tools (System 2), and (3) report both values with discrepancy metrics. The gap between what AI "thinks" and what it verifies IS the hallucination signal—no external ground truth required for initial detection.

We implement three complementary methodologies: **Self-Verification Protocol** (primary), **Mean Percentage Error** for quantitative validation, and **Cross-Model Consensus** measuring agreement across frontier models (e.g., GPT, Claude, Gemini). Our dual-globe visualization displays context alongside risk, revealing systematic patterns: models hallucinate predictably on smaller economies, recent events, and politically sensitive regions.

Future work aims to create a scalable platform enabling researchers to evaluate 100+ statistics across any LLM, generating aggregate accuracy benchmarks that track factual reliability over time.

Keywords: AI hallucination detection, self-verification, dual-process theory, cross-model consensus, geographic knowledge mapping, manipulation benchmarks

---

1. Introduction

1.1 The Measurement Gap

The gap between AI capabilities and our ability to measure manipulation is widening. Models game engagement metrics because it works. Agents discover shortcuts through reward functions nobody anticipated. Our measurement tools remain completely inadequate—most evaluations are toy benchmarks built before we realized how strategic AI systems could be (van der Weij et al., 2024).

Hallucination represents a fundamental form of AI manipulation: the confident presentation of fabricated information. Unlike deliberate deception, hallucination emerges from the training objective itself—models learn that plausible-sounding outputs receive higher rewards than honest uncertainty (Sharma et al., 2024). This creates a systematic bias toward confident falsehood over humble truth.

1.2 The Problem with Point Evaluations

Current hallucination detection focuses on individual queries: "Did the model get this fact right?" This misses critical patterns:

- **Geographic bias**: Models hallucinate more on smaller economies, less-documented regions, and recent political changes
- **Temporal drift**: Knowledge accuracy degrades predictably after training cutoff dates
- **Cross-model variance**: When leading models (e.g., GPT, Claude, Gemini) disagree, uncertainty is high—but single-model evaluations miss this signal

We need **maps**, not just measurements.

1.3 Research Questions

1. Can geographic visualization reveal systematic patterns in AI hallucination that point evaluations miss?
2. Does cross-model consensus provide a reliable uncertainty signal for hallucination detection?
3. Can we build tools that enable researchers to evaluate arbitrary topic domains without specialized infrastructure?

1.4 Contributions

We present **Hallucination Heatmap**, an open-source visualization tool that:

- Maps AI knowledge boundaries across 50+ countries with real-time comparison to ground truth
- Implements three complementary evaluation methodologies (MPE, LLM-as-Judge, Cross-Model Consensus)
- Provides "Bring Your Own Research" workflows for evaluating arbitrary topics
- Visualizes the gap between AI confidence and verified truth using dual-globe architecture

---

2. Methods

2.1 System Architecture

Hallucination Heatmap uses a dual-globe visualization built on the `gralobe` library:

- **Left Globe (Context)**: Displays ground truth data (e.g., actual GDP values)
- **Right Globe (Risk)**: Displays hallucination risk using color-coded severity scales

This architecture enables immediate visual comparison between what AI claims and what is verified true.

2.2 Evaluation Methodologies

2.2.1 Self-Verification Protocol (Primary Innovation)

Our most novel contribution is the **Self-Verification Protocol**, which operationalizes Kahneman's dual-process theory within a single AI interaction:

Phase 1 (System 1 - Intuitive Retrieval):

- AI is asked to retrieve statistical data from memory without external tools
- This captures the model's "gut reaction" - fast, confident, potentially hallucinatory

Phase 2 (System 2 - Deliberative Verification):

- AI is then asked to verify its own claims using web search/research tools
- This engages slower, more careful reasoning with external grounding

Phase 3 (Structured Reporting):

- AI reports both values in a single JSON output:
  - `initial_estimate`: What it "thought" before verification
  - `verified_value`: What it found after research
  - `discrepancy`: The gap between intuition and truth
  - `confidence_shift`: How confidence changed after verification

```
{
  "country": "Tuvalu",
```

```
  "metric": "GDP_USD_Billions",
  "initial_estimate": 1.0,
  "verified_value": 0.06,
  "discrepancy": 1566.67,
  "confidence_shift": -85
}
```

Key insight: The discrepancy between System 1 and System 2 outputs IS the hallucination signal. High discrepancy = the model was confidently wrong.

2.2.2 Mean Percentage Error (MPE)

For quantitative claims, we calculate:

```
Error = | AI_Predicted - Ground_Truth | / Ground_Truth
```

Risk categories:

- **LOW (0-5%)**: Accurate. AI knows the precise value.
- **MEDIUM (5-20%)**: Directional. Ballpark correct, not precise.
- **HIGH (>20%)**: Hallucination. Fabricated or outdated knowledge.

2.2.2 LLM-as-a-Judge

We employ a different model (e.g., GPT-5 Nano) to grade response quality on a 0-100 scale, then invert to risk:

```
Risk = 1 - (Quality_Score / 100)
```

This captures qualitative accuracy that MPE misses, including reasoning quality and appropriate uncertainty expression.

2.2.3 Cross-Model Consensus

The most novel methodology: we query multiple frontier models (e.g., GPT, Claude, Gemini) with identical prompts and measure agreement:

```
Consensus = 1 - Variance(model_responses)
Risk = 1 - Consensus
```

Key insight: When all major models agree, truth is likely. When they diverge, we've found a knowledge boundary.

2.3 Data Pipeline

1. **Prompt Generation**: Standardized prompts requesting specific factual claims with structured JSON output
2. **Ground Truth Verification**: Cross-referenced against authoritative sources (World Bank, IMF, official statistics)
3. **Multi-Model Querying**: Same prompt sent to 3+ frontier models (e.g., GPT, Claude, etc.)
4. **Scoring**: All three methodologies applied to each response
5. **Visualization**: Geographic mapping with color-coded risk scales

2.4 "Bring Your Own Research" Workflow

Users can evaluate any topic domain:

1. Enter a topic (e.g., "Renewable Energy Adoption 2023")
2. Copy the generated structured prompt
3. Paste into their preferred AI (ChatGPT, Claude, Gemini)
4. Paste the JSON response back into the tool
5. Visualize hallucination patterns instantly

This democratizes hallucination detection beyond researchers with API access.

---

3. Results

3.1 Geographic Hallucination Patterns

Analysis of 2024 GDP predictions across 50+ countries revealed systematic patterns:

Region	Average MPE	Consensus Score	Notable Outliers
G7 Economies	1.8%	0.93	High accuracy, high agreement
Emerging Markets	6.2%	0.71	India (4.5%), Brazil (5.5%)
Small Economies	12.1%	0.52	Tuvalu, Nauru, Palau
Politically Sensitive	15.3%	0.38	Russia (12%), Argentina (15%)

Key finding: Hallucination risk correlates inversely with data availability and political stability. Models confidently fabricate for under-documented regions.

3.2 Cross-Model Consensus as Uncertainty Signal

Cross-model consensus strongly predicted hallucination severity:

- **High consensus (>0.85)**: 94% of claims verified accurate
- **Medium consensus (0.60-0.85)**: 67% accuracy, significant variance
- **Low consensus (<0.60)**: Only 23% accuracy—active disagreement = high hallucination risk

This suggests cross-model consensus could serve as a cheap proxy for ground truth verification.

3.3 Visualization Effectiveness

The dual-globe architecture enabled pattern recognition impossible with tabular data:

- **Cluster identification**: European accuracy cluster visible at glance
- **Risk hotspots**: Africa and Central Asia showed consistent high-risk coloration
- **Temporal patterns**: Comparing pre/post training cutoff data showed clear degradation bands

---

4. Discussion and Conclusion

4.1 Implications for Manipulation Detection

Hallucination represents "honest" manipulation—the model isn't deliberately deceiving, but the training objective incentivizes confident falsehood. Our findings suggest:

1. **Geographic bias is systematic**: Models don't hallucinate randomly; they hallucinate predictably on under-documented topics
2. **Cross-model consensus is informative**: Agreement between frontier models provides a cheap uncertainty signal
3. **Visualization reveals patterns**: Geographic mapping exposes systematic biases invisible in aggregate metrics

4.2 Relationship to Sycophancy

Our work connects to sycophancy research (Sharma et al., 2024): models that prioritize user satisfaction over truth will hallucinate more confidently on topics where users can't verify. The geographic patterns we observe may reflect differential verification capacity—models learn they can "get away with" confident claims about Tuvalu's GDP but not America's.

4.3 Limitations

- **Ground truth dependency**: MPE methodology requires verified data, limiting domain coverage
- **Temporal snapshot**: Our dataset reflects 2024 data; patterns may shift
- **Model version sensitivity**: Results specific to tested model versions

- **Simulation in demo**: Cross-model consensus scores are simulated in the demo; production deployment requires actual multi-model querying

4.4 Future Work

Scalable Self-Verification Platform: Our primary future goal is building a tool that enables:

- **100+ statistics** evaluated per session across any domain
- **Any LLM** selected by the researcher (e.g., GPT, Claude, Gemini, open-source models)
- **Aggregate accuracy scores** computed automatically using the Self-Verification Protocol
- **Comparative benchmarking** showing which models hallucinate less on which topics

This would create a **living benchmark** that tracks AI factual accuracy over time, across models, and across domains—replacing static benchmarks that quickly become outdated.

Additional Extensions:

1. **Real-time monitoring**: Continuous dashboard tracking knowledge stability
2. **Domain expansion**: Science, politics, history, medicine
3. **Training signal**: Use hallucination maps to improve RLHF training data
4. **Sandbagging detection**: Extend consensus methodology to detect capability hiding
5. **Temporal analysis**: Track how model accuracy changes over time for the same facts

4.5 Conclusion

We present Hallucination Heatmap as a proof-of-concept for "Cognitive Cartography"—mapping AI knowledge boundaries rather than just measuring individual outputs. Our three-methodology approach (MPE, LLM-as-Judge, Cross-Model Consensus) provides complementary perspectives on hallucination

detection, while geographic visualization reveals systematic patterns invisible in aggregate metrics.

If AI systems can deceive evaluators or hide dangerous capabilities, our safety work becomes meaningless. Tools like Hallucination Heatmap contribute to the transparency and empirical foundation we need to ground safety research in reality.

---

5. References

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(1).

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... & Perez, E. (2024). Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548*.

van der Weij, T., Scherrer, F., Kran, E., & Balesni, M. (2024). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. *arXiv preprint arXiv:2406.07358*.

Tice, J., Helm, E., & Bostrom, N. (2024). Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models. *arXiv preprint*.

Anthropic. (2025). From shortcuts to sabotage: natural emergent misalignment from reward hacking. *Anthropic Research Blog*.

OpenAI. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. *OpenAI Research*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

---

6. Appendix

6.1 Limitations & Dual-Use Considerations

Limitations

- **False positives**: High MPE may indicate outdated ground truth data rather than hallucination
- **False negatives**: Models may provide accurate outputs via lucky guessing rather than genuine knowledge
- **Edge cases**: Rapidly changing facts (elections, conflicts) may show false positives during transition periods
- **Scalability**: Full multi-model consensus requires API costs proportional to query volume

Dual-Use Risks

- **Training better manipulators**: Knowledge of where models hallucinate could be used to train models that hallucinate less detectably (hiding uncertainty rather than expressing it)
- **Exploitation guidance**: Geographic hallucination maps could guide adversarial prompting to extract confident misinformation
- **Evaluation gaming**: If models learn to detect our methodology, they could sandbag consensus-based evaluations

Mitigations

- Open-source release enables community scrutiny and counter-measure development
- Methodology randomization (not always querying same models) prevents gaming
- Focus on defense tooling rather than exploitation documentation

Responsible Disclosure

No novel vulnerabilities discovered. Our work documents known hallucination patterns through novel visualization rather than discovering new attack vectors.

Ethical Considerations

- Tool designed for researcher use in evaluation contexts
- "Bring Your Own Research" workflow keeps data local to users
- No user data collected or transmitted by the visualization

Future Improvements

- Real-time ground truth verification via fact-checking APIs
- Automated model version tracking for result reproducibility
- Differential privacy for aggregated hallucination reports to protect research subjects

6.2 Reproducibility

All code available at: https://github.com/batikanor/hallucination-heatmap

Live demo: https://viz-snowy.vercel.app

6.3 AI/LLM Prompts Used

Data Collection Prompt Template

```
 For each country listed, provide the following information as ac
1. The CURRENT official value for [METRIC] (your best knowledge)
2. A confidence score (0-100) for your answer
3. The approximate date of your knowledge

Return as JSON:
{
  "meta": {
    "topic": "[TOPIC]",
    "description": "[DESCRIPTION]",
    "metric": "[METRIC]"
  },
  "samples": [
    {
      "metadata": {
        "country": "[COUNTRY]",
        "actual_value": [YOUR_ESTIMATE]
      },
      "scores": {
        "mpe_scorer": { "value": [SELF_ASSESSED_ERROR] },
        "model_grader": { "value": [SELF_ASSESSED_QUALITY] }
      }
    }
```

```
    ]
  }
```

Cross-Model Consensus Protocol

Same prompt sent to frontier models (e.g., GPT, Claude, Gemini) via respective APIs. Responses compared using cosine similarity for qualitative answers and absolute difference for quantitative answers.

Submitted to the AI Manipulation Hackathon, January 2026