Kinship Data Benchmark for Multi-hop Reasoning

Anonymous ACL submission

Abstract

Multi-hop kinship reasoning is a natural testbed for LLM compositionality, but existing benchmarks (notably CLUTRR) cover only the descriptive Eskimo system. We introduce KinshipQA, a procedurally-generated benchmark covering seven anthropologically-documented kinship systems (Eskimo, Sudanese, Hawaiian, Iroquois, Dravidian, Crow, Omaha) and up to six reasoning hops, with a tunable simulator horizon that eliminates exact-instance pretraining overlap. Evaluating six LLMs, we find a 40.9% accuracy drop when reasoning shifts from biological multi-hop to culturally-marked classification on the five non-descriptive systems. The drop follows the anthropological hierarchy descriptive > bifurcate > generational > skewing, persists under chain-of-thought and few-shot prompting, and compounds with depth (10.6% at 5–6 hops). Humans given the override rule reach 89.0% versus 50.7% for LLMs, placing the failure in the model rather than the task. Two follow-up experiments suggest distinct contributors. A fictional-rule control swapping system labels and kin terms for invented strings raises accuracy by 6.1%, implicating familiar English surface forms. An in-context-rule probe prepending the override rule helps skewing systems (+17.1%) but *hurts* non-skewing systems whose baseline already exceeds ∼60% (−13.4%), consistent with a missing skewing prior alongside rule interference where the model already has a working approximation.

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.

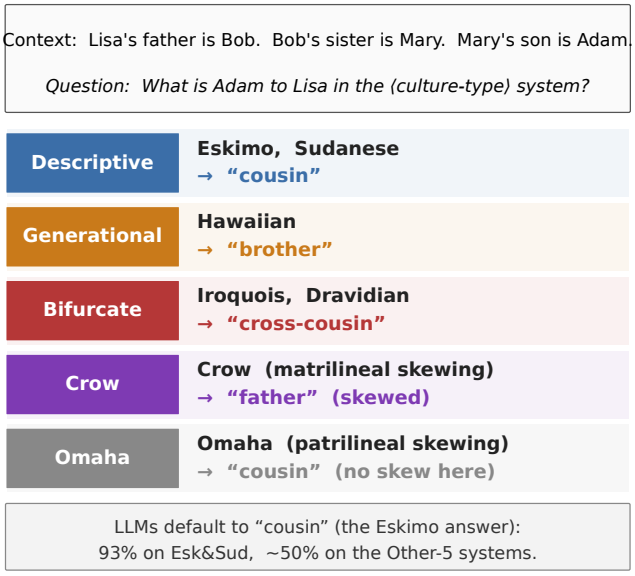

Figure 1: KinshipQA cultural disambiguation: the same kinship configuration (Lisa's father's sister's son Adam) maps to four distinct cultural answers across the typological classes; Crow's matrilineal skewing reclassifies Adam as "father" while Omaha's patrilineal skewing leaves this position unchanged. LLMs default to the descriptive Eskimo answer ("cousin"): 93% on Eskimo+Sudanese (hereafter Esk&Sud), ∼50% on the Other-5 systems.

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, and kin vocabulary belongs to the most stable layers of language (Swadesh, 1952). 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.

Many cultures and their languages lexicalise kin relations that span several nodes in a family tree (e.g. Bulgarian девер, denoting one's husband's brother), letting speakers refer succinctly to specific relatives without spelling out the intermediary chain. Conversely, in some societies the closest translation of a kin term may refer not only to a biological mother but also to maternal aunts. Despite this cross-linguistic variation, any biological kin relation can in principle be described from 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.

This paper makes the following contributions:

- **A procedurally-extensible kinship benchmark generator**: a tunable simulator horizon (50-, 150-, and 200-year) scales population size and chain depth on demand, eliminating exact-instance overlap with pretraining data.
- **KinshipQA**: a 4,062-question benchmark across seven anthropologically-documented kinship systems, four reasoning categories, and reasoning depths from 1 to 6 hops.
- **Cultural disambiguation imposes a robust additional burden beyond multi-hop chaining.** We document a 40.9% Cat. 1–3 $\rightarrow$ Cat. 4 EM drop on the five non-descriptive systems, ordered by anthropological typology (descriptive $>$ bifurcate $>$ generational $>$ skewing). The drop persists under chain-of-thought, few-shot prompting, and depth-matched comparison, and compounds with chain depth, falling to 10.6% at 5–6 hops. A two-annotator human baseline ($n = 280$, IAA 96.8%) reaches 89.0% on this category under matched in-context rule supply; LLMs given the same rule reach only 50.7%, placing the failure in the model rather than the task.
- **Two interventions on the Cat. 4 gap.** A fictional-rule control replaces system labels and kin terms with invented strings and raises Cat. 4 accuracy by 6.1%; an in-context-rule probe prepends the override rule and produces opposite-signed effects across typology, reducing accuracy on non-skewing systems with high baseline and raising it on skewing systems. The two interventions together help distinguish surface-form effects from a possible missing skewing prior; per-(model, system) results in App. H and the orthogonal 2×2 label/kin-term decomposition in App. G.

2 Related Work

Multi-hop reasoning benchmarks. Multi-hop reasoning remains a fundamental challenge for LLMs (Qiao et al., 2023). HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020) face training-data contamination and lack reasoning-chain verification; MuSiQue (Trivedi et al., 2022) mitigates the latter via single-hop composition under a connectivity constraint, while MuSR (Sprague et al., 2024) pushes multi-step reasoning into narrative domains. Recent benchmarks add structural demands or contamination resistance (MoreHopQA, Schnitzler et al., 2024; CofCA, Wu et al., 2024; CompoST, Schmidt et al., 2025), and domain-specific extensions (BioHopR, Kim et al., 2025; FamilyTool, Wang et al., 2025) show that hop count alone is incomplete. Rule-application benchmarks supply rules explicitly (RuleArena, Zhou et al., 2025; Thota et al. 2026); our main task instead names but does not state the cultural rule, with the in-context-rule probe (App. H) comparing both regimes. Li et al. (2024) localise multi-step failures to middle-layer self-attention.

Kinship reasoning benchmarks. There is a history of research using kin data as a machine learning testbed. Earlier knowledge-graph reasoning work used the standard Kinship multi-relational benchmark for path-based reinforcement learning (Lin et al., 2018), predating the LLM-era kinship benchmarks discussed below. CLUTRR (Sinha et al., 2019) uses kinship to evaluate com-

positional reasoning: from semi-synthetic stories (*Alice is Bob's mother; Bob is Carol's father*), models must infer unstated relations (*Alice is Carol's grandmother*). CLUTRR tests *rule induction* over a single (descriptive Eskimo) kinship typology. Two recent works extend kinship reasoning beyond CLUTRR's single-typology setting but along orthogonal axes to ours: Zhang et al. (2024) introduce a Chinese kinship reasoning dataset alongside CLUTRR as one of four evaluation domains, spanning a single descriptive-type system with richer maternal/paternal lexicalisation than English; and Lian et al. (2025) compile LingBench++, a cross-cultural kinship-term puzzle set drawn from the International Linguistics Olympiad archive, covering more than 90 low-resource languages as a static rule-induction benchmark in the IOL paradigm. Our work differs along three axes: we cover Morgan's seven major classification systems selected for typological coverage rather than language coverage; our questions are procedurally generated from RDF/OWL genealogies under formal culture-specific marriage-rule constraints, with a tunable simulator horizon that eliminates exact-instance overlap with pretraining; and we test rule *application* given a named cultural system and an in-context override cue, separating it from rule induction in part via a fictional-rule control (§5.2).

Cultural knowledge in NLP. Cross-cultural NLP has gained attention as researchers recognise that language technologies must account for cultural variation, with kinship identified as an under-researched category where evaluation data are lacking (Liu et al., 2024); existing cultural benchmarks have been criticised for reducing culture to static facts (AlKhamissi et al., 2025). Unlike biological kinship (universal facts of reproduction), cultural kinship operates as a well-defined "constructed, computational system" with formal rules and anthropologically-documented variation (Read, 2012), making it a natural testbed for 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: $f(\mathcal{C}, \mathcal{Q}) \rightarrow \mathcal{A}$.

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. See Appendix A for full anthropological descriptions and per-system override rules.

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

The key challenge lies in distinguishing between two reasoning modes: biological reasoning, which follows universal genealogical facts (e.g., "mother's sister" $\rightarrow$ "aunt"), and cultural reasoning, which requires applying culture-specific classification rules that may override biological defaults (e.g., in Hawaiian kinship, "mother's sister" $\rightarrow$ "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 exact-instance overlap with pretraining data.

Our seven systems span Morgan's descriptive–classificatory dichotomy (Morgan, 1871) as refined by Murdock (Murdock, 1949) into the six-fold classification (Eskimo, Hawaiian, Iroquois, Crow, Omaha, Sudanese), plus the Dravidian elaboration of Trautmann (Trautmann, 1981). Table 1 summarises them; full anthropological descriptions are in App. A.

3.3 Benchmark Generator Pipeline

Figure 2 illustrates our five-stage, fully automated pipeline. We simulate multi-generational family trees under culture-specific marriage constraints over tunable horizons (50-year baseline yields 60–70 individuals; 150/200-year extensions yield 60–360 individuals depending on system marriage rules)—e.g., Eskimo prohibits only sibling mar-

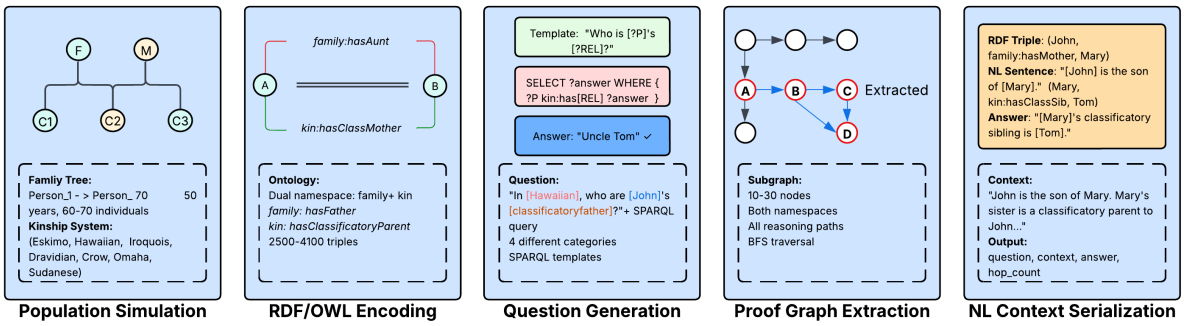

Figure 2: KinshipQA generation pipeline. (1) **Population Simulation** — multi-generational family trees under culture-specific marriage rules. (2) **RDF/OWL Encoding** — dual namespaces (`family:` biological; `kin:` cultural). (3) **Question Generation** — four categories at controlled hop depths. (4) **Proof Graph Extraction** — minimal reasoning subgraph per question. (5) **NL Context Serialisation** — RDF subgraph → NL for the LLM.

riage; Iroquois allows cross- but not parallel-cousin marriage; Dravidian prefers cross-cousin marriage; Crow and Omaha enforce clan-based prohibitions under matrilineal and patrilineal descent. Trees are encoded as RDF with a dual-namespace ontology—`family:` for biological relationships and `kin:` for cultural classifications—enabling precise evaluation of whether models distinguish biological facts from cultural categories. Questions are generated by mapping biological relationship paths (e.g., mother → sister → child) to system-specific kinship terms, ensuring genuine multi-hop reasoning rather than memorised labels; ground truth is computed by SPARQL queries against the ontology, and the minimal sub-graph supporting each question is serialised to natural language by templates (e.g., "Alice is the mother of Bob"). The benchmark therefore has a machine-checkable closure: a deterministic SPARQL executor over the released RDF/OWL ontology returns the gold answer on 100% of the 4,062 questions by construction, so the LLM gaps reported below arise from natural-language interpretation and reasoning over the supplied facts rather than gold inconsistency or evaluator ambiguity (the natural-language surface realisation is validated separately by the human baseline in App. I). The full generator, the 4,062-question benchmark, and the evaluation harness are released in the anonymised supplementary repository.[1]

3.4 Question Categories

We define four question categories of progressively more complex reasoning. **Cat. 1 (Fact Retrieval)** is a 1-hop lookup of an explicitly stated relation (*Who is John's father?*), a baseline for context extraction. **Cat. 2 (Multi-Hop Biological)** chains 2–4 biological steps (*paternal grandfather* = father→father), testing compositional reasoning over universal genealogical structure. **Cat. 3 (Counting/Filtering)** adds logical operations on top of Cat. 2: counting (*How many siblings does Mary have?*), filtering (*List X's male cousins*), and comparison. **Cat. 4 (Cultural Disambiguation)** tests whether LLMs can deploy culture-specific classification rules that diverge from the biological default. The model receives the system name and an in-context cultural query (*According to Hawaiian kinship, who are X's parents?* relies on the convention that mother's sisters are classified as "mothers"). *Explicit rule statements are never given in the prompt*: Cat. 4 therefore tests application of a *named latent rule system*—combining recall of the cultural prior with override behaviour—rather than application of an explicitly supplied in-context rule. The fictional-rule control (§5.2) probes the contribution of cultural-prior interference to the residual failure.

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 a seed population of four individuals and growing to 60–70 individuals under the hard-wired offspring constraints (full simulator parameters in the supplementary repository). The resulting trees are used to evaluate and compare a number of LLMs popular at present.

Table 2 reports overall statistics; per-cell counts (system × category × hop) are released alongside the dataset. The largest cells (e.g. Cat. 2 at 2–3

[1] https://anonymous.4open.science/r/Kinship-Data-Benchmark-for-Multi-hop-Reasoning-35F1/; CC BY 4.0.

Table 2: KinshipQA dataset statistics for the 1–4 hop subset (3,134 questions); 5–6 hop Cat. 2 coverage adds 708 further questions across all 7 systems (App. E). "Override" is the fraction of questions requiring cultural reclassification (e.g. mother's sister → classificatory mother in Hawaiian); only Cat.4 questions on the Other-5 systems are override-eligible. Sudanese is 0% because its relational structure aligns with the biological default—its lexical uniqueness (a distinct term per kin relation) does not constitute structural override.

	Questions	%	Override
By System Type			
Eskimo, Sudanese	974	31.1	0%
Remaining five	2,160	68.9	30.6%
By Category			
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%
By Complexity			
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%**

hops on each system) contain 100–350 questions. The cultural-override path library covers 2-, 3-, and 4-hop paths per Other-5 system, anthropologically grounded in Morgan's typology (e.g., Hawaiian: parallel-uncle's grandchild → classificatory child; Crow: father's sister's grandchild → skewed matrilineage member; full path table in Appendix C.4). At 4-hop, this contributes 34–50 Cat. 4 questions per Other-5 system (Hawaiian/Iroquois/Dravidian: 50; Crow: 36; Omaha: 34; 220 in total)—sufficient for per-cell mean estimates with bootstrap CIs as reported in §5.1.

Generation controls. Templates are organised by category and hop depth (4 / 12 / 6 / per-system override templates for Cat. 1–4 respectively); names are sampled uniformly from a fixed roster of common first names with 4 surnames per tree; and context paragraphs include only the minimal subgraph required for the SPARQL ground-truth derivation, with no added distractor facts—reasoning difficulty derives from chain length and cultural-rule application rather than information-filtering noise.

4 Results and Evaluation

4.1 Experimental Setup

Models and protocols. We evaluate six LLMs: three open-source 27–32B models — Qwen3-32B (Yang et al., 2025), Gemma3-27B (Team et al., 2025), DeepSeek-R1-Distill-Qwen-32B (hereafter DeepSeek-R1-32B; DeepSeek-AI et al., 2025) — and three closed-source mid-tier APIs — GPT-4o-mini (OpenAI et al., 2024), Claude-Haiku-4.5 (Anthropic, 2025), Gemini-2.5-Flash (Comanici et al., 2025). All models receive identical zero-shot prompts under greedy decoding (temperature = 0). We additionally evaluate the open-source models and two closed-source models (GPT-4o-mini, Claude-Haiku-4.5) under zero-shot and few-shot CoT; Gemini-2.5-Flash is omitted from the CoT matrix due to free-tier quota.

Metric and infrastructure. We report **Exact Match (EM)** as our primary metric; multiple-valid-answer questions (43.8%) are scored set-based.[2] Because the symbolic SPARQL executor returns the gold on 100% of questions by construction (§3.5), the LLM gaps reported below isolate natural-language interpretation and reasoning rather than evaluator drift. Open-source models run via Ollama on a single H100 80GB GPU (Q4_K_M quantization, *max_tokens* = 512 for direct / 2048 for CoT); closed-source models use vendor APIs with matched settings. Model identifiers, parsers, prompt templates, and the dataset generator are released in the anonymised supplementary repository (§3.5).

Table 3 presents our main findings under zero-shot direct prompting. Table 4 extends to all three protocols (zero-shot direct, zero-shot CoT, few-shot CoT) on the five models with closed-source coverage stable enough to run the full matrix; Gemini-2.5-Flash is omitted from CoT protocols due to API quota.

Few-shot CoT is best on overall accuracy for 4 of 5 models (Avg. 84.6 → 88.1%) and raises Cat. 4 Other-5 accuracy by +3.6% on average. Gemma3 and GPT-4o-mini drop slightly on Cat. 4 under FS-CoT (Gemma3 65.3 → 63.6%, GPT-4o-mini 62.6 → 56.1%) despite gains on Cat. 1–3, consistent with demonstrations drawn from other systems biasing these models toward the demonstrated systems' rules. In every protocol on every model, the Esk&Sud descriptive controls reach 92–100% Cat. 4 accuracy while the Other-5 systems plateau at 37–67% (§5.1).

[2] Bootstrap CIs use N=10,000 percentile-method resamples (unit: question for hop-matched comparisons, (system, model) cell for ablations; unpaired, stratified within group).

Table 3: Main results on KinshipQA (Exact Match %; F1 in brackets for the multi-answer categories Cat. 3 and Cat. 4). **Scope:** Cat. 1–3 are computed on all 7 systems; Cat. 4 is computed only on the Other-5 systems (Hawaiian, Iroquois, Dravidian, Crow, Omaha), since Esk&Sud (Eskimo+Sudanese) are descriptive controls without override paths by construction. Δ Gap = Other-5 − Esk&Sud over Cat. 1–4 (negative entries indicate a drop on the Other-5 systems). All six models are evaluated under zero-shot direct prompting with greedy decoding (temperature = 0); the Avg. row is the unweighted mean over the six models. Under F1 (partial-credit set match), Cat. 4 Other-5 scores rise by 8–11%, but the cultural-override gap persists: the F1 gap remains +18 to +43% across models, preserving the same descriptive → skewing hierarchy.

	By Culture			By Category			
Model	Esk&Sud	Other 5	Δ Gap	Cat.1	Cat.2	Cat.3 (F1)	Cat.4 (F1)
Closed-Source Models							
GPT-4o-mini	92.3	81.2	−11.1	100.0	90.4	86.1 (86.1)	62.6 (72.5)
Claude-Haiku-4.5	98.3	85.3	−13.0	100.0	98.3	87.7 (87.9)	65.0 (72.8)
Gemini-2.5-Flash	98.2	84.3	−13.9	100.0	99.5	86.1 (86.1)	61.4 (70.8)
Open-Source Models							
Qwen3-32B	92.2	77.9	−14.3	99.4	96.0	86.1 (86.1)	44.8 (53.2)
Gemma3-27B	96.8	84.7	−12.1	100.0	98.4	84.9 (85.7)	65.3 (73.4)
DeepSeek-R1-32B	92.7	73.8	−18.9	89.1	94.6	82.1 (88.1)	41.1 (52.0)
Avg.	95.1	81.2	−13.9	98.1	96.2	85.5 (86.7)	56.7 (65.8)

Table 4: Accuracy by protocol on the five models with full protocol-matrix coverage (Exact Match %). **Overall** averages all four categories on all 7 systems; **Cat. 4** reports Other-5 Cat. 4 EM. Few-shot CoT raises Overall on 4 of 5 models and Cat. 4 Other-5 on 3 of 5; the cultural-override gap persists in every cell. Per-system protocol breakdown in App. F.

	Overall (%)			Cat.4 (%)		
Model	Dir.	CoT	FS	Dir.	CoT	FS
GPT-4o-mini	84.4	**89.1**	87.5	62.6	**65.0**	56.1
Claude-Haiku-4.5	89.0	89.3	**89.6**	65.0	65.0	**67.0**
Gemma3-27B	88.2	60.7	**88.9**	65.3	37.0	63.6
DeepSeek-R1	79.2	74.1	**86.9**	41.1	39.2	**52.9**
Qwen3-32B	82.0	83.8	**87.8**	44.8	49.8	**57.6**
Avg.	84.6	79.4	**88.1**	55.8	51.2	**59.4**

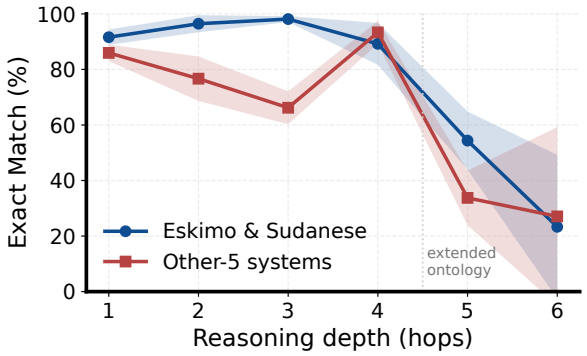

Figure 3: Overall performance by reasoning depth (EM %, zero-shot direct, mean across the three open-source models; shaded bands: ± cross-model std), split into Esk&Sud (descriptive controls) and the Other-5 systems where cultural override applies. The 1–4 hop average mixes Cat. 1–4; the 5–6 hop average covers 708 Cat. 2 questions on all 7 systems plus 249 Cat. 4 questions on the Other-5 systems (see §4.2 and Appendix E). The vertical dotted line at 4.5 marks the simulator boundary between the 50-year ontology and the extended 150/200-year ontologies.

4.2 Performance by Reasoning Complexity

To probe scaling beyond four hops, we regenerate ontologies with extended genealogies for all 7 systems (150 years for Eskimo, Sudanese, Iroquois, Dravidian; 200 years for Hawaiian, Crow, Omaha, whose more restrictive marriage rules require a longer horizon). This yields 708 new 5–6 hop Cat. 2 questions across all 7 systems and 249 5–6 hop Cat. 4 cultural-override questions on the Other-5 systems (Esk&Sud have no override paths by construction). Per-hop Cat. 4 accuracy on the 50-year ontology is visualised in App. E, Figure 5.

Within Cat. 4 at 1–4 hops, accuracy on the Other-5 systems drops monotonically (58.4 → 42.5 → 18.9%) while Esk&Sud declines only modestly (96.0 → 94.4 → 85.5%); the cultural-override gap widens from 37.6% (2-hop) to 66.5% (4-hop), suggesting that cultural disambiguation and chain depth compound. At 5–6 hops, Other-5 Cat. 4 EM falls further to 10.6% (−39.9% from the 1–4 hop baseline of 50.4%; per-system drops range −7 to −52%, App. E). Cross-model divergence is sharp at this depth: on Cat. 2, Gemma3 retains 72% at 6-hop while DeepSeek and Qwen3 collapse to 4% and 3% (App. E). The within-Other-5 ordering partially scrambles at this depth (Omaha 25.3%, Crow 0.0% on a small 16-question cell with multi-name targets; we read the Crow EM against its F1 of 7.9% rather than in isolation), reflecting per-system

Table 5: Performance by kinship system (EM %, mean ± std across 3 open-source models; ordered descriptive → skewing). "All Cat." averages Cat. 1–4. The right block reports Cat. 4 with cultural override (applicable only to Other-5 systems; Esk&Sud entries "–") against the per-system mean over Cat. 1–3. Humans given the same override rule reach 89.0% on Other-5 Cat. 4 vs. 50.7% for LLMs (App. I). The descriptive → skewing ordering holds under F1 (per-system Cat. 4 F1 in App. H).

System	All Cat.	w/ override	Cat. 1–3	Δ
Eskimo	94.1 ± 2.1	–	94.8 ± 1.9	–
Sudanese	93.7 ± 3.0	–	94.8 ± 1.9	–
Hawaiian	82.6 ± 3.0	61.9 ± 7.2	91.8 ± 2.7	29.9
Iroquois	78.2 ± 5.5	50.3 ± 14.8	90.6 ± 1.6	40.3
Dravidian	81.2 ± 6.8	60.4 ± 16.6	90.3 ± 3.3	30.0
Crow	78.4 ± 6.7	47.0 ± 15.7	92.2 ± 3.1	45.3
Omaha	73.7 ± 5.8	32.6 ± 13.7	91.8 ± 2.7	**59.2**
Other-5 avg.		50.4 ± 11.8	91.3 ± 0.8	40.9

path-structure differences at the extended-horizon ontology boundary. On the overall 1–4 hop curve (Figure 3), apparent local non-monotonicities are category-composition shifts rather than depth reversals, since Cat. 4 tapers from the Other-5 average as depth grows; full per-model curves in App. E, consistent with the compositionality limits of Dziri et al. (2023) and Lin et al. (2025).

4.3 Performance by Kinship System

Table 5 reveals a clear hierarchy across system types, ordered descriptive → bifurcate → generational → skewing: Eskimo and Sudanese sit near ceiling, the bifurcate (Iroquois, Dravidian) and generational (Hawaiian) systems are intermediate, and the two skewing systems (Crow, Omaha) are hardest—Omaha bottoming at 32.6% Cat. 4 EM (45.6% F1). Comparing per-system Cat. 4 (cultural rule application) against Cat. 1–3 (biological reasoning) highlights the empirical cost of the override setting: switching from biological to cultural reasoning drops accuracy by 40.9% EM (and 31.8% F1) on average across the Other-5 systems, with the largest gaps in skewing systems (Omaha 59.2% EM, Crow 45.3% EM) and the smallest in Hawaiian and Dravidian (29.9 and 30.0% EM); under F1 the same descriptive → skewing ordering holds (Omaha 45.6, Crow 53.2, Iroquois 63.2, Dravidian 65.9, Hawaiian 69.8). Generational merging (Hawaiian) and cross-cousin rules (Dravidian) thus appear more learnable than generational skewing on both metrics. Human annotators (App. I; $n = 280$, two graduate-student volunteers, IAA 96.8%) reach 89.0% on Other-5 Cat. 4 when given

Table 6: Hop-matched comparison of Cat.2 (biological multi-hop, all 7 systems) vs Cat.4 (cultural override, Other-5 systems) on identical input format. The gap **widens** from 38.0% at 2-hop to 56.3% at 3-hop, opposite to a lookup-disruption prediction and suggesting that cultural-rule application adds difficulty beyond chain depth. Per-model breakdown shows the gap holds across all three open-source models. (DeepSeek-R1-32B's small Cat.2 increase from 2-hop (89.7) to 3-hop (95.4) reflects per-cell sample variance; the cross-model average and the cultural-gap pattern are unchanged by it.)

Model	Hop	Cat.2	Cat.4	Gap
Gemma3-27B	2	100.0	76.7	23.3
	3	100.0	53.9	46.1
DeepSeek-R1-32B	2	89.7	48.8	40.9
	3	95.4	33.3	**62.1**
Qwen3-32B	2	100.0	50.3	49.7
	3	100.0	39.4	60.6
Average	**2**	**96.6**	**58.6**	**38.0**
Average	**3**	**98.5**	**42.2**	**56.3**

the same one-sentence rule as the in-context-rule probe, compared with 50.7% for LLMs under the same rule-supplied setting. This indicates the questions are intelligible under matched rule access rather than ambiguous.

Each step up the hierarchy adds a structural demand: descriptive systems impose no override, generational and bifurcate require a one-generation reclassification (sex-and-generation matching; parallel-vs-cross), and skewing requires generation-spanning reclassification keyed to lineage membership — the most counterintuitive override and the rarest in pretraining corpora, consistent with the in-context-rule probe's evidence for a weaker skewing prior (§5.2).

5 Cultural Findings

5.1 Robustness: Not a Prompting or Depth Artefact

Prompting. The Esk&Sud → Other-5 gap **persists under all three protocols** for the five models with full protocol-matrix data (App. F). Few-shot CoT raises Cat. 4 Other-5 accuracy by +3.6% on average but shifts the gap inconsistently across models. In every protocol on every model, the descriptive controls reach 92–100% Cat. 4 accuracy while the Other-5 systems plateau at 37–67%, consistent with Sprague et al. (2025)'s finding that CoT helps mainly on symbolic-format reasoning.

Chain depth. Could the Cat.2→Cat.4 drop simply reflect task difficulty rather than cultural varia-

Table 7: Fictional-rule control: orthogonal decomposition of the +6.1% Cat. 4 gain into system-name and kin-term swaps (Other-5 systems, mean across 3 open-source models, Δ vs. real-label baseline; 95% bootstrap CI over 15 (system, model) cells). Either swap alone is null; the joint swap is +6.1, indicating an interaction.

Condition	Real EM	Δ vs. Real	95% CI
Real (baseline)	50.4	–	–
+ system-name swap	51.7	+1.3	[−1.2, +4.1]
+ kin-term swap	49.2	−1.2	[−4.4, +2.3]
+ both (full)	56.5	**+6.1**	**[+2.8, +9.4]**

Table 8: In-context-rule probe: mean ΔEM (%) by override typology and model baseline (n = model×system pairs). Skewing cells uniformly help; non-skewing high-baseline cells uniformly hurt. Full per-cell data in App. H.

Baseline EM	Non-skewing	Skewing
$\geq 60\%$	-13.4 (n=4)	$+12.0$ (n=1)
$< 60\%$	$+4.9$ (n=5)	$+17.9$ (n=5)

tion? Table 6 compares Cat.2 and Cat.4 **at matched hop depths**. The cultural gap is substantial at every cell (23.3–62.1%) and *widens* with depth (38.0% at 2-hop → 56.3% at 3-hop), opposite to a lookup-disruption prediction. These results suggest that cultural override and multi-hop complexity compound.

Phrasing, decoding, and frontier scale. A rule-paraphrased Cat. 4 subsample on Gemma3-27B preserves accuracy (78.0 → 84.0%; App. K), and multi-seed decoding (T=0.7, n=5) is 3–14% lower than greedy but preserves the system-type ordering (App. E). At frontier scale, Claude Opus 4.7 reaches 100% on Esk&Sud but 69% Other-5 under zero-shot direct (gap 31%, narrowing to 21.6% under FS-CoT, App. J); the depth × override interaction also widens at frontier scale (22.4 → 47.6% from 2- to 3-hop).

5.2 Probing Cat. 4 Failure: Two Interventions

Cat. 4 errors are dominated by biological-default leakage: 53.3% of the 1,000 non-EM predictions are bare-default substitution (32.8%) or over-inclusion (20.5%); only 9.8% are wrong-disjoint (App. D). Two interventions probe this gap.

Fictional-rule control. Replacing both the system label and the kin terms with phonotactically distinct invented strings raises Cat. 4 Other-5 accuracy by +6.1% (95% CI [+2.8, +9.4]); replacing either alone is null (Table 7), so the gain is a surface-feature *interaction* rather than additive. A pure knowledge deficit predicts the opposite direction; accuracy instead rises on all three open-source models (per-model Other-5 means: Gemma3 +1.1, DeepSeek +5.7, Qwen3 +11.5), strongest on the hardest systems (Iroquois, Dravidian, Omaha) and the two reasoning-focused models, pointing to familiar surface forms as one contributor to the Cat. 4 gap (per-cell breakdown in App. G).

In-context rule probe. If failure were instead a deficit of rule *recall*, supplying the rule in context should close the gap. We prepend a one-sentence override rule to each Cat. 4 question and evaluate the three open-source models on the full Other-5 set (1,254 questions/model; App. H). The effect splits along a typology × baseline axis (Table 8). Skewing systems show a uniform rule-recall deficit (6/6 pairs HELP, mean +17.1%; per-model on Crow/Omaha: Gemma3 +12/ + 28, DeepSeek +14/ + 19, Qwen3 +6/ + 23). Non-skewing systems above ∼60% baseline EM are *hurt* by the rule (−9 to −18%; 4/4 pairs, mean −13.4), consistent with interference between the supplied rule and an internalised approximation; below ∼60% the rule helps or is flat (descriptive-control noise floor ≤ 7.5%).

6 Conclusion

KinshipQA spans seven anthropologically-documented kinship systems with a cultural-disambiguation category. The 40.9% Cat. 1–3 → Cat. 4 drop on the five non-descriptive systems follows the typological ordering of Table 5, persists under chain-of-thought, few-shot, stochastic decoding, depth-matched controls, and frontier-scale evaluation, and compounds with chain depth (10.6% at 5–6 hops). Two interventions help distinguish two contributing factors: removing familiar surface forms raises accuracy by 6.1% (pointing to surface forms as one contributor); supplying the override rule helps skewing systems but reduces accuracy on non-skewing systems with high baseline (consistent with a knowledge gap and rule interference respectively). Humans given the same rule reach 89.0% vs. 50.7% for LLMs — a 38.3% gap locating the failure in the model rather than the task. Benchmarks limited to a single descriptive typology may therefore overestimate cultural capabilities.

Limitations

KinshipQA covers seven kinship systems from Morgan's typology, implemented as idealised anthropological models in English; the named systems are analytical abstractions, not rigid descriptions of contemporary practice, and the generator assumes a binary-sex two-parent biological family structure. Our "cultural disambiguation" construct captures performance on stylised English glosses of Morgan-typology overrides; pretraining-frequency effects of these specific glosses are not separately measured. Contexts contain only the minimal subgraph required for SPARQL ground truth (no distractor facts), so KinshipQA is a controlled diagnostic for cultural disambiguation over symbolic relational chains. Closed-source coverage in our headline tables is the cost-controlled mid-tier (GPT-4o-mini, Claude-Haiku-4.5, Gemini-2.5-Flash); we evaluate Claude Opus 4.7 at frontier scale on a Cat. 4 subset (Appendix J). Fine-tuning and retrieval-augmented regimes are out of scope. Symbolic gold answers are exact by construction; a two-annotator human baseline ($n = 280$, IAA 96.8%; App. I) reaches 95.4% overall and 89.0% on Other-5 Cat. 4 with the same in-context rule used in the in-context-rule probe (§5.2); the like-for-like LLM-with-rule mean is 50.7%. Multilingual native-speaker validation is left to future work.

Acknowledgments

Generative AI tools (Claude, Anthropic) were used during revision for language polishing and code review; all research contributions and substantive analyses were authored and verified by the human authors.

References

Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. Hire Your Anthropologist! Rethinking Culture Benchmarks Through an Anthropological Lens.

Anthropic. 2025. Introducing Claude Haiku 4.5. Released October 15, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, 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, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and Fate: Limits of Transformers on Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. 2025. BioHopR: A Benchmark for Multi-Hop, Multi-Answer Reasoning in the Biomedical Domain. *Preprint*, arXiv:2505.22240. ArXiv:2505.22240.

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.

Da-Chen Lian, Ri-Sheng Huang, Pin-Er Chen, Chunki Lim, You-Kuan Lin, Guan-Yu Tseng, Zi-Cheng Yang, Zhen-Yu Lin, Pin-Cheng Chen, and Shu-Kai Hsieh. 2025. LingBench++: A Linguistically-Informed Benchmark and Reasoning Framework for Multi-Step and Cross-Cultural Inference with LLMs. *Preprint*, arXiv:2507.16809. ArXiv:2507.16809.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-Hop Knowledge Graph Reasoning with Reward Shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3253. ArXiv:1808.10568.

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.

George Peter Murdock. 1949. *Social Structure*. Macmillan, New York.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, 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. 2025. CompoST: A Benchmark for Analyzing the Ability of LLMs to Compositionally Interpret Questions in a QALD Setting. In *Proceedings of the 24th International Semantic Web Conference (ISWC 2025)*, Nara, Japan. Springer.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. MoreHopQA: More Than Multi-hop Reasoning. *Preprint*, arXiv:2406.13397. ArXiv:2406.13397.

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.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *International Conference on Learning Representations (ICLR)*. ArXiv:2310.16049.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or Not to CoT? Chain-of-Thought Helps Mainly on Math and Symbolic Reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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].

Yogeswar Reddy Thota, Setareh Rafatirad, Houman Homayoun, and Tooraj Nikoubin. 2026. When Models Ignore Definitions: Measuring Semantic Override Hallucinations in LLM Reasoning. *Preprint*, arXiv:2602.17520. ArXiv:2602.17520.

Thomas R. Trautmann. 1981. *Dravidian Kinship*. Cambridge University Press, Cambridge.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554. ArXiv:2108.00573.

Yuxin Wang, Yiran Guo, Yining Zheng, Zhangyue Yin, Shuo Chen, Jie Yang, Jiajun Chen, Yuan Li, Xuanjing Huang, and Xipeng Qiu. 2025. FamilyTool: A Multi-hop Personalized Tool Use Benchmark. *Preprint*, arXiv:2504.06766. ArXiv:2504.06766.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2024. CofCA: A STEP-WISE Counterfactual Multi-hop QA benchmark. *Preprint*, arXiv:2402.11924. ArXiv:2402.11924; ICLR 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen 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.

Ge Zhang, Mohammad Ali Alomrani, Hongjian Gu, Jiaming Zhou, Yaochen Hu, Bin Wang, Qun Liu, Mark Coates, Yingxue Zhang, and Jianye Hao. 2024. Path-of-Thoughts: Extracting and Following Paths for Robust Relational Reasoning with Large Language Models. *Preprint*, arXiv:2412.17963. ArXiv:2412.17963.

Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang

Wang. 2025. RuleArena: A Benchmark for Rule-Guided Reasoning with LLMs in Real-World Scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. ArXiv:2412.08972.

A Kinship Systems and Override Mappings

A.1 System Descriptions

Eskimo (Lineal). 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.

A.2 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

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 (descriptive lineal) system, demonstrating that chain complexity—not cultural knowledge—is the limiting factor.

Example 1 illustrates a chain-tracking failure in the descriptive (Eskimo) system; the cultural-default and generational-skewing failure modes are quantified across all 1,000 non-EM Cat. 4 predictions in the automatic error taxonomy of §D.

C Prompt Templates and 4-hop Cultural-Override Paths

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 Zero-Shot Chain-of-Thought 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:
End with the line "FINAL ANSWER: <answer>".
```

C.3 Few-Shot CoT Prompt (sketch)

For few-shot CoT we prepend three worked examples per question category to the zero-shot CoT template above. Each worked example contains a short context, a question, the five reasoning steps, and a `FINAL ANSWER` line, illustrating the expected reasoning format and an instance of cultural-rule application where applicable.

C.4 4-hop Cultural-Override Paths

The cultural-override path library includes anthropologically grounded paths at hop depths 2, 3, and 4 per Other-5 system, supporting within-Cat.4 hop-scaling analysis in §4.2. At 4-hop, the library produces ∼50 Cat.4 questions per Other-5 system. The per-system path mappings extend the same logic as the 2-hop overrides documented in Table 9 (App. A.2): same-sex chains remain parallel in Iroquois/Dravidian; FZ-line / MB-line members remain skewed across generations in Crow/Omaha; G±1 relatives become own parents/children in Hawaiian. Full mappings are in the supplementary repository.

Cross-model variance at 4-hop. The cross-model std rises sharply at 4-hop (e.g., Hawaiian ±24.6, Omaha ±29.2; visualised per-system in Figure 4). Per-model breakdowns show a stark split: Gemma3-27B retains 34–67% accuracy on 4-hop Other-5 Cat.4 while DeepSeek-R1-32B and Qwen3-32B collapse to 0–11%. The two reasoning-focused models are more affected by the depth × override interaction than the general-purpose Gemma3, an observation we leave to future work.

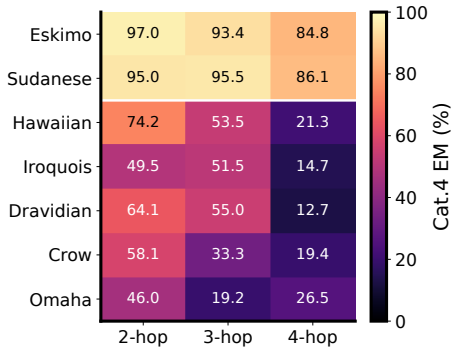

Figure 4: Per-system Cat.4 EM (%) by hop (zero-shot direct, mean across the three open-source models). Eskimo and Sudanese (above the white divider) serve as descriptive controls; the five Other-5 systems below decline sharply with depth, with the bifurcate systems (Dravidian, Iroquois) and Omaha hitting the lowest 4-hop accuracies.

D Error Analysis Detail

Automatic Cat. 4 error decomposition. We automatically classify all 1,000 non-EM Cat. 4 predictions across the three open-source models (zero-shot direct, Other-5 systems) into six error types using ground-truth-vs-prediction set arithmetic and relationship-label structure (Table 10). *Default substitution* (32.8%) and *over-inclusion* (20.5%) together account for 53.3% of all Cat. 4 errors—both forms of biological-default leakage in which the model fails to apply the cultural override correctly. *Default substitution* is overwhelmingly concentrated in reverse-mapping (Pattern C) questions, accounting for 83.9% of errors when the model is asked to name the underlying biological relationship behind a cultural kin term. Genuinely wrong-disjoint answers (no overlap with the gold) account for only 9.8% of errors; the remaining 35.8% are empty/format errors, concentrated in the reasoning models (DeepSeek/Qwen) where long chains exceed the answer-line parser's recovery range. The 53.3% biological-default leakage figure is the mechanism signature targeted by the two interventions in §5.2. A parallel manual annotation of 200 CoT responses from GPT-4o-mini and Gemma3-27B yielded a consistent 30.4% cultural-variation share (cultural-default + over-inclusion), giving a lower-bound corroboration of the automatic estimate.

Table 10: Cat. 4 error taxonomy across the 1,000 non-EM predictions on the Other-5 systems (3 OS models, zero-shot direct). Categories are mutually exclusive. "Default sub." = bare biological term (uncle, aunt, ...) where the gold requires a paternal/maternal/cross/parallel/classificatory marker; "over-inclusion" = pred $\supset$ gold (biological default emitted alongside cultural members); "partial under" = pred $\subset$ gold; "partial overlap" = mixed; "wrong / disjoint" = no overlap with gold; "empty / format" = no parseable answer. The 32.8% default substitution rises to 83.9% on reverse-mapping Pattern C questions.

Error type	N	%
Biological-default leakage (53.3%)		
Default substitution	328	32.8
Over-inclusion	205	20.5
Set-shape errors (1.1%)		
Partial under-coverage	7	0.7
Partial overlap	4	0.4
Other (45.6%)		
Wrong / disjoint	98	9.8
Empty / format	358	35.8
Total Cat. 4 errors	**1000**	**100**

E Hop Scaling and Stochastic Decoding

Table 11 reports the per-model hop curves underlying Figure 3. DeepSeek-R1-32B and Qwen3-32B collapse to single-digit accuracy at 6 hops while Gemma3-27B partially recovers to 66%, indicating divergent failure modes beyond the original 4-hop ceiling.

Table 11: Hop scaling by model (Exact Match %, zero-shot direct, averaged across systems with data at each depth).

Model	1	2	3	4	5	6
Gemma3-27B	89	92	81	85	41	72
DeepSeek-R1-32B	83	80	71	89	58	**4**
Qwen3-32B	90	84	75	82	44	**3**
Average	88	85	76	85	48	27

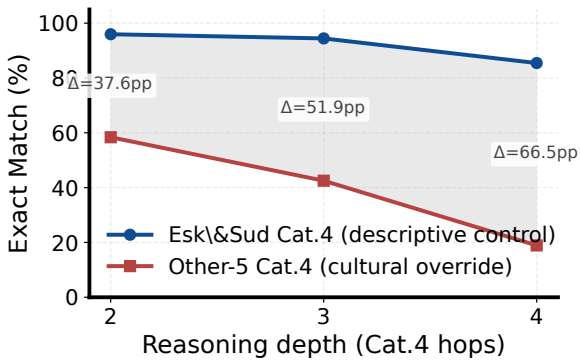

Figure 5: Cat. 4 accuracy by reasoning depth on the 50-year ontology (zero-shot direct, mean across the three open-source models; per-cell counts in §3.5). The cultural-override gap between Esk&Sud (descriptive control) and the Other-5 systems widens monotonically with depth, from 37.6% at 2-hop to 66.5% at 4-hop. 5–6 hop Cat. 4 data are reported in Appendix E.

Multi-seed robustness check. The accuracies reported in the main paper use greedy decoding (temperature $= 0$) for reproducibility, with the $\pm$ values in Tables 5, 6, and 14 reflecting cross-model dispersion across the three open-source models. To confirm that the system-type hierarchy is not a single-trajectory artefact, we evaluate Cat.4 on a stratified subsample of 40 questions per kinship system (280 total) using temperature $= 0.7$ and $n = 5$ samples per question. Each per-question accuracy is the fraction of the 5 samples that exact-match the ground truth.

Multi-seed accuracies are 3–14% lower than the greedy numbers in Table 5, consistent with greedy decoding selecting the modal output (more often correct on this benchmark) and stochastic sampling exposing the underlying decision-noise. Critically,

Table 12: Multi-seed Cat.4 evaluation (temperature = 0.7, $n = 5$ samples, 40-question stratified subsample per system). Mean EM across the three open-source models with ± cross-model dispersion; per-model bootstrap 95% CIs in the supplementary repository. The typological ordering of Table 5 is preserved.

System	Type	Multi-seed EM (%)
Eskimo	Descriptive	91.0 ± 5.8
Sudanese	Descriptive	88.7 ± 10.8
Hawaiian	Generational	57.3 ± 5.1
Iroquois	Bifurcate	43.2 ± 18.3
Dravidian	Bifurcate	61.0 ± 21.4
Crow	Mat. Skewing	49.8 ± 19.1
Omaha	Pat. Skewing	**33.2 ± 18.1**

the relative ordering is preserved: descriptive systems (Eskimo, Sudanese) remain $\geq 88\%$, while skewing systems (Crow, Omaha) remain at the bottom, with Omaha still the lowest at 33.2%. The cross-model dispersion is wider on the bifurcate and skewing systems—further evidence that these are the categories where models genuinely diverge in capability rather than agree on a wrong answer.

5–6 hop Cat. 4 detail. Table 13 reports per-system × per-model accuracy on the 249 5–6 hop cultural-override questions discussed in §4.2. The Other-5 mean EM is 10.6% (F1 18.9%), down from 50.4% at 1–4 hop (Table 5). The per-system ordering at 5–6 hop is Omaha > Hawaiian > Dravidian > Iroquois > Crow, which differs from the Hawaiian/Dravidian/Iroquois/Crow/Omaha ordering observed at 1–4 hop. Crow's cell ($n = 16$) uses six-name set targets; EM is 0% and F1 is 7.9%, the latter reflecting partial overlap with the two ground-truth clusters.

Table 13: 5–6 hop Cat. 4 cultural-override accuracy (EM % / F1 %, zero-shot direct, max_tokens = 8192). n is the per-system question count. Mean is the per-system unweighted average over the three open-source models.

System	n	Gemma3	DeepSeek	Qwen3	EM	F1
Hawaiian	40	35.0/58.4	2.5/28.6	7.5/13.8	15.0	33.6
Iroquois	80	7.5/12.7	3.8/ 6.3	1.2/ 2.2	4.2	7.1
Dravidian	80	17.5/24.3	2.5/ 5.0	5.0/ 5.2	8.3	11.5
Crow	16	0.0/ 7.5	0.0/16.2	0.0/ 0.0	0.0	7.9
Omaha	33	36.4/48.3	21.2/32.4	18.2/22.2	25.3	34.3
Other-5 mean		18.5/28.1	5.2/13.6	5.6/ 7.5	10.6	18.9

Empty responses (truncation before the answer line) are 0/249 on Gemma3, 5/249 on DeepSeek-R1-32B, and 4/249 on Qwen3-32B, indicating that truncation does not account for the accuracy floor.

F Protocol Sensitivity

Table 14 reports Cat.4 accuracy by kinship system under zero-shot direct and few-shot CoT, supporting the claim in §5 that the typological ordering persists under enhanced prompting. Few-shot CoT raises Other-5 Cat.4 accuracy modestly on every system, but the descriptive controls (Esk&Sud) reach near-ceiling, so the system-type gap is preserved.

Table 14: Performance by kinship system under two protocols (Cat.4 Exact Match %, mean ± std across the three open-source models).

System	Type	Direct	Few-Shot
Eskimo	Descriptive	93.0 ± 3.3	99.6 ± 0.5
Sudanese	Descriptive	92.0 ± 5.2	99.5 ± 0.8
Hawaiian	Generational	61.9 ± 5.9	69.7 ± 1.9
Dravidian	Bifurcate	60.4 ± 13.6	63.6 ± 3.4
Iroquois	Bifurcate	50.3 ± 12.1	62.4 ± 7.7
Crow	Mat. Skewing	47.0 ± 12.8	50.0 ± 9.3
Omaha	Pat. Skewing	**32.6** ± 11.2	**44.4** ± 4.3

Per-model gap by protocol. Per-model breakdown of the overall (all-category) Esk&Sud → Other-5 gap across protocols, referenced from §5.1.

Table 15: Overall Esk&Sud → Other-5 accuracy gap by protocol (EM %, mean across all four categories). The gap persists under both protocols for all five models; few-shot CoT raises both sides similarly and shifts the gap by at most a few points in either direction.

Model	Protocol	Esk&Sud	Other 5	Gap
GPT-4o-mini	Direct	92.3	81.2	11.1
	Few-Shot	98.2	82.6	**15.5**
Claude-Haiku-4.5	Direct	98.3	85.3	13.0
	Few-Shot	97.6	87.9	**9.8**
Gemma3-27B	Direct	96.8	84.7	12.1
	Few-Shot	97.6	85.0	**12.6**
DeepSeek-R1	Direct	92.7	73.8	18.9
	Few-Shot	98.4	81.8	**16.6**
Qwen3-32B	Direct	92.2	77.9	14.3
	Few-Shot	98.2	83.1	**15.1**

G Fictional-Rule Control

To test whether the Cat.4 drop reflects a deficit of cultural *knowledge* or interference from learned cultural *priors*, we construct a relabelled variant of Cat.4 in which the system name and the cultural kin term are replaced with phonetically distant invented labels of identical structural shape. Hawaiian, Iroquois, Dravidian, Crow, and Omaha become Zorblax, Krendari, Glabran, Vespian, and Tarskan;

"classificatory mother/father/sibling/. . . " become "tier-mater/pater/sibling/. . . "; "cross-cousin" and "parallel cousin" become "off-axis cousin" and "on-axis cousin." The biological context, anchor, target persons, ground-truth biological relation, and reasoning depth are preserved exactly; only the cultural-naming surface is changed.

We evaluate the three open-source models on the relabelled Cat.4 questions under zero-shot direct prompting. Table 16 reports the per-system, per-model exact-match comparison.

Table 16: Cat.4 accuracy on real cultural labels vs. fictional isomorphic labels (Exact Match %, zero-shot direct). Fictional EM exceeds real EM in 11 of 15 cells; the average gain is +6.1% (95% bootstrap CI [+2.8, +9.4] over the 15 cells).

Model	System	Real	Fictional	Δ
Gemma3-27B	Hawaiian	68.9	64.4	−4.5
	Iroquois	66.7	72.0	+5.3
	Dravidian	79.5	83.3	+3.8
	Crow	63.6	59.1	−4.5
	Omaha	47.7	53.0	+5.3
DeepSeek-R1-32B	Hawaiian	62.1	61.4	−0.8
	Iroquois	37.9	46.2	**+8.3**
	Dravidian	51.5	65.2	**+13.6**
	Crow	32.6	30.3	−2.3
	Omaha	21.2	31.1	**+9.8**
Qwen3-32B	Hawaiian	54.5	61.4	+6.8
	Iroquois	46.2	62.9	**+16.7**
	Dravidian	50.0	63.6	**+13.6**
	Crow	44.7	53.8	**+9.1**
	Omaha	28.8	40.1	**+11.4**
Average		**50.4**	**56.5**	**+6.1**

Interpretation. If Cat.4 difficulty reflected a pure deficit of cultural *knowledge*, removing the cultural label should leave accuracy unchanged or reduce it (since any helpful prior from training would be lost). Instead, accuracy systematically *rises* when the labels are replaced with invented ones, with the strongest gains on the hardest systems (Iroquois, Dravidian, Omaha) and on the two reasoning-focused models (DeepSeek-R1-32B and Qwen3-32B). This pattern is consistent with a cultural-prior interference account, in which cultural-system labels preferentially activate a learned default (Eskimo-style classification) that competes with the in-context reclassification rule. The orthogonal decomposition below further rules out the simplest lexical/parser explanations. We note that several alternative accounts—altered answer-space expectations, reduced ambiguity in the fictional prompts, or differential parsing behavior around culturally-loaded words—are not directly excluded, and we therefore frame the result as evidence consistent with interference rather than a decisive isolation.

Threats. The fictional labels ('Zorblax', 'Vespian', etc.) may themselves carry weak associations (e.g., reading as fantasy or science fiction); we mitigate by sampling phonotactically distinct strings with no Wikipedia presence, but cannot rule out residual lexical effects. We address the lexical-channel question directly below via an orthogonal decomposition of the two surface swaps.

Orthogonal decomposition (detail). The summary 2×2 of the system-name vs. kin-term swap interaction appears as Table 7 in §5.2. We evaluate the three open-source models under three conditions: *system-only* (only system names swapped, kin terms real), *term-only* (only kin terms swapped, system name real), and *full-fictional* (both swapped). Numbers pool over the $5 \times 3 = 15$ (system, model) cells. The interaction is largest for Qwen3-32B (cell-level interactions of +16 to +23% on most systems); per-cell deltas are in the per-system per-model table above. The model's behaviour does not depend on either surface feature in isolation, but on the joint absence of a recognisable cultural anchor.

H In-Context Rule Probe: Full Per-Model Data

Table 17 reports per-(model, system) baseline and +rule Cat. 4 accuracy on the full Other-5 set (1,254 questions per model). Each row is one kinship system; columns give baseline (no rule) and +rule EM/F1 (%) for each of the three open-source models, plus the per-system mean ΔEM across models. Descriptive controls (Eskimo, Sudanese) are included as a noise-floor sanity check: their prompts are byte-identical between baseline and +rule runs (no rule is injected, since the rule dictionary covers only the Other-5), so any non-zero delta reflects run-to-run non-determinism (0–0% on Gemma3, ∼4% on DeepSeek, ∼7–8% on Qwen3 under temperature = 0 batched inference). All deltas reported in §5.2 exceed this noise floor.

Per-system replication detail. The typology × baseline summary appears as Table 8 in §5.2. Hawaiian is the only non-skewing system where rule supply HURTS on both models with baseline > 60% (Gemma3 −11.2, DeepSeek

Table 17: In-context-rule probe: per-(model, system) Cat. 4 EM / F1 (%) under baseline (no rule) vs. +rule. Three open-source models, zero-shot direct prompting, `max_tokens` = 8192. "Δ mean" is the cross-model average of ΔEM. Empty-response rates (truncation before answer): 0/249 Gemma3, 5/249 DeepSeek-R1-32B, 4/249 Qwen3-32B on the Other-5 cells.

System	Gemma3		DeepSeek		Qwen3		Δ mean
	base	+rule	base	+rule	base	+rule	EM%
Eskimo (desc. ctrl)	96.8/99.4	96.8/99.4	93.6/95.3	97.3/99.5	88.8/93.0	96.3/97.9	+3.7
Sudanese (desc. ctrl)	98.4/99.7	98.4/99.7	92.0/95.7	97.9/99.4	85.6/90.5	92.5/95.0	+4.3
Hawaiian (Gen.)	68.9/76.1	57.7/68.2	62.1/70.7	46.7/58.2	54.5/63.7	51.1/61.5	−10.0
Iroquois (Bif.)	66.7/78.1	57.7/71.1	37.9/53.4	46.7/58.7	46.2/62.4	52.7/65.8	+2.1
Dravidian (Bif.)	79.5/79.5	61.5/67.1	51.5/61.7	53.3/62.1	50.0/61.8	61.0/68.8	−1.7
Crow (Skewing)	63.6/70.0	75.6/78.9	32.6/45.8	46.4/57.2	44.7/56.7	50.6/61.5	+10.6
Omaha (Skewing)	47.7/63.5	75.3/79.6	21.2/35.1	40.4/52.0	28.8/41.1	51.8/61.0	+23.2
Other-5 mean	65.3	65.6	41.1	46.7	44.8	53.4	+4.9

−15.4) — a cross-model replication of the interference signature on a single Generational system. Per-cell deltas above the descriptive noise floor (max +7.5% on Qwen3) are exactly the HELPS/HURTS counts in Table 8.

Cross-benchmark dissociation. The baseline column of Table 17 also exhibits a Gemma3 > DeepSeek-R1-Distill-Qwen-32B ordering on Cat. 4 (Other-5 means 65.3% vs. 41.1%), which reverses the same models' published ordering on standard reasoning benchmarks: on MATH, GPQA, and LiveCodeBench, the reasoning-distilled DeepSeek-R1-Distill leads Gemma3-27B by roughly 5–28% (DeepSeek-AI et al., 2025; Team et al., 2025). This dissociation between reasoning-benchmark performance and cultural-rule-application accuracy is consistent with our positioning of Cat. 4 as a capability axis distinct from generic multi-hop reasoning.

Rule statements. The one-sentence override rule used per system is reproduced in Appendix C. Rules are derived from standard anthropological descriptions (Murdock, 1949; Trautmann, 1981); rule-variant ablations (paraphrasing, length, position) are not run here.

I Human Baseline

To establish a human ceiling for KinshipQA and to verify that residual model failures are not artefacts of question ambiguity, we ran a two-annotator human evaluation. We sampled 280 questions stratified by system × category (10 per cell, 7 systems × 4 categories) and asked two graduate-student annotators with no anthropological background to answer each question from the natural-language context alone. Annotators received the same one-sentence override-rule summary used in the in-context-rule probe (§5.2) for each Other-5 question; Esk&Sud questions were answered without rule supply. We report per-cell exact-match accuracy and inter-annotator agreement on item-level responses.

Table 18: Human baseline per system × category (% correct, 10 questions per cell, two annotators A1/A2). Inter-annotator agreement (IAA) on item-level responses is 96.8% over all 280 questions (per-category: Cat.1 97.1%, Cat.2 98.6%, Cat.3 97.1%, Cat.4 94.3%).

	Cat. 1		Cat. 2		Cat. 3		Cat. 4	
System	A1	A2	A1	A2	A1	A2	A1	A2
Eskimo	100	100	100	100	100	100	100	100
Sudanese	100	100	100	100	100	100	100	100
Hawaiian	100	100	100	100	100	100	100	100
Iroquois	100	100	100	100	100	100	100	100
Dravidian	100	100	90	100	100	100	90	90
Crow	90	100	100	90	90	80	70	80
Omaha	100	90	90	90	80	80	80	80
Mean	**98.6**		**97.1**		**95.0**		**90.7**	

Humans reach 95.4% across all 280 questions (annotator means 95.7% and 95.0%), with category-level accuracy decreasing along the same hierarchy that LLMs exhibit (Cat. 1: 98.6%; Cat. 2: 97.1%; Cat. 3: 95.0%; Cat. 4: 90.7%). On the cultural-override category (Cat. 4), human accuracy on the Other-5 systems averages 89.0% (annotator means 88% and 90%), compared to the open-source LLM mean of 50.4% (Table 5) — a 38.6% human–model gap. The per-system pattern within Cat. 4 also matches the typological ordering: humans saturate the descriptive controls and the generational/bifurcate systems where the override is structurally simple (100% on Eskimo, Sudanese, Hawaiian, Iroquois) and drop on the skewing systems (Crow 75%, Omaha 80%) and on the marriage-rule-constrained Dravidian (90%). The two skewing systems are also the lowest-scoring cells under inter-annotator agreement, consistent with their structural difficulty rather than question ambiguity. Multilingual native-speaker validation is left to future work.

J Frontier-scale Ceiling Check

To test whether the cultural-override pattern persists at frontier scale, we evaluate **Claude Opus 4.7** (`claude-opus-4-7`, Anthropic's flagship reasoning-augmented model) on a stratified Cat. 4 subsample: 100 questions per system (700 total) drawn with seed 42 across the same dataset used for the open-source evaluation, under the same zero-shot direct prompt template (Appendix C) and parser. Greedy decoding (temperature = 0) is used

where supported by the API.

Table 19: Frontier-scale Cat. 4 spot-check: Claude Opus 4.7 (zero-shot direct and few-shot CoT) vs. the three open-source models. Open-source numbers are per-system Cat. 4 EM under zero-shot direct from Table 14; the Opus 4.7 columns use 100 stratified Cat. 4 questions per system. The typological ordering of Table 5 is preserved at frontier scale; few-shot CoT narrows the cultural-override gap but does not close it.

System	Type	Open-src	Opus Dir.	Opus FS-CoT	Δ
Eskimo	Descriptive	93.0	**100.0**	**100.0**	+0
Sudanese	Descriptive	92.0	**100.0**	**100.0**	+0
Hawaiian	Generational	61.9	74.0	84.0	+10
Iroquois	Bifurcate	50.3	70.0	75.0	+5
Dravidian	Bifurcate	60.4	87.0	84.0	−3
Crow	Mat. Skewing	47.0	60.0	76.0	+16
Omaha	Pat. Skewing	32.6	54.0	**73.0**	+19
Esk&Sud avg.		92.5	**100.0**	**100.0**	+0
Other-5 avg.		50.4	69.0	78.4	+9.4
Cultural-override gap		**42.1**	31.0	**21.6**	−9.4

Three observations sharpen the paper's central claim. *First,* frontier-scale reasoning lifts Cat. 4 accuracy by $+18.6\%$ on average across the Other-5 systems and saturates the descriptive controls (Eskimo and Sudanese both reach 100%); the cultural-override gap narrows from 42.1% to 31.0% under zero-shot direct. *Second,* few-shot CoT at frontier scale further narrows the gap to 21.6%—about half of the open-source baseline closes when scale and demonstrations are combined—with the largest per-system gains on the hardest systems (Omaha $+19$, Crow $+16$). *Third,* the gap nevertheless persists: even Opus 4.7 with few-shot CoT misses 16–27% of Cat. 4 questions on the bifurcate, generational, and skewing systems, and the per-system relative ordering is fully preserved. The cultural-override phenomenon is therefore not a mid-tier-only artefact: it persists at frontier scale, is partially mitigated by combining scale with few-shot demonstrations, and shows a stable cross-system structure.

Two-factors-compound pattern at frontier scale. Binning the Opus 4.7 spot-check by reasoning depth (Table 20) shows the cultural-override gap also *widens with hop count* at frontier scale—from 22.4% at 2-hop to 47.6% at 3-hop, a +25.2% widening. The open-source models widen by a comparable amount in absolute terms (+17.4% from 37.1→54.5%), indicating that the depth × override interaction documented in §5.1 for the open-source models is not eliminated by frontier-scale reasoning capacity.

Table 20: Cat. 4 cultural-override gap by reasoning depth, Opus 4.7 vs. the three open-source models. The gap widens with hop in both, with similar absolute widening ($\sim$17–25%).

Cohort	Hop	Esk&Sud	Other-5	Gap
Opus 4.7	2	100.0	77.6	22.4
	3	100.0	52.4	**47.6**
Open-source (3-model avg)	2	95.7	58.6	37.1
	3	96.7	42.2	**54.5**

K Paraphrase Robustness Probe

To test whether the cultural-override gap reflects sensitivity to the specific question template used in the headline benchmark (a possible "phrasing artefact" explanation), we re-run Gemma3-27B (the highest-accuracy open-source model on Cat. 4) on a 50-question Cat. 4 subsample (10 per Other-5 system, seed 20260523, stratified across 2-hop and 3-hop) with the question text replaced by a rule-paraphrased variant.

Paraphrase construction. Paraphrases preserve entity names, the system identifier, and the kin term verbatim, and vary only the surrounding framing (Table 21). Patterns cover the four most common Cat. 4 question shapes in the dataset: forward-singular ("In the X kinship system, who is A's KT?"), forward-plural ("In the X kinship system, who are A's KTs?"), forward-list ("List all of A's KTs according to the X system"), and reverse-mapping ("In the X system, A is called B's 'KT'. What is their actual biological relationship?"). The context paragraph (NL serialisation of the proof subgraph) is held identical to the original.

Table 21: Example paraphrases (one per pattern). All four templates preserve entity names, system label, and kin term verbatim.

Original	Paraphrase
In the Hawaiian kinship system, who is Lisa W.'s classificatory mother?	Under the Hawaiian kinship convention, identify Lisa W.'s classificatory mother.
In the Iroquois kinship system, who are M.W.'s cross-cousins?	Under the Iroquois kinship convention, enumerate M.W.'s cross-cousins.
List all of P.J.'s classificatory siblings according to the Hawaiian system.	Within the Hawaiian system, give the complete list of P.J.'s classificatory siblings.
In the Hawaiian system, M.W. is called D.J.'s 'classificatory mother'. What is their actual biological relationship?	Per Hawaiian kinship conventions, M.W. holds the role of 'classificatory mother' relative to D.J. State their actual biological relationship.

Results. Table 22 reports Gemma3-27B EM on the 50 originals (anchor) and on the paraphrased variants. Aggregate EM rises from 78.0% to 84.0% (+6.0%); F1 rises from 87.1% to 88.3% (+1.2%). Five questions flip from incorrect to correct under paraphrase and two flip the other way, leaving net +3 correct out of 50. Per-system, four of five systems are stable or improve (Hawaiian, Iroquois +10%; Omaha +20%; Dravidian unchanged at 100%); Crow drops 10% (one question, $n = 10$). The direction of change is the opposite of what a brittle template-artefact account would predict; the absence of degradation is consistent with the Cat. 4 cultural-override gap reflecting a cultural-knowledge axis rather than dependence on the specific phrasing used in our headline benchmark.

Table 22: Paraphrase robustness on Gemma3-27B (zero-shot direct, $n = 10$ per system, total $n = 50$). EM is computed against the same gold answers used in the headline runs. Net flips = (orig-wrong, para-right) minus (orig-right, para-wrong).

System	Orig EM	Para EM	Δ
Hawaiian	90.0	100.0	+10.0
Iroquois	90.0	100.0	+10.0
Dravidian	100.0	100.0	+0.0
Crow	70.0	60.0	−10.0
Omaha	40.0	60.0	+20.0
Other-5 avg.	**78.0**	**84.0**	+6.0
Net flips	+5 wrong→right, −2 right→wrong		

Caveats. The probe covers a single open-source model ($n = 1$) on a single stratified subsample ($n = 50$); the +6.0% aggregate is within sampling noise for $n = 50$, and the per-system entries should not be over-interpreted at $n = 10$ per cell. The probe is a controlled-direction stress test against the template-artefact hypothesis, not a broad robustness study. Paraphrase generation is rule-based via four templates (one per question pattern), which avoids LLM-paraphraser contamination but limits the surface diversity probed. Full paraphrase set and per-question predictions are released alongside the dataset.