Yamit Ehrlich - 313600751

GENEVAL (Genealogical Evaluation): Evaluating Multi-Hop Reasoning Limitations in Large Language Models

Abstract. I present GENEVAL, a benchmark of 501 questions across 7 reasoning tasks over the British Royal Family genealogy. Evaluating Gemini 2.5 Flash and Pro, I find that while both models achieve ~90% on simple queries, while multi-hop reasoning* degrades significantly-converging to ~40% at 6 hops regardless of model size.

*Multi-hop reasoning: traversing multiple relationship links to answer a query (e.g., "Who is the grandfather of X's cousin?").

1. Introduction. LLMs are increasingly used to explore historical data and are often treated as authoritative truth, and the British Royal Family, widely discussed online and extensively documented across centuries, provides a rich domain to test whether such perceived factual reliability holds under reasoning.
I found that LLMs achieve 90%+ accuracy on simple genealogical queries, demonstrating they possess the underlying knowledge. However, performance degrades sharply when traversing multiple relationship hops - the limitation is reasoning, not knowledge. No existing benchmark tests this specifically, so I created **GENEVAL**: 501 questions over 314 British Royal Family members from Wikidata [3].

Recent work supports my findings: Yang et al. [1] found multi-hop evidence is "substantial for the first hop but only moderate for subsequent hops," and Wang et al. [2] showed that scaling up model size does not improve multi-step reasoning.

2. GENEVAL Dataset:

1. **Data Collection**: I queried Wikidata [3] using SPARQL to extract 314 British Royal Family members into a CSV with structured metadata: birth/death dates, parents, spouses, reign periods, and royal house.
2. **Question Generation**: The raw data is preprocessed to compute derived relationships (e.g., siblings from shared parents, lifespans from dates). Task generators then programmatically create questions by sampling people and relationships, with ground-truth answers computed directly from the structured data. Each question also carries metadata (royal house, time period, gender, etc.) derived from the people it references, enabling analysis across multiple dimensions.
3. **GENEVAL Benchmark**: 501 questions, 7 task types, equal difficulty distribution (33/33/33).

Task Type	Count	Difficulty Criteria	Example
Multi-Hop Reasoning	99 (20%)	Easy: 1-2 hops; Medium: 3-4 hops; Hard: 5-6 hops	Who is the mother of Henry VIII's grandfather?
Temporal Reasoning	72 (14%)	Easy: 2-person; Medium: 3-person; Hard: 4-5 person ranking	Who was born first: Elizabeth I or Mary I?
Negative Reasoning	72 (14%)	Easy: 2-3 options; Medium: 4-5 options; Hard: nested negations	Which of these was NOT a child of George III?
Sibling Inference	72 (14%)	Easy: direct sibling; Medium: half-sibling; Hard: sibling count	Are Edward VI and Elizabeth I full or half siblings?
Constraint Satisfaction	63 (13%)	Easy: 2 constraints; Medium: 3; Hard: 4+ constraints	Name a king who ruled before 1400 and lived past 60
Adversarial Ambiguity	63 (13%)	Easy: unique name; Medium: 2 same-name; Hard: 3+ same-name	Which Edward died first: the father or the son?
Comparative Lifespan	60 (12%)	Easy: simple comparison; Medium: reign overlap; Hard: multi-person	Who lived longer, Henry VII or Henry VIII?

3. Methodology.

- **Models:** Gemini 2.5 Flash and Gemini 2.5 Pro

- **Evaluation:** Accuracy using LLM-as-judge [4] for answer equivalence
- **Secondary metric:** Accuracy by Hop Count (compositional degradation curve)

4. Results.

E/M/H = Easy/Medium/Hard accuracy breakdown. Our analysis tool also segments by royal house, time period, century, people involved, monarch status, gender composition, relationship type, dynasty, and hop count.

Task Type	Flash	Flash E/M/H	Pro	Pro E/M/H
Adversarial Ambiguity	100%	100/100/100	98%	100/100/93
Temporal Reasoning	97%	100/100/90	99%	100/100/95
Sibling Inference	97%	100/92/100	90%	100/81/94
Comparative Lifespan	95%	100/92/100	**78%**	100/78/58
Negative Reasoning	89%	100/100/**66**	82%	96/95/**54**
Constraint Satisfaction	**81%**	76/71/95	**73%**	66/76/**76**
Multi-Hop Reasoning	**71%**	**90/81/39**	**66%**	**81/63/51**
Overall	**89%**	**94/92/81**	**83%**	**89/85/75**

Key findings: 1. **Hard Multi-Hop is the weakest category** for both models (39% Flash, 51% Pro), confirming multi-hop reasoning as a fundamental limitation 2. **Hard Negative Reasoning also reveals weakness** (66% Flash, 54% Pro), suggesting nested negations are challenging 3. **Both models achieve near-perfect accuracy** on Adversarial Ambiguity and Temporal Reasoning (97-100%) 4. **Flash outperforms Pro by 6% overall** (89% vs 83%)

Multi-Hop Degradation Analysis.

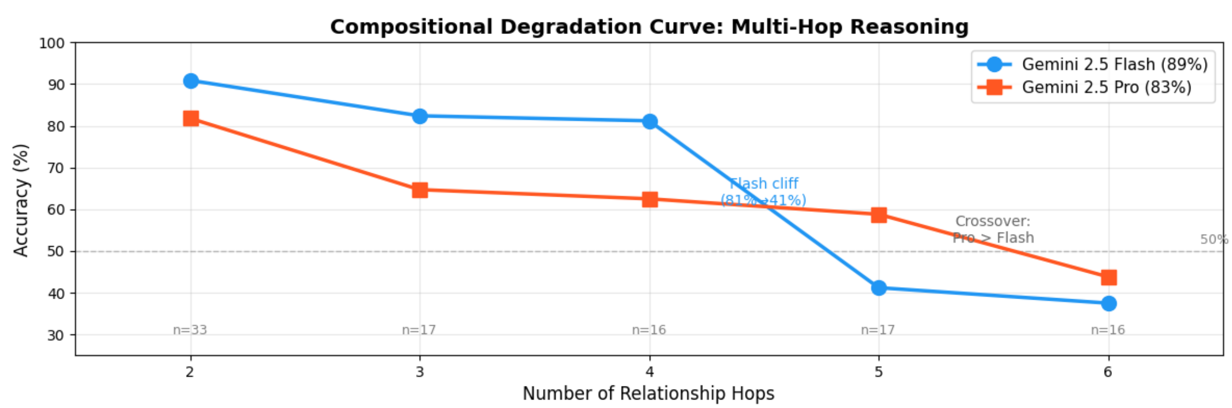

Figure 1: Compositional degradation curves showing Flash's sharp cliff at 5 hops and Pro's gradual decline starting at 3 hops.

Key Finding: Both models degrade as hop count increases, converging to ~40% accuracy at 6 hops - a fundamental ceiling for multi-hop genealogical reasoning. Flash maintains >80% through 4 hops before a sharp drop, while Pro degrades more gradually starting at 3 hops.6.

5. Contributions.

1. **Novel benchmark:** GENEVAL, the first genealogy-specific multi-hop reasoning dataset (501 questions, created from scratch) 2. **Fundamental limitation identified:** Both models converge to ~40% accuracy at 6 hops, revealing a ceiling in multi-hop genealogical reasoning 3. **Degradation patterns:** Performance degrades as hop count increases, with both models dropping below 50% at 5+ hops 4. **Practical implication:** For reliable genealogical queries, limit relationship chains to 2-3 hops.

References

[1] Yang, S., et al. (2024). *Do Large Language Models Latently Perform Multi-Hop Reasoning?* ACL 2024.

[2] Wang, P., et al. (2024). *Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability.* arXiv:2407.15720.

[3] Wikidata. https://www.wikidata.org/

[4] Zheng, L., et al. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.* NeurIPS 2023.