

Kinship Data Benchmark for Multi-hop Reasoning

Tianda Sun and Dimitar Kazakov

University of York

Department of Computer Science

Heslington, York

YO10 5DD

Abstract

Large language models (LLMs) are increasingly evaluated on their ability to perform multi-hop reasoning, i.e., to combine multiple pieces of information into a coherent inference. We introduce KinshipQA, a benchmark designed to probe this capability through reasoning over kinship relations. The central contribution of our work is a generative pipeline that produces, on demand, large-scale, realistic, and culture-specific genealogical data: collections of interconnected family trees that satisfy explicit marriage constraints associated with different kinship systems. This allows task difficulty, cultural assumptions, and relational depth to be systematically controlled and varied. From these genealogies, we derive textual inference tasks that require reasoning over implicit relational chains. We evaluate the resulting benchmark using six state-of-the-art LLMs, spanning both open-source and closed-source models, under a uniform zero-shot protocol with deterministic decoding. Performance is measured using exact-match and set-based metrics. Our results demonstrate that KinshipQA yields a wide spread of outcomes and exposes systematic differences in multi-hop reasoning across models and cultural settings.

1 Introduction

Large language models (LLMs) are increasingly evaluated not only on their ability to recall isolated facts, but also on their capacity to combine multiple pieces of information into a coherent answer. A particularly challenging instance of this broader goal is multi-hop reasoning: the ability to infer a correct conclusion by chaining together several intermediate relations, including such that are not explicitly stated. In this paper, we propose a novel way to test this ability by making use of data about kinship relations.

Our approach is based on producing computer-generated realistic datasets representing a collection of interconnected family trees constructed to

comply with a set of constraints regulating marriage in a given society. These constraints, often perceived as cultural or legal ‘taboos’, determine which kin relations are possible, forbidden, or socially marked. By controlling these constraints, we can generate large, internally consistent genealogical datasets tailored to specific cultural settings.

Family trees are a particularly suitable domain for evaluating multi-hop reasoning. First, kinship is a natural and ubiquitous topic of human communication. Second, genealogical structures offer a controlled way to vary task complexity: simple relations can be defined over one or two edges in a tree, while more complex relations require chaining over many intermediate nodes. This makes kinship an ideal testbed for probing how well an LLM can integrate multiple facts into a single inference.

Crucially, many cultures and their languages lexicalise kin relations that span several nodes in a family tree. English second cousins, for instance, are defined as individuals who share great-grandparents but not grandparents. Bulgarian *девер* (*dever*) denotes one’s husband’s brother. Such terms allow speakers to refer succinctly to a specific relative without spelling out all intermediary relations. In principle, combinations of such terms could further shorten descriptions of complex relationships; in practice, however, they may also lead to confusion or over-complication. Continuing with the Bulgarian example, the relationship between a man’s *шурей* (*shooray*) and his *балдъза* (*balduza*) reduces to a simple brother–sister relation, since both are siblings of his wife.

There is cross-linguistic variation in kinship systems where concepts lexicalised in one language have no direct equivalent in another: Bulgarian, for example, lacks a single word corresponding to English *sibling*, instead using the phrase ‘brothers and sisters’. Conversely, a single term may cover distinct concepts across cultures. In some societies, the closest translation of a given kin term may be

used to refer not only to a biological mother, but also to all maternal aunts. Despite these differences, any biological kin relation can in principle be described using elementary notions such as biological father and mother; what varies is the ease, compactness, and conventionality of such descriptions.

These observations suggest that LLM performance on kinship inference tasks may vary depending on the language and culture involved, and on the model’s implicit knowledge of the relevant kinship vocabulary and concepts. Nevertheless, the task itself remains well defined: inferring the relationship between two individuals from a set of stated facts. As such, it provides a robust and interpretable way to assess multi-hop reasoning ability.

It is also reasonable to expect that both human and artificial reasoners will perform better on kinship distinctions that are salient in their cultural training data. For example, in societies with strong Orthodox Christian traditions, distinctions between second cousins (prohibited as marriage partners) and third cousins (permitted by the Church) are more likely to be explicitly discussed than in Protestant contexts, where even first-cousin marriage may not be religiously forbidden. Similarly, in populations with weakened extended family ties—due, for instance, to high labour mobility, there may simply be fewer occasions to encounter or talk about relations such as one’s wife’s sister’s husband (*bacanak* in Turkish); in societies where one-child families are common, the offspring of such single children have no aunts, uncles or cousins. In short, we expect systematic variation in performance as a function of culture- and population-specific exposure to kinship structures.

For an evaluation framework to be informative, it must yield a sufficient spread of outcomes: it is of limited value if most models either pass all tests or fail entirely. As a core component of any culture, kinship terms belong to the most stable layers of vocabulary and have long been used in historical and comparative linguistics—most notably in the Swadesh lists—to identify genealogical relationships between languages and to infer common origin (Swadesh, 1952). If changes in kinship systems and their associated vocabularies lead to measurable differences in LLM performance, this should be taken as a broader warning that model behaviour may not be consistent across cultures and domains.

In this paper, we introduce a pipeline for generating culture-specific, realistic sets of family trees spanning an arbitrary number of individuals

and generations. Such data have many potential applications—for instance, providing genealogical backgrounds for non-player characters in computer games—and we hope our tool will find wide adoption beyond the present study. Here, however, we focus specifically on its use as a testbed for evaluating the ability of LLMs to perform multi-hop reasoning over textual descriptions of culturally grounded relational data, which we demonstrate on a number of popular LLMs.

2 Related Work

2.1 Multi-Hop Reasoning Benchmarks

Multi-hop reasoning—chaining multiple inference steps to answer complex questions—remains a fundamental challenge for LLMs (Qiao et al., 2023). Benchmarks like HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020) evaluate this capability, but face critical limitations: training data contamination and lack of reasoning chain verification. MRKE addresses contamination through knowledge editing, revealing that GPT-4’s accuracy drops from 69.3% to 53.2% on edited questions, with only 36.3% of responses following correct reasoning chains (Zhou et al., 2024). CompoST demonstrates that LLMs struggle with compositional SPARQL-mapped questions even when they understand atomic components, with F1 scores degrading from 0.45 to 0.09 as structural complexity increases (Schmidt et al., 2024; Li et al., 2024) identify three prevalent error types: hasty answers, incomplete reasoning chains, and logical inconsistencies.

2.2 Kinship Reasoning Benchmarks

There is a history of research using kin data as a machine learning testbed. CLUTRR (Sinha et al., 2019) is a recent attempt to use kinship relations for the evaluation of compositional reasoning in natural language understanding systems. Given semi-synthetic stories describing family relationships, models must infer unstated relations (e.g., given “Alice is Bob’s mother” and “Bob is Carol’s father,” infer Alice is Carol’s grandmother). CLUTRR tests *rule induction*: whether models can learn logical rules from examples and generalise to longer reasoning chains.

2.3 Cultural Knowledge in NLP

Cross-cultural NLP has gained attention as researchers recognise that language technologies

must account for cultural variation. Here kinship is a fundamental, but under-researched category where diverse training and evaluation datasets are still lacking (Liu et al., 2024). AlKhamissi et al. (2025) has recently criticised existing cultural benchmarks for reducing culture to static facts or homogeneous values, in contradiction with anthropological accounts that emphasize culture as dynamic and enacted in practice.

It should be noted that unlike biological kinship (universal facts of reproduction), the concepts of cultural kinship can vary, yet it operates as a well-defined, “constructed, computational system” with formal rules and anthropologically-documented variation (Read, 2012). This makes kinship systems an excellent testbed for systematic evaluation of multi-relational reasoning.

3 Methodology

3.1 Task Formulation

We frame kinship reasoning as a reading comprehension task. Given a natural language context \mathcal{C} describing family relationships and a question \mathcal{Q} about a specific individual, the model must produce an answer \mathcal{A} that is either a single entity or a set of entities. Formally: $f(\mathcal{C}, \mathcal{Q}) \rightarrow \mathcal{A}$.

The key challenge lies in distinguishing between two reasoning modes: biological reasoning, which follows universal genealogical facts (e.g., “mother’s sister” → “aunt”), and cultural reasoning, which requires applying culture-specific classification rules that may override biological defaults (e.g., in Hawaiian kinship, “mother’s sister” → “mother”). Our benchmark tests both capabilities across controlled complexity levels measured by the number of relationship hops (n-hops) required to derive the answer.

3.2 Kinship Systems as Test Domains

We select kinship systems as our test domain for three reasons. First, kinship provides naturally controlled complexity: relationships form tree structures where reasoning depth (n-hops) is precisely measurable. Second, kinship systems exhibit well-documented cultural variation with formal, rule-based classifications—ideal for systematic evaluation. Third, the domain enables procedural generation of novel instances, eliminating training data contamination.

Our benchmark encompasses seven anthropologically documented kinship systems representing

Table 1: Overview of seven kinship classification systems in KinshipQA. F=Father, FB=Father’s Brother, FZS=Father’s Sister’s Son, MB=Mother’s Brother, MBS=MB’s Son.

System	Type	Key Rule
Eskimo	Descriptive	$F \neq FB$
Sudanese	Descriptive	All terms unique
Hawaiian	Generational	$F = FB$
Iroquois	Bifurcate	Parallel ≠ Cross
Dravidian	Bifurcate	Cross = Spouse
Crow	Mat. Skewing	$FZS = F$
Omaha	Pat. Skewing	$MBS = MB$

the major classification patterns identified by Morgan (Morgan, 1871): Eskimo, Hawaiian, Iroquois, Dravidian, Crow, Omaha, and Sudanese. Table 1 summarises these systems. Detailed anthropological descriptions are provided in Appendix A.

3.3 Benchmark Generator Pipeline

Figure 1 illustrates our five-stage, fully automated pipeline for generating the KinshipQA benchmark. We briefly describe each stage below.

Population Simulation We generate multi-generational family trees using a given initial population size, and providing the time span (in years) of the generated data, and the type of kinship system as parameters. Each kinship system enforces different constraints, e.g. Eskimo prohibits only sibling marriage, Iroquois allows cross-cousin but not parallel-cousin marriage, Dravidian prefers cross-cousin marriage, and Crow/Omaha enforce clan-based prohibitions following matrilineal and patrilineal descent, respectively.

RDF/OWL Encoding Family structures are formalised as RDF ontologies using a dual-namespace architecture. The family: namespace captures biological relationships (hasFather, hasMother, hasSibling), while the kin: namespace captures cultural classifications (hasClassificatoryParent, hasCross-Cousin). This separation enables precise evaluation of whether models can distinguish biological facts from cultural categories.

Question Generation We employ path-based generation, where biological relationship paths (e.g., mother → sister → child) are mapped to kinship terms that vary by cultural system. This approach ensures questions test genuine multi-hop reasoning rather than memorised relationship labels. Questions span four categories with controlled n-hop complexity (1-4 hops).

Ground Truth Generation For each question, we

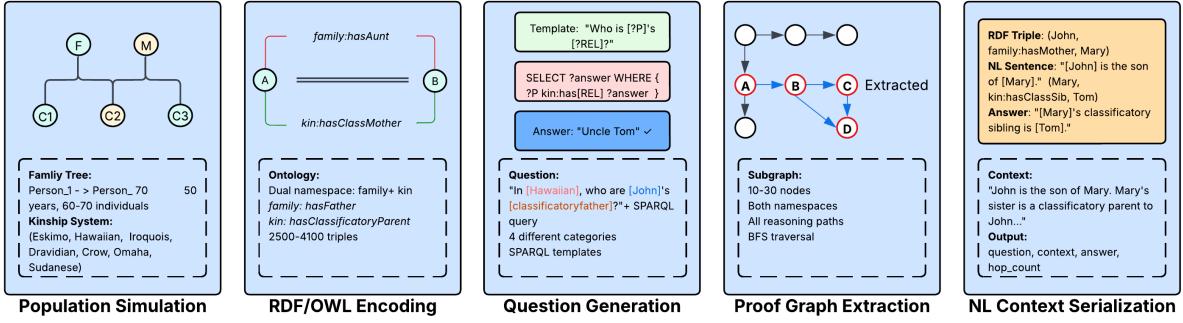


Figure 1: KinshipQA generation pipeline. (1) **Population Simulation** generates multi-generational family trees (60–70 individuals) respecting culture-specific marriage rules. (2) **RDF/OWL Encoding** formalizes structures using dual namespaces: `family:` for biological relationships and `kin:` for cultural classifications. (3) **Question Generation** creates questions across four categories with controlled n-hop complexity using path-based templates. (4) **Proof Graph Extraction** derives minimal reasoning subgraphs required to answer each question. (5) **NL Context Serialisation** converts the relevant RDF subgraph into natural language context for LLM evaluation.

Table 2: Question categories in KinshipQA.

Cat	Type	Hops	Example
1	Fact Retrieval	1	Who is John's father?
2	Multi-hop Biol.	2-4	Who is John's paternal uncle?
3	Order/Counting	1-2	How many siblings does Mary have?
4	Cultural	2-4	In the Hawaiian system, X is John's 'classificatory father'. What is X's biological relation?

execute SPARQL queries against the RDF ontology to obtain ground-truth answers with 100% accuracy. This symbolic approach eliminates annotation ambiguity and enables automatic verification of model reasoning chains.

Natural Language Serialisation The minimal subgraph required to answer each question is serialised into natural language context. We use template-based generation for consistency, converting RDF triples to sentences (e.g., “Alice is the mother of Bob”).

3.4 Question Categories

Table 2 presents our four question categories, designed to test progressively complex reasoning capabilities.

Category 1 (Fact Retrieval) tests single-hop lookup of directly stated relationships, serving as a baseline to verify that models can extract explicit information from context.

Category 2 (Multi-Hop Biological) requires chaining 2-4 biological relationships to derive answers. For example, finding a "paternal grandfather" re-

quires traversing `father → father`. These questions test compositional reasoning over universal genealogical structures.

Category 3 (Counting/Filtering) introduces logical operations: counting entities ("How many siblings does X have?"), filtering by attributes ("List X's male cousins"), and comparisons ("Who is older, X or Y?"). These questions test whether models can handle queries of Category 2 augmented with order and counting during graph traversal.

Category 4 (Cultural Disambiguation) represents our core contribution. These questions require models to apply culture-specific classification rules that diverge from biological relationships. For example, "According to Hawaiian kinship, who are X's parents?" requires recognising that mother's sisters are classified as "mothers." This category directly tests whether LLMs can override default biological assumptions and reason about culturally constructed categories.

3.5 Comparing LLMs with KinshipQA

We have used our pipeline to generate one dataset for each type of kinship system. All datasets span 50 years, starting from an initial population of four individuals. The currently hard-wired constraints about number of offspring results in an overall population of 60-70 individuals generated in that time span. The data is then used to evaluate and compare a number of LLMs popular at present.

Table 3 presents the statistics of the dataset. We generate balanced distributions across categories and n-hop complexity levels, with a minimum of 80 questions per cell to ensure statistical reliability.

Table 3: KinshipQA dataset statistics. “Override” indicates questions requiring cultural reclassification knowledge, e.g. mother’s sister → mother.

	Questions	%	Override
<i>By System Type</i>			
Eskimo, Sudanese	974	31.1	0%
Remaining five	2,160	68.9	100%
<i>By Category</i>			
Cat 1: Fact	350	11.2	—
Cat 2: Multi-hop	1,050	33.5	—
Cat 3: Order & count	700	22.3	—
Cat 4: Cultural	1,034	33.0	63.8%
<i>By Complexity</i>			
1-hop	985	31.4	—
2-hop	877	28.0	—
3-hop	812	25.9	—
4-hop	460	14.7	—
Total	3,134	100	21.1%

4 Results and Evaluation

4.1 Experimental Setup

Models. We use KinshipQA to evaluate six large language models spanning both open-source and closed-source state-of-the-art systems. Open-source models include **Qwen3-32B** (Yang et al., 2025), **Gemma3-27B** (Team et al., 2025), and **DeepSeek-R1-32B** (DeepSeek-AI et al., 2025), representing diverse model families from Alibaba, Google, and DeepSeek respectively. Closed-source models include **GPT-4o-mini** (OpenAI et al., 2024), **Claude-3.5-Haiku** (Anthropic, 2024), and **Gemini-2.5-Flash** (Comanici et al., 2025).

Evaluation Protocol. All models receive identical zero-shot prompts containing the natural language context and question. We use greedy decoding (temperature=0) to ensure reproducibility. For questions with multiple valid answers (43.8% of the dataset), we compute set-based metrics. We report **Exact Match (EM)** as our primary metric with a detailed analysis of the difference by category and culture.

Evaluation Infrastructure. We implement a unified evaluation pipeline supporting multiple API backends. For open-source models, we use Ollama for local inference.

Table 4 presents our main findings across all evaluated models.

4.2 Performance by Reasoning Complexity

Figure 2 presents performance as a function of reasoning complexity measured by n-hops. Contrary

to expectations, we observe that 4-hop questions achieve higher accuracy than 3-hop questions for systems with override. This counterintuitive pattern arises because our 4-hop questions primarily test biological relationship chains, while 3-hop questions more frequently involve cultural classification steps.

Table 5 quantifies this pattern. The non-override/override gap is minimal at 1-hop (6.2%) and 4-hop (0.4%), but peaks at 3-hop (27.4%). This demonstrates that the difficulty of override-type kinship reasoning is not simply a function of chain length—it specifically arises when cultural classification rules must be applied during multi-hop inference.

4.3 Performance by Kinship System

Table 6 presents detailed performance breakdowns by kinship system, revealing systematic patterns in model capabilities.

Skewing Systems are Most Challenging. The Crow and Omaha systems, which employ generational skewing rules (where individuals on one parent’s side are reclassified across generations), prove most difficult for LLMs. Omaha (patrilineal skewing) yields the lowest Category 4 accuracy at **44.1% ± 4.6%**—barely above random chance for binary classification. Crow (matrilineal skewing) follows at 57.8%. These systems require models to apply non-intuitive rules where relatives are classified into different generations than their biological position (e.g., father’s sister’s son is classified as “father” in Crow systems).

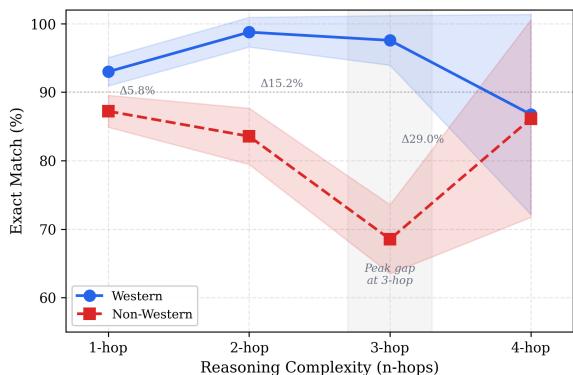


Figure 2: Performance degradation by reasoning complexity (n-hops). Non-override systems maintain high accuracy across all hop counts, while the rest show degradation at 3-hop complexity.

Table 4: Main results on KinshipQA (Exact Match %). ‘Other 5’ includes Hawaiian, Iroquois, Dravidian, Crow, and Omaha. Δ Gap = Esk&Sud – Other 5. Cat.1,2,3 are computed on all 7 systems. Cat.4 scores are computed on ‘other 5’ systems only, where cultural override rules apply.

Model	By Culture			By Category			
	Esk&Sud	Other 5	Δ Gap	Cat.1	Cat.2	Cat.3	Cat.4
<i>Closed-Source Models</i>							
GPT-4o-mini	92.8	81.2	+11.1	100.0	92.6	92.0	62.6
Claude-3.5-Haiku	87.1	75.1	+11.9	100.0	88.7	86.8	47.3
Gemini-2.5-Flash	98.2	84.3	+13.9	100	99.5	86.1	61.4
<i>Open-Source Models</i>							
Qwen3-32B	96.7	83.7	+13.0	99.7	98.1	86.1	61.3
Gemma3-27B	96.9	84.5	+12.4	100.0	98.4	84.6	64.5
DeepSeek-R1-32B	94.3	77.3	+17.0	91.4	93.9	82.9	51.0
Average	96.0	81.8	+14.1	97.0	96.8	84.5	58.9

Table 5: Performance by n-hops complexity (% of exact matches averaged across models). The non-override/override gap peaks at 3-hop, where cultural reasoning is most prevalent.

System Type	1-hop	2-hop	3-hop	4-hop
non-override	93.0	98.8	97.6	86.7
override	87.2	83.6	68.5	86.1
Δ Gap	5.8	15.2	29.0	0.6

Generational Merging is Relatively Easier. Hawaiian kinship, which employs generational merging (all relatives of the same generation receive the same term), achieves 65.9% on Category 4—still significantly below Western systems but higher than skewing systems. This suggests that “lumping” rules may be easier to learn than “skewing” rules, possibly because generational merging follows a more regular pattern.

Cross-Cousin Systems Show Intermediate Difficulty. Iroquois (59.8%) and Dravidian (70.4%) systems, both based on parallel/cross distinctions between the mother’s and the father’s side, show intermediate performance. Dravidian’s higher accu-

Table 6: Performance by kinship system (Exact match [%], mean \pm std across 3 open-source models). Omaha shows the lowest Cat.4 performance at 44.1%.

System	Type	Overall	Cat.4
Eskimo	Descriptive	94.3 ± 3.9	93.8 ± 7.3
Sudanese	Descriptive	94.2 ± 3.6	93.8 ± 6.2
Hawaiian	Generational	82.8 ± 3.2	63.9 ± 4.0
Iroquois	Bifurcate	80.2 ± 4.2	57.6 ± 8.9
Dravidian	Bifurcate	84.0 ± 4.8	71.6 ± 9.7
Crow	Mat. Skewing	80.6 ± 3.5	55.2 ± 5.2
Omaha	Pat. Skewing	77.4 ± 2.9	44.1 ± 4.6

racy may reflect greater representation in training data due to its association with South Asian languages spoken by over a billion people.

4.4 Cultural Override Effect

Table 7 isolates the effect of cultural override by comparing Category 4 questions where cultural classification differs from biological relationship versus those where they align. The results reveal a striking pattern: when cultural rules require overriding biological intuitions, accuracy drops by **23.5 percentage points** on average (from 90.8% to 67.3%).

This gap is most pronounced in skewing systems: Omaha shows a 39.3-point drop, and Crow shows a 29.5-point drop. In contrast, Hawaiian (16.9 points) and Dravidian (11.7 points) show smaller gaps, suggesting that generational merging and cross-cousin rules are more learnable than generational skewing.

5 Discussion

Our results reveal two distinct factors that independently affect LLM performance on kinship reason-

Table 7: Cultural override effect on Category 4 questions (Exact Match %). “w/ Override” indicates questions where cultural classification differs from biological relationship; “w/o Override” indicates alignment.

System	w/ Override	w/o Override	Gap
Hawaiian	72.4 ± 4.0	90.2 ± 3.0	17.7
Iroquois	68.5 ± 7.2	88.7 ± 2.9	20.2
Dravidian	78.5 ± 8.0	88.0 ± 3.4	9.6
Crow	60.9 ± 5.0	91.4 ± 3.0	30.5
Omaha	51.9 ± 4.4	91.7 ± 2.4	39.8
Average	66.4 ± 9.2	90.0 ± 1.4	23.6

ing: **reasoning complexity** (measured by n-hops) and **cultural variation** (differences across kinship systems). We analyse each factor below, then examine their interaction.

5.1 Factor 1: Reasoning Complexity (N-hops)

Performance degrades predictably with chain length across all kinship systems. Table 5 shows that accuracy decreases from 1-hop to 3-hop questions in all systems. The recovery at 4-hops reflects our dataset design: 4-hop questions primarily test biological relationship chains, while 3-hop questions more frequently involve cultural classification steps. This confirms that chain tracking is a genuine challenge for LLMs, independent of cultural knowledge.

Importantly, multi-hop errors occur even in Eskimo and Sudanese systems, where cultural familiarity should be maximal. The “off-by-one generation” error pattern—where models stop one hop short of the correct answer—appears across all systems. This demonstrates that reasoning complexity imposes a fundamental limitation independent of cultural factors.

5.2 Factor 2: Cultural Variation

Beyond reasoning complexity, kinship system type independently affects performance. Holding n-hops constant, we observe systematic accuracy differences across cultural systems.

Performance Varies by System Type. Table 6 reveals a clear hierarchy: the best results are achieved for the descriptive systems (Eskimo, Sudanese), followed by bifurcate systems (Iroquois, Dravidian), generational systems (Hawaiian), and, finally, the skewing systems (Crow, Omaha).

This ordering reflects both training data exposure and rule complexity. Eskimo/Sudanese terms dominate English-language corpora (“aunt,” “uncle,” “cousin”), while skewing rules require non-local reasoning that depends on clan membership across generations.

The Cultural Override Effect. The most striking evidence for cultural factors appears in Category 4 questions where cultural classification differs from biological relationship. Table 7 shows that when cultural rules require overriding biological intuitions, accuracy drops by **23.6 percentage points** on average (from 90.0% to 66.4%). This gap is the greatest for skewing systems, followed by bifurcate and generational systems.

Table 8: Error Types by Primary Factor

Error Type	N	%
<i>Reasoning Complexity (53.9%)</i>		
Incomplete chain	114	29.8
Counting error	92	24.1
<i>Cultural Variation (30.4%)</i>		
Cultural default	81	21.2
Over-inclusion	35	9.2
<i>Other (15.7%)</i>		
Hallucination	48	12.6
Other	12	3.1
Total	382	100

The magnitude of this effect demonstrates that cultural variation is not merely a function of training data frequency—it reflects genuine difficulty in applying rules that contradict encoded defaults.

5.3 Interaction: Cultural Rules at Multi-Hop Complexity

The two factors interact: cultural classification rules typically apply at 2–3 hop complexity, precisely where multi-hop reasoning becomes challenging. Figure 2 shows the Eskimo and Sudanese verse the rest systems gap peaks at 3-hops (29.0%), the complexity level where cultural rules most frequently apply.

This interaction explains a key finding: models struggle not because they lack cultural knowledge, but because they must *apply* cultural rules while simultaneously tracking multi-hop relationship chains. The cognitive load of multi-hop reasoning amplifies the difficulty of cultural rule application.

5.4 Error Analysis

To understand *how* these two factors manifest in model behaviour, we manually analysed 200 incorrect responses from GPT-4o-mini and Gemma3-27B using chain-of-thought prompting. Table 8 presents error types organised by their primary cause.

Reasoning Complexity Errors (53.9%). The majority of errors reflect chain-tracking failures:

- **Incomplete chain:** Models stop one hop short of the correct answer (“off-by-one generation”);
- **Counting error:** Models fail to enumerate all members of a relationship set.

These errors occur across all kinship systems, confirming that multi-hop reasoning imposes fundamental limitations.

Cultural Variation Errors (30.4%). A substantial minority of errors specifically involve cultural knowledge:

- **Cultural default:** Models apply Eskimo terminology to other systems (e.g., calling a “classificatory mother” an “aunt”)
- **Over-inclusion:** Models include biological relatives alongside classificatory relatives, failing to recognise that cultural terms *redefine* rather than merely *expand* categories

Crucially, cultural default errors occur *exclusively* in systems other than Eskimo or Sudanese (Table 6), suggesting that LLMs likely encode Eskimo (*aka* ‘Western’) kinship structures as defaults. A case study with three examples of different errors can be found in Appendix B.

5.5 Implications

Our two-factor analysis has implications for both LLM evaluation and deployment:

1. **Multi-hop reasoning limitations are fundamental:** Performance degrades with chain length regardless of cultural familiarity. This confirms that compositional reasoning remains a core challenge for LLMs.
2. **Cultural variation is not reducible to training frequency:** The cultural override effect (23.6%) persists even when models have declarative knowledge of rules, suggesting deeper architectural or training limitations.
3. **The factors interact multiplicatively:** Cultural classification rules that apply at multi-hop complexity are particularly challenging, as models must simultaneously track relationship chains and override default assumptions.
4. **Standard benchmarks may overestimate capabilities:** Evaluations focused on Western-centric knowledge structures miss systematic performance variation that emerges in culturally diverse contexts.

These findings suggest that addressing cultural competence in LLMs requires more than expanded training data—it may require architectural innovations that enable genuine rule-following over

deeply-encoded defaults, particularly in multi-hop reasoning contexts.

6 Conclusion

We presented KinshipQA, a contamination-proof benchmark for evaluating multi-hop reasoning across culturally diverse kinship systems. Our experiments reveal two independent factors limiting LLM performance: **reasoning complexity** and **cultural variation**. Multi-hop reasoning degrades with chain length regardless of cultural familiarity, while cultural variation independently reduces accuracy from 96.0% on Eskimo/Sudanese systems to 81.8% on the rest of systems—a 14.1% gap that persists across model families. The cultural override effect (23.6% drop when cultural rules contradict biological intuitions) confirms this gap reflects genuine difficulty applying non-default rules, not merely training frequency.

These factors interact: the gap between the descriptive systems and the rest peaks at 3-hop complexity (29.0%), where cultural classification rules most frequently apply. Error analysis reveals a declarative-procedural gap—models cite cultural rules correctly but fail to apply them—suggesting that cultural competence requires architectural innovations beyond expanded training data. Future work includes extending KinshipQA to additional systems and investigating whether similar two-factor patterns emerge in other culturally variable domains.

Limitations

Our benchmark covers seven kinship systems from Morgan’s typology, representing major classification patterns but not the full diversity of documented systems worldwide. Our implementations follow idealised anthropological models that may not capture regional variations or hybrid systems.

KinshipQA is English-only, which may favor models trained predominantly on English text and does not test reasoning with indigenous kinship terminology. Our evaluation uses zero-shot prompting exclusively; few-shot or fine-tuning experiments may reveal different capability profiles. Due to budget constraints, we tested mid-tier closed-source models rather than flagship variants, though consistent bias patterns across model families suggest findings would generalise. Finally, we lack human baseline data, which would help contextualise task difficulty.

Acknowledgments

References

- Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. [Hire Your Anthropologist! Rethinking Culture Benchmarks Through an Anthropological Lens](#).
- Anthropic. 2024. [The Claude 3 Model Family: Opus, Sonnet, Haiku](#).
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marrs, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities](#). *arXiv preprint*. ArXiv:2507.06261 [cs].
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhushu Li, Ziyi Gao, and 181 others. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *arXiv preprint*. ArXiv:2501.12948 [cs].
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps](#). *Preprint*, arXiv:2011.01060.
- Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. [Understanding and Patching Compositional Reasoning in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9668–9688, Bangkok, Thailand. Association for Computational Linguistics.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. [Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art](#). *arXiv preprint* arXiv:2406.03930.
- Lewis Henry Morgan. 1871. Systems of Consanguinity and Affinity of the Human Family. *Smithsonian Contributions to Knowledge*, 17.
- OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. [GPT-4o System Card](#). *arXiv preprint*. ArXiv:2410.21276 [cs].
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. [Reasoning with Language Model Prompting: A Survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.
- Dwight Read. 2012. Cultural Kinship as a Computational System: From Bottom-up to Top-down Forms of Social Organization. *Computational and Mathematical Organization Theory*, 18(2):135–174.
- David Maria Schmidt, Raoul Schubert, and Philipp Cimiano. 2024. CompoST: A Benchmark for Analyzing the Ability of LLMs to Compositionally Interpret Questions in a QALD Setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. [CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.
- M. Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452–463.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#). *arXiv preprint*. ArXiv:2503.19786 [cs].
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Bin Yuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *arXiv preprint*. ArXiv:2505.09388 [cs].
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Yiming Zhou, Pei Zhang, Baosong Yang, and others. 2024. MRKE: The Multi-hop Reasoning Evaluation of LLMs by Knowledge Edition. *arXiv preprint* arXiv:2402.11924.

A Kinships System Description

Eskimo (Western/Lineal). The dominant system in English-speaking societies. Distinguishes lineal

relatives (parents, grandparents) from collateral relatives (aunts, uncles, cousins). Mother’s sister and father’s sister both = “aunt.”

Hawaiian (Generational). Merges all relatives of the same generation and sex. All female relatives of mother’s generation = “mother”; all male relatives of father’s generation = “father”; Cousins = siblings.

Iroquois (Bifurcate Merging). Distinguishes parallel relatives (same-sex parents’ sibling) from cross relatives (opposite-sex parents’ sibling). Father’s brother = “father”; mother’s brother = “uncle.” Parallel cousins = siblings; cross cousins distinguished.

Dravidian (Bifurcate with Cross-Cousin Marriage). Similar to Iroquois, but cross-cousins are prescribed marriage partners, classified as potential spouses rather than kin.

Crow (Matrilineal Skewing). Members of father’s matrilineage are “skewed” upward generationally. Father’s sister’s children are classified as “fathers/female fathers” regardless of actual generation.

Omaha (Patrilineal Skewing). Mirror image of Crow. Members of the mother’s patrilineage are skewed upward. Mother’s brother’s children are classified as “mothers/male mothers.”

Sudanese (Descriptive). Maximally descriptive system with unique terms for each biological relationship. No merging or classification rules—each kin type receives a distinct label.

B Qualitative Examples

Example 1: Chain Tracking Failure (Eskimo System)

Question: Who are Lisa Williams’s paternal uncle’s grandchildren?

Context: Lisa’s father is Robert Williams. Robert’s brother is Larry Williams. Larry’s children are Patricia, Justin, and Mark. Mark’s children are Samantha, Jerry, and Jason. Justin’s children are Betty, Nicholas, and Ashley.

Ground Truth: 6 people (Samantha, Jerry, Jason, Betty, Nicholas, Ashley)

Prediction: Patricia, Justin, Mark (3 people)

Analysis: The model returns Larry’s *children* (3 hops) instead of *grandchildren* (4 hops). This error occurs in the Eskimo (Western) system, demonstrating that chain complexity—not cultural knowledge—is the limiting factor.

Example 2: Cultural Default Error (Crow System)

Question: List all of Samantha Williams’s classificatory mothers according to the Crow system.

Context: Samantha’s mother is Lisa Williams. Lisa’s sister is Michelle Williams.

Ground Truth: Michelle Williams

Prediction: Lisa Williams, Michelle Williams

Analysis: The model correctly identifies that mother’s sister (Michelle) is a classificatory mother. However, it *also* includes the biological mother (Lisa), failing to recognise that the question asks specifically for *classificatory* mothers—a category that excludes the biological mother. This over-inclusion pattern reflects cultural default thinking.

Example 3: Generational Skewing Failure (Omaha System)

Question: In the Omaha kinship system, what kinship term does John use for his mother’s brother’s son (MBS)?

Ground Truth: “mother’s brother” / “uncle” (due to patrilineal skewing, MBS = MB)

Prediction: “cousin”

Analysis: Omaha patrilineal skewing collapses the mother’s brother’s line: MBS is classified as MB (mother’s brother), not as a cousin. The model applies Western/Eskimo terminology where MBS would be a cousin, demonstrating failure to apply generational skewing rules.

C Prompt Templates

C.1 Zero-Shot Evaluation Prompt

Answer the following question based on the given context.
Be concise and provide only the answer without explanation.

Context: {context}

Question: {question}

Answer:

C.2 Chain-of-Thought Error Analysis Prompt

You are analyzing kinship relationships in a family. Read the context carefully and answer the question by showing your complete reasoning process.

Context: {context}

Question: {question}

Please think through this step-by-step:

STEP 1 - IDENTIFY KEY PERSON(S):

Who is the question asking about?

What relationship are we looking for?

STEP 2 - EXTRACT RELEVANT FACTS:

List the specific family relationships from the context that are relevant.

STEP 3 - TRACE THE REASONING CHAIN:

Work through the relationships step by step.

Show each connection.

STEP 4 - APPLY CULTURAL RULES (if mentioned):

If the context mentions any specific kinship system rules (like clan membership, classificatory relationships, or cultural terminology), apply them here.

STEP 5 - FINAL ANSWER:

Based on your reasoning, provide the answer.

D Cultural Override Mappings

Table 9 shows how biological relationship paths map to different kinship terms across cultural systems.

Table 9: Cultural override mappings for key relationship paths. Note that cl. = classificatory

Bio Path	Esk./Sud.	Hawaiian	Iroq./Drav.
F→Brother	pat. uncle	cl. father	cl. father
M→Sister	mat. aunt	cl. mother	cl. mother
F→Sister	pat. aunt	cl. mother	cross-aunt
M→Brother	mat. uncle	cl. father	cross-uncle
F→Bro→Child	cousin	cl. sibling	parallel cousin
M→Sis→Child	cousin	cl. sibling	parallel cousin
F→Sis→Child	cousin	cl. sibling	cross-cousin
M→Bro→Child	cousin	cl. sibling	cross-cousin