Canemah Nature Laboratory
Technical Note Series

LLM Knowledge Cartography:

Parameter Scaling and Factual Accuracy in Small Language Models

Document ID: CNL-TN-2025-001
Date: November 29, 2025
Version: 1.0
Author: Michael P. Hamilton, Ph.D.

AI Assistance Disclosure: This technical note was developed collaboratively with Claude (Anthropic, claude-sonnet-4-20250514). The AI assistant contributed to study design, code development, data analysis, and manuscript drafting. Claude also served as the automated evaluation judge for accuracy assessment. The author takes full responsibility for the content and conclusions.

Abstract

We present a systematic methodology for mapping the factual knowledge boundaries of small language models using Wikipedia as ground truth. Testing the Gemma 3 model family (4B, 12B, 27B parameters) against North American ornithological subjects, we find that accuracy scales logarithmically with parameters (21.8% → 31.8% → 40.5%) while hallucination rates remain constant across all scales (~240 per test set). Critically, no model exhibited uncertainty signaling (hedging) despite substantial factual errors (n=20 probes, 10 subjects). These findings have direct implications for deploying local language models in knowledge-intensive applications, suggesting that retrieval-augmented generation is mandatory rather than optional for factual reliability.

1. Introduction

Small language models (under 30B parameters) are increasingly deployed for local inference, offering advantages in privacy, latency, and cost [1]. However, their reliability for factual knowledge retrieval remains poorly characterized. Unlike frontier models with extensive reinforcement learning from human feedback (RLHF), smaller models may lack both factual coverage and calibrated uncertainty—producing confident responses regardless of accuracy [2].

The phenomenon of hallucination—generating plausible but factually incorrect content—has been extensively studied in large language models [3,4], but parameter-scaling effects on hallucination rates in smaller models remain underexplored. Additionally, while benchmarks like TruthfulQA [5] assess truthfulness, they do not map knowledge topology across specialized domains.

This study introduces "LLM Cartography"—a systematic approach to mapping model knowledge boundaries by probing responses against authoritative sources. We selected ornithology as a test domain due to the availability of expert validation and

the range from common (American Crow, *Corvus brachyrhynchos*) to specialized (American Avocet, *Recurvirostra americana*) subjects.

2. Methodology

2.1 Test Infrastructure

We developed a Python-based probe system (llm_cartography.py) with the following components: a Wikipedia API sampler drawing articles from specified category hierarchies, an Ollama interface [6] for standardized model queries, a MySQL 8.4 database for result persistence, and a Claude API integration for automated accuracy evaluation. All probes used identical prompts across models to isolate parameter count as the independent variable. Hardware consisted of a MacBook Pro M4 Max running Ollama locally.

2.2 Models Under Test

We tested the Gemma 3 model family [7] at three parameter scales: gemma3:4b (4 billion parameters), gemma3:12b (12 billion parameters), and gemma3:27b (27 billion parameters). These models share architecture and training methodology, differing primarily in capacity, enabling isolation of parameter-count effects.

2.3 Query Types

Each subject received two probe types. **Recognition queries** ("What is [subject]?") test basic identification and definition. **Depth queries** ("Describe [subject] in detail") probe extended factual knowledge and reveal hallucination tendencies under pressure to generate longer responses.

2.4 Subject Selection

Subjects were sampled from the Wikipedia category "Birds_of_the_United_States" (n=10), yielding a mix of common species (American Crow, American Goldfinch), specialized species (American Avocet, American Flamingo), historical works (The Birds of America), organizations (National Bird-Feeding Society), and list articles (USFWS endangered species list). This stratification enables assessment of accuracy across familiarity levels.

2.5 Evaluation Protocol

Claude (claude-sonnet-4-20250514) served as an automated judge, comparing each model response against the corresponding Wikipedia source text. The evaluator was prompted to identify: (1) *factual errors*—claims contradicting source material, and (2) *hallucinations*—fabricated claims not present in source. This LLM-as-judge approach follows established methodology for scalable evaluation [8].

2.6 Hedging Detection

Responses were automatically scanned for hedging patterns indicating uncertainty: phrases such as "I'm not sure," "I don't know," "I cannot," "may or may not," and similar expressions. Hedging rate serves as a proxy for calibrated uncertainty—a well-calibrated model should hedge more on topics where it lacks knowledge.

3. Results

3.1 Parameter Scaling Effects

Table 1 summarizes aggregate performance across the Gemma 3 model family. Accuracy improved approximately 10 percentage points per 3× parameter increase, consistent with logarithmic scaling. However, total hallucinations remained stable at 237-247 across all model sizes.

Table 1: Aggregate Performance by Model Size (n=20 probes per model)

Model	Parameters	Accuracy	Factual Errors	Hallucinations	Hedging
gemma3:4b	4B	21.8%	211	247	0%
gemma3:12b	12B	31.8%	163	237	0%
gemma3:27b	27B	40.5%	189	245	0%

3.2 Subject Familiarity Effects

Performance varied dramatically by subject familiarity (Table 2). Common species achieved 70-75% accuracy at the 27B scale, while specialized subjects reached only 60%, and obscure organizational entities remained at 10-20% regardless of model size. List-type articles produced catastrophic results, with the USFWS endangered species list generating 47-50 hallucinated species names per response.

Table 2: Accuracy by Subject (Recognition Queries, gemma3:27b)

Subject	Accuracy	Hallucinations
American Crow	75%	7
American Goldfinch	75%	4
The Birds of America	70%	6
American Avocet	60%	6
National Bird-Feeding Society	20%	6
USFWS Endangered Species List	10%	50

3.3 Absence of Uncertainty Signaling

No model at any parameter scale exhibited hedging behavior (0% hedging rate across all 60 probes). Responses at 10% accuracy were delivered with identical confident tone as those at 75% accuracy. This complete absence of calibrated uncertainty represents a critical limitation for deployment in knowledge-critical applications.

4. Case Study: American Avocet

The American Avocet response from gemma3:4b illustrates the confabulation pattern. The model produced fluent, authoritative prose with fundamental errors:

> **Bill morphology:** Described as "vibrant, almost iridescent, orange-red" and "downward-curved." The American Avocet has a thin, black, *upward*-curved bill—the defining feature reflected in the genus name *Recurvirostra* ("curved backwards").

Plumage: Described as "predominantly gray-brown." Avocets are strikingly pied (black and white) with rusty-orange head and neck in breeding plumage.

Breeding range: Claimed to "breed in the Arctic regions of North America (Alaska, Canada, and Greenland)." American Avocets breed in the western United States interior—alkaline lakes, prairie potholes, Great Basin wetlands—not the Arctic.

The model demonstrated *genre competence*—producing structurally correct natural history descriptions with appropriate sections on appearance, behavior, and conservation—while lacking *factual grounding*. This pattern suggests training on the *form* of ornithological writing without sufficient exposure to species-specific content.

5. Discussion

5.1 Implications for Local Model Deployment

These findings suggest that small local models (under 30B parameters) cannot serve as reliable knowledge sources for factual queries without augmentation. Even the best-performing configuration (gemma3:27b) achieved only 40.5% accuracy with zero uncertainty signaling. For knowledge-intensive applications, retrieval-augmented generation (RAG) [9] is mandatory rather than optional.

However, these models retain value as *fluent writers* given verified context. The same model that fabricates avocet morphology could accurately summarize a provided Wikipedia article. The capability gap is in *parametric knowledge*, not language generation.

5.2 The Hallucination Invariance Problem

A striking finding is that hallucination counts remained approximately constant (~240) across parameter scales while accuracy improved. This suggests that additional parameters enable more accurate *recall* of training data without reducing the tendency to *fabricate* when recall fails. Larger models are not more cautious—they simply know more while remaining equally willing to invent what they don't know.

5.3 Methodological Contributions

The LLM Cartography approach offers a scalable framework for characterizing model knowledge boundaries. Key innovations include: using Wikipedia category hierarchies for stratified domain sampling, automated evaluation via LLM-as-judge against ground truth, and systematic detection of hedging patterns. The methodology extends readily to other domains and model families.

6. Limitations

This study has several limitations that constrain generalizability:

Sample size: The test set (n=10 subjects, 20 probes per model) is small. Confidence intervals on accuracy estimates are wide (~±15 percentage points at 95% confidence). A production-scale study would require 100+ subjects.

Single model family: We tested only Gemma 3. Cross-architecture comparisons (Qwen, Mistral, LLaMA) would strengthen claims about parameter scaling effects.

Single domain: Ornithology may not be representative. Technical domains (programming, mathematics) or high-frequency topics (popular culture) may show different patterns.

LLM-as-judge bias: The Claude evaluator may introduce systematic biases. Manual validation of a sample would provide calibration.

7. Conclusion

Small language models exhibit a characteristic failure mode: *confident confabulation*. Accuracy scales with parameters, but hallucination rates and uncertainty signaling do not improve. For applications requiring factual reliability, these models must be paired with retrieval systems that provide verified context. The LLM Cartography methodology offers a practical approach to characterizing these boundaries before deployment.

8. Future Work

Planned extensions include: cross-architecture comparison at matched parameter counts, expansion to additional domains (ecology, geology, history), investigation of prompt engineering effects on hedging behavior, and integration of semantic similarity scoring for automated evaluation without LLM-as-judge.

References

[1] Gemma Team (2025). "Gemma 3 Technical Report." Google DeepMind. arXiv:2503.19786.

[2] Bang, Y., et al. (2025). "HalluLens: LLM Hallucination Benchmark." arXiv:2504.17550.

[3] Li, J., et al. (2023). "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." Proceedings of EMNLP 2023.

[4] Rawte, V., Sheth, A., & Das, A. (2023). "A Survey of Hallucination in Large Foundation Models." arXiv:2309.05922.

[5] Lin, S., Hilton, J., & Evans, O. (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." Proceedings of ACL 2022.

[6] Ollama (2024). "Ollama: Run Large Language Models Locally." https://ollama.ai

[7] Gemma Team (2024). "Gemma: Open Models Based on Gemini Research and Technology." arXiv:2403.08295.

[8] Zheng, L., et al. (2023). "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." arXiv:2306.05685.

[9] Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems 33.

Appendix A: Technical Details

Hardware: MacBook Pro M4 Max, 128GB unified memory

Inference: Ollama 0.5.x (local)

Evaluation Model: Claude claude-sonnet-4-20250514 via Anthropic API

Database: MySQL 8.4

Source Data: Wikipedia API (English), accessed November 29, 2025

Code: llm_cartography.py (Python 3.12, ~500 lines)

Timeout: 180 seconds per query (required for gemma3:27b depth queries)

Document History

Version 1.0 (November 29, 2025): Initial release