

Schema-Guided Extraction of Interaction Knowledge Bases from Narrative Text: A Folktale Case Study

Sergey Kovalchuk^{1,2}, Anastasia Ugleva³, Oleg Kubryak^{1,4}

¹Sirius University of Science and Technology, Sirius Federal Territory, Russia,

²ITMO University, Saint Petersburg, Russia, ³Higher School of Economics, Moscow, Russia,

⁴National Research University “Moscow Power Engineering Institute”, Moscow, Russia

Correspondence: kovalchuk@itmo.ru

Abstract

We present MAGIC, a six-dimensional framework for extracting structured knowledge bases of hero and magic entity interactions from narrative texts of folktales and fairytales: each interaction between a story’s protagonist and a magical helper, villain, or object is decomposed along trigger modality, actor profile and agency balance, interaction mechanics, outcome valence, narrative function, and narrative position. We build an LLM-based extraction pipeline and apply it to a corpus of 1306 folktales collected from around the world, then show the schema supports rich structural analysis: unsupervised clustering over the extracted features recovers five recurring, readily nameable interaction archetypes – invoked and unsolicited helpers, a morally tested gatekeeper, a hero-controlled magic object, and an adversarial obstacle – that are stable across the three full-corpus extraction models (0.95 mean cluster-centroid similarity) and cut across both tale type and regional tradition rather than reproducing existing typological catalogues. We further characterize the schema’s reliability under LLM extraction (inter-model, not against a human gold standard): parse success and schema-conformance rates, a decomposition of extraction drift into field-confusion versus vocabulary-incompleteness failure modes, and cross-model field-level agreement, which varies substantially across the six dimensions and is sensitive to how many tales are compared across models. Despite this field-level variability, the archetype-level structure remains stable, suggesting that knowledge bases built from LLM extraction should be evaluated at the level of the downstream structure they support.

1 Introduction

Building a knowledge base from unstructured text requires a schema precise enough to be extracted consistently, yet expressive enough to capture what matters about the source material. This tension

is easier to manage when the target schema stays close to the surface form of the text (entities, relations, dates) – not because such extraction is free of interpretive judgment, but because its judgments are usually anchored to an explicit textual mention – and harder when the schema requires inference with no such anchor: who holds power in an exchange, or what narrative role an event plays, rather than simply what was stated. Recent work on LLM-driven knowledge graph construction often faces this tension. For example, [Shrimal et al. \(2025\)](#) show that static, hand-specified extraction schemas produce unreliable output and propose optimizing the schema itself against extraction performance, with [Han and Lai \(2026\)](#) arguing that symbolic, closed-vocabulary relation schemas lose exactly the contextual nuance an LLM could otherwise represent in natural language.

Narrative text is close to a worst case for this problem: its structure is implicit, culturally variable, and has been the subject of a century of disagreement among human narratologists about how to formalize it at all. Within narrative text, folktales and fairytales are a particularly attractive target for knowledge base construction. Transmitted orally across generations and borders, they act as a condensed carrier of cultural knowledge: recurring plots and character configurations encode moral norms and archetypal patterns of human behavior that remain remarkably stable across regions and centuries ([Lüthi, 1986](#); [Uther, 2004](#); [da Silva and Tehrani, 2016](#)). This stability is precisely why folktales attracted the earliest systematic formalization efforts in narratology ([Propp, 1968](#); [Greimas, 1983](#)) mainly with expert-designed schemas for narrative structure, hand-applied decades before machine extraction was conceivable. Among the structures these formalisms identify, the encounter between a hero and a magic-bearing entity (a helper, a donor, a villain, an enchanted object) is where much of a tale’s “wisdom content” is staged: it is the lo-

cus of moral testing, conditional aid, asymmetric agency, and reward, and it recurs across otherwise unrelated traditions. This makes the hero–magic interaction both a natural unit for an interaction-level knowledge base and a demanding testbed for schema-guided LLM-based extraction.

We introduce MAGIC (a recursive acronym for MAGIC as Actor-Governed Interaction Cycles), a framework with a six-dimensional schema for one recurring narrative structure focused on the interaction between a story’s protagonist and a non-protagonist actor or object with unusual capabilities (a magic helper, a villain, an enchanted item) that often appear in folktales and fairytales. Unlike character-centered frameworks that describe the protagonist’s overall arc (Campbell, 1949) or fixed-order plot grammars (Propp, 1968), MAGIC treats each such interaction as the unit of analysis, decomposed along six orthogonal dimensions: trigger modality (D1), actor profile and agency balance (D2), interaction mechanics (D3), outcome valence (D4), narrative function (D5), and narrative position (D6). This yields an interaction-level knowledge base rather than a single story-level label, dense enough to support downstream analysis – clustering, cross-corpus comparison – that a coarser schema would not.

Why the interaction, rather than the story or the character? The established formalisms attach their labels at coarser units: ATU (Aarne–Thompson–Uther) assigns one type per tale (Uther, 2004), the monomyth one arc per hero (Campbell, 1949), and character-network corpora one edge per co-occurring pair (Schmidt et al., 2021). Each is well suited to the question it was built to answer, and none of them can express, for a given exchange, what set it in motion, what it cost, on what condition the aid was granted, and what the hero was left holding afterwards which is where a tale’s normative content actually sits. A single tale stages several such exchanges with different answers to those questions, so a story-level label cannot be refined into them after the fact; and a character-level edge discards the conditionality that separates a gift from a bargain. We therefore add a layer beneath the existing typologies rather than proposing a replacement for them, and the evaluation below is correspondingly about whether that layer can be populated repeatably, not about whether it classifies tales better than ATU does.

To evaluate the schema, we extract interactions from 1306 folktales from a publicly available cor-

pus (Kaggle, 2024) with five LLMs of varying scale and provenance (DeepSeek V3.2, Gemini 2.5 Flash, Qwen 3.6 35B, GPT-5.2, GPT-OSS-120B): what structure does the schema reveal, and how reliably can that structure be extracted? For the first, unsupervised clustering over the extracted features recovers five recurring, readily nameable interaction archetypes, stable across the three full-corpus models and cutting across both tale type and regional tradition. For the second, we measure extraction reliability (parse and schema-conformance rates) and cross-model field-level agreement (Fleiss’ κ (Fleiss, 1971)) in two conditions (all five models on a shared 50-tale subset; three full-corpus models on all 1306 tales in the corpus). The central empirical finding connects the two: the archetype-level structure is highly stable even though no individual dimension is reliably extracted in isolation, suggesting individual field values may be less trustworthy than the aggregate structure they compose – a distinction current single-model evaluations do not surface.

Our contributions are: (1) the MAGIC schema and a reproducible, multi-model extraction pipeline that produces schema-conformant interaction knowledge bases from narrative text at corpus scale; (2) an unsupervised taxonomy of hero–magic interaction archetypes, shown to be stable across extraction models and orthogonal to existing tale-type and regional classifications despite weak field-level agreement; and (3) a reliability and cross-model agreement evaluation, run at two sample sizes, that identifies which parts of a narrative-interaction schema are robustly extractable and which are not, including a decomposition of schema drift into *field confusion* versus *vocabulary-incompleteness* failure modes, and a demonstration that fine-grained field-reliability rankings from a few dozen shared items are considerably less stable than the coarser conclusions they are usually read alongside.

2 The MAGIC Schema

The MAGIC framework analyzes each interaction between a story’s hero and a magic-bearing entity – a helper, villain, or object – along six orthogonal dimensions. Each dimension decomposes into multiple categorical or ordinal attributes, for 23 interaction-level attributes in total, plus two story-level fields (hero name, hero nature) that are constant across a tale’s interactions.

2.1 Schema design

The six dimensions come from two sources borrowing from existing narratological descriptions and revision against pilot extractions.

Three dimensions have direct antecedents. D5 reuses Propp’s function vocabulary (Propp, 1968) in condensed form, restricted to the seven functions that a hero-magic exchange can realize rather than the full inventory of 31, most of which describe plot moves with no magic-bearing counterpart. D2 takes from Greimas’s actantial model (Greimas, 1983) the separation between what an actor *is* in the story world and what it *does* in a given episode; in our schema this becomes the actor_type/actor_functional_role pair. D6’s story-arc values follow the departure-to-return progression common to Propp (1968) and Campbell (1949).

The other three dimensions have no direct antecedent, and for a structural reason: the unit those formalisms describe is larger than ours. Once the interaction rather than the tale is the coded object, three questions become answerable that a story-level schema cannot pose – who set this particular exchange in motion (D1), on what operational terms the magic was available (D3), and what the hero concretely gained and gave up (D4). These are the dimensions along which a helper’s gift differs from a donor’s bargain and from a gatekeeper’s test, distinctions that a coarser encoding collapses.

Two rounds of pilot extraction then revised the vocabulary, and both revisions are visible in the version reported here. A binary requires_payment field could not distinguish a material price from a demanded virtue and was replaced by the cost_structure/cost_timing/moral_condition triple; a single magic_nature field conflated where a power came from with what made it fire, and was split into magic_source and activation_condition. In both cases the trigger for revision was extractions that were fully schema-conformant yet visibly lossy against the source text – the schema had no way to record a distinction the tale plainly made.

Here, we use “orthogonal” the term in a specific and modest sense: fixing a value on one dimension does not determine a value on any other, so the dimensions are free to vary independently. The clearest case is D1 against D2, where the prompt states as an explicit rule that a villain (a D2 value) causing an inter-

action yields trigger_source=external_event and not a villain-specific trigger; the same holds within D2, where a villain may take the functional role trickster_helper. We do not claim statistical independence in the extracted data – helper actors and zero-cost interactions plainly co-occur, and Section 4.3 is in part a description of which combinations are common. The claim is only that the schema does not force those co-occurrences by construction.

It’s worth mentioning that six dimensions is a working decomposition, not a proof of sufficiency. It is adequate in the sense that we did not need a residual or “other” dimension to code the interactions encountered in this corpus and dense enough to support the analysis in Section 4.3. Dimensions we considered and left out (like affective register, the hero’s stated motivation, repetition structure across parallel episodes) are candidates for extension rather than evidence against the present set.

The extraction prompt elicits one further field, interaction_actions, which lists the observable actions performed during the exchange. It is an auxiliary extraction field rather than a MAGIC attribute. We consider it for further diagnostics and analysis and do not include it in the current experimental analysis.

2.2 Schema description

D1: Trigger Modality (2 attributes) separates *who* caused an interaction to begin (trigger_source: hero agency, actor agency, or an external event) from *how* it was expressed (trigger_mode: request, discovery, autonomous action, imposition, or crisis), so that, e.g., a villain-imposed trigger is not conflated with the villain’s identity.

D2: Actor Profile & Agency (7 attributes) captures the actor’s stable structural type (actor_type: helper, villain, or object), its functional role in this specific interaction (actor_functional_role: ally, donor, trickster-helper, tester, obstacle, or neutral – deliberately allowed to diverge from structural type, e.g. a villain who tests then rewards the hero), its physical/ontological nature, its observed autonomy, and the balance of control between hero and actor.

D3: Magic Mechanics (6 attributes) covers the operational constraints on the magic itself: magic_source (inherent, bestowed, environmental, stolen, or transferred), activation_condition (what makes it fire), cost_structure and cost_timing (what the

Dim.	Value
D1 Trigger	hero_agency · request
D2 Actor	helper (animal), donor
D2 Autonomy	medium · balanced
D3 Magic	inherent · moral_test
D3 Cost	material, upfront · kindness
D3 Duration	single_use
D4 Outcome	test_passed
D4 Gain / Cost	object + information / material
D5 Function	NF-2 (Provision)
D6 Position	threshold; follows [1], precedes [4,5]
Actions	gives_object, gives_advice

Table 1: Worked example: the “black cat” interaction from *Baba Yaga and the Orphans*, showing how one interaction event populates all six MAGIC dimensions.

hero pays and when), *moral_condition* (which virtue is demanded, where the cost is behavioral), and duration (how long the effect persists).

D4: Outcome Valence (3 attributes) separates the primary narrative result (success, qualified success, a passed test, failure, or an unintended consequence) from what the hero concretely gains and loses, each a multi-value field, so that a productive-looking interaction with a hidden cost is not collapsed into a single valence label.

D5: Narrative Function (1 attribute) assigns one or more of seven Proppian-derived function codes (NF-1 Reconnaissance through NF-7 Obstacle) capturing the interaction’s role in the plot.

D6: Narrative Position (4 attributes) situates the interaction in the story arc (departure, journey, threshold, climax, resolution) and records direct causal dependencies (precedes/follows) on other interactions in the same tale, encoding a lightweight causal chain rather than only story order.

A complete definition of attributes of six dimensions is provided in Appendix A.

An optional *actor network* extension records explicit relationships between non-hero actors (e.g., a witch who controls an enchanted object later given to the hero), with fields for relationship type, structural role (mediation, subordination, opposition, cooperation, or provision chain), direction, and chain depth to the hero. We describe it because it is part of the released knowledge base, but it is deliberately outside the scope of the evaluation: none of the three experiments in Section 4 operates on it, and no claim in this paper depends on it. A complete definition of actor network is provided in Appendix B.

Table 1 shows one interaction coded against the

full schema extracted for *Baba Yaga and the Orphans*¹: a black cat, requested for help by two orphans, tests them morally before rewarding correct behavior with an object and information at the cost of a material payment – a single event that would collapse into an undifferentiated “helper gives item” edge in a coarser schema, but here carries an explicit trigger, cost structure, moral condition, and position in the tale’s causal chain. The same tale’s non-hero actors are also linked into an actor network (Appendix B, Figure 2): the orphans receive help from three animal donors and use two magic objects to escape a villain.

3 Extraction Pipeline

We extract from a corpus of 1306 folktales spanning many cultural traditions and regions, with a maximum concentration from German-language sources, including the Brothers Grimm collection (Kaggle, 2024). Each tale is passed once to an LLM with a single prompt (see Appendix C) specifying all 25 fields, their closed vocabularies, and worked examples of ambiguous distinctions (e.g., *actor_type* vs. *actor_functional_role*); the model returns one JSON object per tale, an interaction list plus the ATU tale-type index (Uther, 2004). We ran this pipeline with three models: DeepSeek V3.2, Gemini 2.5 Flash, and Qwen 3.6 35B over the full corpus. Additionally, we ran GPT-5.2 and GPT-OSS-120B over a fixed 50-tale subset shared by all five models for extended cross-model comparison. The 50 tales were drawn uniformly at random from the corpus with a fixed seed before any extraction was run, and the same 50 were used for all five models. The pipeline is resumable and validates each response against the schema’s vocabulary before accepting it (Section 4.1).

4 Experimental evaluation

4.1 Extraction Conformance and Schema Drift

For each model, we measure whether its output parses as JSON at all, and, among parsed interactions, whether every field is present and every value falls within the schema’s declared vocabulary (Table 2). Parse success is high across the board (90–99%), but schema conformance – whether the *values* inside a successfully parsed interaction obey the schema – ranges from 80.6% (GPT-OSS-120B)

¹Here and further we refer to folktales’ titles giving examples for narrative processing as they were given in the dataset.

Model	Parse	Conform.	Int./tale
DeepSeek V3.2	96.7%	93.9%	5.10
Gemini 2.5 Flash	98.3%	88.3%	6.54
Qwen 3.6 35B	99.0%	90.0%	4.42
GPT-5.2 [†]	98.0%	99.2%	5.43
GPT-OSS-120B [†]	90.0%	80.6%	4.00

Table 2: Extraction reliability by model. Parse = fraction of outputs parsing as JSON; Conform. = fraction of interactions with zero missing or out-of-vocabulary fields; Int./tale = mean interactions extracted per tale. [†]: 50-tale subset.

Model	Drift	Confusion	%
DeepSeek V3.2	424	244	57.5%
Gemini 2.5 Flash	1057	986	93.3%
Qwen 3.6 35B	617	576	93.4%
GPT-5.2 [†]	2	2	100.0%
GPT-OSS-120B [†]	41	14	34.1%

Table 3: Schema drift decomposed into field confusion vs. novel values. Drift = total out-of-vocabulary values; Confusion = count valid for a different field; % = confusion share. [†]: 50-tale subset.

to 99.2% (GPT-5.2, though only measured on the smaller 50-tale subset), with the three full-corpus models spanning 88.3–93.9%. Mean interactions extracted per tale also vary substantially (4.42–6.54, for models run on the whole corpus), a first sign that models do not agree on the granularity at which a tale’s magic interactions should be segmented, a point we return to in Section 4.2.

Field confusion vs. vocabulary incompleteness.

We further classify every out-of-vocabulary value as *field confusion* (the value is valid for a *different* MAGIC field – e.g. duration=always_active, a valid activation_condition value) or a *novel value* (not in any field’s vocabulary at all – e.g. hero_gain=transportation), by checking every drifted value against the full cross-field vocabulary index.

This split (Table 3) is strongly model-dependent: field confusion accounts for 93.3% and 93.4% of drift for Gemini and Qwen respectively, but only 57.5% for DeepSeek and 34.1% for GPT-OSS-120B, meaning roughly half to two-thirds of these two models’ drift is genuinely novel, invented values rather than confusion with an existing field – a pattern that does not track model scale in any simple way. Critically, the two failure modes are not evenly spread across fields. Novel-value drift concentrates almost entirely

Field	κ_{50}	κ_{1306}	Δ
actor_type (D2)	0.44	0.45	+0.01
outcome_type (D4)	0.25	0.35	+0.11
cost_structure (D3)	0.15	0.33	+0.18
actor_functional_role (D2)	0.28	0.32	+0.04
trigger_source (D1)	0.32	0.32	−0.01
narrative_function (D5)	0.29	0.31	+0.03
story_phase (D6)	0.19	0.27	+0.08

Table 4: Fleiss’ κ for one representative field per MAGIC dimension, in two independent evaluation conditions: κ_{50} – 5 models, 50-tale shared subset; κ_{1306} – the 3 full-corpus models only, all 1,306 tales. $\Delta = \kappa_{1306} - \kappa_{50}$. Rows ordered by κ_{1306} .

in interaction_actions and moral_condition across *all five* models – models keep inventing plausible action verbs (gives_information, alliance_formation, provides_remedy) and moral conditions (obedience, loyalty, faith, purity) beyond the schema’s fixed lists, suggesting these two fields need either an open vocabulary or a substantially expanded closed one. Field-confusion drift, by contrast, concentrates in a specific, recurring set of field pairs: magic_source receiving none (not a valid value for that field, but the default for several cost-related fields, and the single largest drift category overall at 410 occurrences for Gemini alone), duration receiving activation_condition values, actor_type receiving actor_nature values (e.g. animal, spirit), and trigger_source receiving trigger_mode values (e.g. crisis) – boundaries between specific fields that are evidently ambiguous to extractors independent of which model is extracting.

4.2 Cross-Model Field Agreement

Interaction segmentation differs across models (Table 2), so interaction-to-interaction alignment across models is unreliable without itself imposing a modeling assumption. We instead compare models at the *tale* level, using their agreement as a lens on schema reliability rather than as a target metric in its own right: for each tale, we take each model’s dominant (most frequent) value per field across the interactions it found, then measure multi-rater agreement with Fleiss’ κ (Fleiss, 1971), computed directly from category counts, since the NO_INTERACTIONS vote described below needs a rater-by-category matrix that packaged implementations do not accept, and cross-validated against the statsmodels reference where both ap-

ply. James (2026) note that metric choice materially affects reported reliability in NLP annotation; we chose Fleiss’ κ specifically because it corrects for chance agreement, which matters here (see below). A model that extracted zero interactions from a tale is treated as voting NO_INTERACTIONS for every field, since whether a model judges a tale to contain any magic interactions at all is itself part of what we measure; this also folds in a small number of outright parse failures, which is a limitation we return to in Section 6.

We report this in *two* conditions rather than one, because a single 50-tale estimate turned out not to be trustworthy on its own. κ_{50} uses all five models on the 50-tale subset GPT-5.2 and GPT-OSS-120B were restricted to; κ_{1306} uses only the three full-corpus models (DeepSeek, Gemini, Qwen) but all 1,306 tales – trading two raters for a $26\times$ larger item count. Table 4 shows both.

No field reaches better than moderate agreement in either condition, and *actor_type* is the single highest-agreement field in both ($\kappa_{50} = 0.44$, $\kappa_{1306} = 0.45$) – these two conclusions are robust to sample size and rater set, and together form the paper’s central reliability finding. Beyond that, however, the picture that motivated our first attempt at this section – a clean, monotonic gradient from “observable” to “interpretive” fields – does not survive the larger sample. *cost_structure* moves from the *worst*-agreement field at $n=50$ ($\kappa_{50} = 0.15$) to solidly mid-table at $n=1,306$ ($\kappa_{1306} = 0.33$, $\Delta = +0.18$, the largest shift of any field), and at $n=1,306$ six of the seven fields cluster in the “fair” band (0.27–0.35) with only *actor_type* clearing into “moderate” – none falls to “slight,” unlike at $n=50$. The Spearman rank correlation between the two conditions’ field orderings is only 0.21 (mean $|\Delta| = 0.06$, max $|\Delta| = 0.18$): the fine-grained ranking of the five middle fields is not a stable property of the schema at $n=50$, only the two extremes are (*actor_type* highest in both; *story_phase* lowest or second-lowest in both, $\kappa_{50} = 0.19$, $\kappa_{1306} = 0.27$). We read this as a caution that generalizes beyond this paper: cross-model κ estimates over few dozen shared items – a scale forced on us here by API cost, and not unusual in similar reliability studies – can rank-order schema fields in ways that a substantially larger sample overturns, even when the sample size is enough to establish the coarser, headline conclusion reliably.

Possible mislead in agreement metrics. In both

conditions, *cost_structure* shows a large gap between κ and raw agreement: at $n=50$, 80% of tales have ≥ 4 -of-5 models agreeing on the dominant value despite $\kappa_{50} = 0.15$; at $n=1,306$, 67.5% of tales have all 3 models agreeing despite $\kappa_{1306} = 0.33$. A possible explanation can be that *cost_structure=none* is an overwhelming majority class (84.8% of tale-level votes at $n=50$), so models trivially “agree” on none near the rate the base rate alone would produce – exactly what a chance-corrected statistic discounts and raw agreement does not. *story_phase*, by contrast, has both low raw agreement (28% at $n=50$) and low κ : its five values are far more evenly used (journey 44%, departure 20%, climax/threshold/resolution splitting the rest), so its low κ reflects genuine disagreement rather than a base-rate artifact. Reporting only raw agreement would have overstated *cost_structure* as reliable in both conditions – a distinction visible only once a chance-corrected and a raw statistic are reported side by side, independent of which of the two field-ranking pictures above one trusts.

A secondary, weaker signal points the same direction as the rest of this section: the mean pairwise Spearman correlation of per-tale interaction counts is 0.573 across all five models at $n=50$ but rises to 0.747 across the three full-corpus models at $n=1,306$ – another instance of a cross-model estimate looking noisier at the smaller sample size. Interaction-count means still differ substantially at full scale (4.4 for Qwen to 6.4 for Gemini per tale) – models do not agree on *how much* is happening in a given tale, before any question of how to characterize it arises, and this disagreement is not merely small-sample noise.

Four further checks (see Appendix D) support this reading. A bootstrap CI and a same-model-set control together confirm the *cost_structure* shift is a genuine sample-size effect, not resampling noise or which models happen to be included. Pairwise Cohen’s κ across all model pairs shows disagreement broadly distributed rather than concentrated on one model, ruling out a single idiosyncratic rater as the explanation for weak aggregate agreement. A same-pipeline calibration field (*hero_nature*, nearly unambiguous from text) reaches $\kappa = 0.64$ –0.79, well above every MAGIC dimension at matching sample size. These dimensions are specifically hard to determine from text, not uniformly hard to agree on under this pipeline. Finally, a worked disagreement case

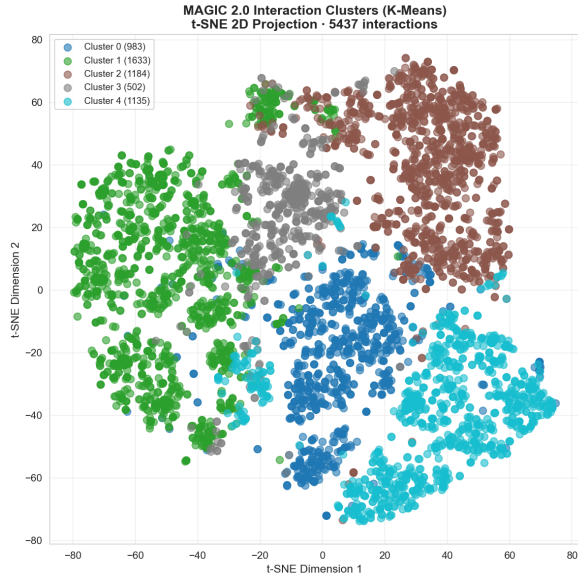


Figure 1: t-SNE projection of the five-cluster taxonomy discovered in the extraction.

shows some field-level “disagreement” is a segmentation artifact: models can agree on every interaction’s content and still disagree on the tale-level summary if they extract different numbers of interactions overall.

4.3 Taxonomy Induction and Cross-Model Stability

Beyond individual-field agreement (Section 4.2), we ask a complementary question: does a higher-level structure (as an unsupervised taxonomy of interaction types) emerge from the extracted features, and is it stable across models? We build a 93-dimensional feature vector per interaction (one-hot categorical fields, ordinal fields normalized to $[0, 1]$, multi-value fields binarized) and run k -means. The optimal number of clusters $k=5$ was identified with the elbow method from $k \in [2, 10]$ using DeepSeek extraction and was fixed for cross-model comparability. Clustering was run independently on each of the three full-corpus extractions, excluding villain interactions (5437 / 6620 / 4097 interactions for DeepSeek, Gemini, and Qwen respectively). Figure 1 shows the resulting projection for DeepSeek.

The description that follows is a coherence check rather than evidence of the schema’s superiority: clusters induced from MAGIC features are necessarily describable in MAGIC terms. What is not settled in advance is *which* dimensions organize the largest clusters. The two distinctions the schema draws most deliberately (D1’s trigger modality and D2’s split of structural type from functional role)

Archetype	N	Signature
Invoked helper	983	hero-requested (76%), helper (97%), NF-1/2/4 balanced
Bestowed object	1633	hero-controlled (91%), no cost (91%), NF-2, success (92%)
Moral-test gate	1184	moral cost (63%), NF-2/NF-3 mixed, mixed outcome
Object obstacle	502	actor-dominant (64%), NF-7 (57%)
Unsolicited helper	1135	actor-initiated (98%), no cost (99%), NF-1

Table 5: The five interaction archetypes discovered in the extraction.

could have been swamped by the other four; that they are instead the separating axes indicates they carry real variance in the corpus. A stronger test, clustering the same interactions under an alternative encoding (Proppian functions alone, or text embeddings), we leave to future work.

The five archetypes (Table 5) are readily nameable, recur across the corpus, and separate two schema dimensions that a coarser taxonomy would conflate. *Invoked helper* (e.g. *The Six Servants*) and *unsolicited helper* (*The Journey to the House of Herr Korbes*) share almost every attribute (helper actor, no cost, success) and differ essentially only in D1: whether the hero asked (76% hero-initiated) or the actor acted first (98% actor-initiated, e.g. a talking animal that volunteers advice before being asked) – exactly the trigger-modality distinction Section 2 introduces to avoid conflating an actor’s identity with how an exchange began. *Bestowed object* (*Jack the Giant-Killer*) and *object obstacle* (*The Stolen Moon*) make the same separation on D2’s actor-functional-role axis: both are dominated by actor_type=object, but one is almost purely helpful (91% hero-controlled, 92% success) while the other is almost purely adversarial (57% NF-7 Obstacle, actor- rather than hero-dominant). The same structural category (a non-sentient magic item) plays opposite functional roles depending on the interaction, which is exactly why D2 keeps actor_type and actor_functional_role as separate fields rather than one. *Moral-test gate* (*The Three Citrons*) is the one archetype defined by a reciprocal exchange rather than a one-way gift or obstacle: an actor (helper or object) withholds aid until the hero demonstrates a prescribed virtue, mixing NF-2 (Provision) and NF-3 (Threshold Test) roughly evenly and producing one of the least con-

centrated outcome distributions of the five (44% success, second only to *object obstacle*’s 34%; the rest split across qualified success, failure, and unintended consequence). The schema’s cost/moral-condition fields exist specifically to capture this kind of conditional aid, which a simpler success/failure outcome label would collapse into ordinary success. We match clusters across the three models via Hungarian assignment on cosine distance between centroids as the underlying one-hot feature space is defined by the fixed schema vocabulary, not by which model produced the data, so centroids from different models occupy the same coordinate space without any additional alignment step. We take centroid similarity as the primary criterion because what should be stable is the *content* of each archetype, which its centroid directly expresses. Partition-agreement measures (ARI, AMI) do not apply here: they compare two labelings of one item set, and the three extractions share no item set, since each model segments a tale into a different number of interactions (Section 4.2). An internal measure such as silhouette does apply, and would report separation within each partition rather than similarity across them; we report only the latter. The result is a mean matched-cluster cosine similarity of 0.953 across all three model pairs (DeepSeek–Gemini 0.964, DeepSeek–Qwen 0.944, Gemini–Qwen 0.949; minimum over all 15 individual matched-cluster pairs 0.858), with matched clusters sharing both their dominant attributes and, qualitatively, their example tales.

Additionally, we conducted analysis of the tale type and region (see Appendix E). Each archetype draws examples from a similar spread of ATU tale types – ATU 425 (*The Search for the Lost Husband*) is the single most frequent type in all five clusters (24–92 occurrences), and four more types (ATU 313, present in all five clusters’ top five; ATU 300 and 750, each in three of five; ATU 301 and 480, each in two of five) recur across multiple archetypes rather than separating them; only a handful of types (e.g. ATU 563, 513, 470) are top-five for exactly one archetype. Region follows a similar pattern: German- and Scottish-sourced tales are the top two regions in every one of the five archetypes, so no archetype is a proxy for a single regional tradition. This indicates the MAGIC taxonomy captures a structurally distinct axis of variation (what kind of exchange this is) that cuts across tale type and regional tradition (which story this is and where it is from). Given this, we see

the schema intended to complement the existing typological literature rather than replace or rename it (Section 5).

4.4 Reconciling Experiments

The taxonomy is stable at the level of *interaction archetypes* even though individual fields disagree at the level of *individual interactions*. We read this as evidence that most of the cross-model disagreement in Section 4.2 is about exactly where to draw a boundary within a shared underlying structure – e.g., precisely which *story_phase* an interaction falls into – rather than about whether that structure exists at all. For a knowledge base built this way, this suggests that aggregate, downstream structure derived from LLM extraction may be considerably more trustworthy than any single extracted field value, a distinction that reliability evaluations reporting only field-level or triple-level accuracy do not capture.

5 Related Work

Structural narrative theory. Propp’s fixed-order function grammar (Propp, 1968) and Greimas’s actantial model (Greimas, 1983) remain foundational accounts of folktale structure, and Campbell’s monomyth (Campbell, 1949) traces the hero’s arc at the story level. All three center the hero’s overall trajectory or the plot’s fixed sequence; MAGIC instead treats each hero–actor interaction as a bounded, independently-coded event, closer in spirit to a knowledge-base schema than to a narrative grammar, while reusing Propp’s function vocabulary as one of its six dimensions (D5).

Folktale and narrative corpora. The ATU index (Uther, 2004) remains the standard typological catalogue; recent annotated corpora include the multilingual Bag-of-Tales (Hagedorn and Darányi, 2022), the German-language FairyNet character-network corpus (Schmidt et al., 2021), and a syntactically/semantically annotated French fairy-tale corpus (El Maarouf and Villaneau, 2012). These resources are richly annotated, but largely at the story or character level. MAGIC’s interaction-level, cross-culturally scoped schema therefore extends this line of work rather than competing with it. Most directly related is concurrent work on narrative-centric knowledge structures for long-form text understanding (Tian et al., 2026), which builds entity- and storyline-level graphs to support retrieval-augmented reasoning over narrative text;

our schema operates at a finer grain (the individual interaction) and is evaluated for extraction reliability rather than downstream QA performance, making the two efforts complementary within the same broader problem.

LLM-based structured extraction. Generative information extraction with LLMs has become the dominant paradigm for populating knowledge bases from text (Xu et al., 2023; Bian, 2025). Closest to our reliability findings, Shrimal et al. (2025) show that static schemas cause unreliable extraction and propose iteratively optimizing the schema against measured extraction performance – our field-confusion/novel-value decomposition (Section 4.1) is a diagnostic of exactly the kind of schema weakness their approach is designed to correct, applied post-hoc rather than in a training loop. Srivastava et al. (2025) find that textual annotation guidelines materially improve LLM event extraction and cross-schema generalization, consistent with our prompt’s reliance on worked examples for ambiguous field pairs (Section 3) and suggestive that our identified confusion pairs (e.g. trigger_source vs. trigger_mode) are precisely where such guidelines should be strengthened. Han and Lai (2026) argue that closed symbolic relation vocabularies lose the contextual nuance LLMs could otherwise represent in natural language; our finding that two of fourteen closed-vocabulary fields (interaction_actions, moral_condition) absorb the great majority of novel-value drift is direct empirical evidence for exactly this failure mode.

Agreement metrics for LLM annotation. James (2026) survey inter-annotator agreement metric selection for NLP annotation and evaluation, cautioning against uncritical use of raw agreement; our cost_structure result (Section 4.2) is a concrete instance of the failure mode such guidance warns against. We build directly on the classical chance-corrected agreement literature (Fleiss, 1971; Landis and Koch, 1977).

6 Conclusion and Future Work

The two central results – a taxonomy of interaction archetypes that is highly stable across extraction models, obtained even though no individual schema dimension is reliably extracted in isolation – together suggest a practical design principle for LLM-based knowledge base construction from schemas that, like ours, mix observable and interpretive

fields: validate and report reliability at the level of the structure a knowledge base will actually be used for (clusters, types, aggregate statistics), not only at the level of individual extracted fields, since the latter can look considerably less trustworthy than the former. Practically, our field-confusion analysis also gives schema designers a concrete lever: fields whose drift is dominated by confusion with one specific neighboring field (magic_source, duration, actor_type, trigger_source) are candidates for clearer boundary definitions or worked disambiguating examples in the prompt (Srivastava et al., 2025), while fields whose drift is dominated by novel, plausible values (interaction_actions, moral_condition) are candidates for an explicitly open or substantially expanded vocabulary (Han and Lai, 2026) rather than further prompt engineering against a fixed list.

We see four directions. First, a human-annotated gold subset: the results here measure consistency under LLM extraction, and only annotation against the same schema can turn them into statements about accuracy or expose correlated error. Second, sensitivity of the taxonomy: how matched-centroid similarity and internal cluster quality behave across k , whether the merges predicted in Section 4.3 occur, and whether the archetypes survive restriction to tales all three models segment identically. Third, comparison against alternative encodings of the same interactions – Proppian functions alone, or text embeddings – separating what is specific to MAGIC’s distinctions from what any reasonable representation recovers. Fourth, replacing mode-based tale-level aggregation with interaction alignment (Appendix D); and beyond the evaluation, folkloristic analysis of the induced archetypes with domain experts, and propagation of the schema’s core configuration (a protagonist engaging an entity of bounded, extraordinary capability under explicit triggers, costs and conditions) to other domains of complex human behavior.

Acknowledgments

This work was supported by the grant of the state program of the «Sirius» Federal Territory «Scientific and technological development of the «Sirius» Federal Territory» (Agreement №34-03 date 26.08.2026).

Limitations

All reliability and agreement results in this paper are inter-model rather than model-versus-ground-truth: we have no human-annotated gold standard at this stage, and they should be read as measuring the consistency of the schema under LLM-based extraction, not extraction accuracy. The distinction matters more than the phrasing suggests. Agreement among extractors is compatible with all of them being wrong in the same way, and models with overlapping pretraining data are exactly the population in which such correlated error is plausible; a shared prior about what a folktale helper does would produce convergence indistinguishable, by our measurements, from convergence on the text. This applies to the taxonomy-stability result as well as to the field-level κ values. The taxonomy result mitigates the concern partially, by showing that structure survives a change of extractor and so is not any single model’s idiosyncrasy, and the hero_nature calibration shows the pipeline can reach high agreement where the text is unambiguous – but neither substitutes for ground truth, and we do not treat them as doing so. A human-annotated subset is the first item of future work for this reason.

Two of the five models were only run on a 50-tale subset for cost reason. Section 4.2 reports agreement in two conditions specifically to compensate: all five models on small subset, and the three full-corpus models alone on all 1,306 tales. However, the taxonomy-stability result (Section 4.3) still uses the three full-corpus models only.

The taxonomy is induced at a single value of k , selected by the elbow method and held fixed across models for comparability. The archetypes in Table 5 are therefore a description of the corpus at one granularity, and a coarser or finer partition would yield a different, plausibly also stable, set of archetypes. The claim we defend is that the partition is reproducible across extractors, not that five is the correct number of interaction types.

Finally, our corpus obtained from (Kaggle, 2024), while multi-regional, is not a completely random or representative sample of world folk narrative, and the taxonomy in Section 4.3 should be read as a description of recurring patterns in this specific corpus rather than a claim of narratological universality.

We consider these limitations as the directions for further work in the scope of ongoing research

project.

References

- Haonan Bian. 2025. [LLM-empowered knowledge graph construction: A survey](#). *Preprint*, arXiv:2510.20345.
- Joseph Campbell. 1949. *The Hero with a Thousand Faces*. Pantheon Books, New York.
- Sara Graça da Silva and Jamshid J. Tehrani. 2016. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science*, 3(1):150645.
- Ismail El Maarouf and Jeanne Villaneau. 2012. A french fairy tale corpus syntactically and semantically annotated. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 345–350. European Language Resources Association (ELRA).
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Algirdas Julien Greimas. 1983. *Structural Semantics: An Attempt at a Method*. University of Nebraska Press, Lincoln, NE. Translated by D. McDowell, R. Schleifer, and A. Velie. Original work published 1966.
- Julian Hagedorn and Sándor Darányi. 2022. [Bearing a bag-of-tales: An open corpus of annotated folktales for reproducible research](#). *Journal of Open Humanities Data*, 8(16):1–10.
- Kanyao Han and Yushang Lai. 2026. [From symbolic to natural-language relations: Rethinking knowledge graph construction in the era of large language models](#). *Preprint*, arXiv:2601.09069.
- Joseph James. 2026. [Counting on consensus: Selecting the right inter-annotator agreement metric for NLP annotation and evaluation](#). *Preprint*, arXiv:2603.06865.
- Kaggle. 2024. 1000+ Folk Stories around the World. <https://www.kaggle.com/datasets/chayanonc/1000-folk-stories-around-the-world>.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Max Lüthi. 1986. *The European Folktale: Form and Nature*. Indiana University Press, Bloomington, IN. Translated by J. D. Niles.
- Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press, Austin, TX. Translated by L. Scott, revised by L. A. Wagner. Original work published 1928.

- Diandra Schmidt, Albin Zehe, Janis Lorenzen, Lena Sergel, Sabine Dücker, Markus Krug, and Frank Puppe. 2021. The FairyNet corpus – character networks for german fairy tales. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 49–56. Association for Computational Linguistics.
- Anubhav Shrimall, Aryan Jain, Soumyajit Chowdhury, and Promod Yenigalla. 2025. [PARSE: LLM driven schema optimization for reliable entity extraction](#). *Preprint*, arXiv:2510.08623.
- Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. [Instruction-tuning LLMs for event extraction with annotation guidelines](#). *Preprint*, arXiv:2502.16377.
- Qiuyu Tian, Fengyi Chen, Yiding Li, Youyong Kong, Fan Guo, Yuyao Li, Jinjing Shen, Zhijing Xie, Yiyun Luo, Xin Zhang, Yingce Xia, and Zequn Liu. 2026. [Narrative knowledge weaver: Narrative-centric retrieval-augmented reasoning for long-form text understanding](#). *Preprint*, arXiv:2606.05724.
- Hans-Jörg Uther. 2004. *The Types of International Folktales: A Classification and Bibliography*. Number 284–286 in FF Communications. Finnish Academy of Science and Letters, Helsinki.
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2023. [Large language models for generative information extraction: A survey](#). *Preprint*, arXiv:2312.17617.

A MAGIC Attribute Definitions

Table 6 gives the full value set and definition for every interaction-level attribute across the six MAGIC dimensions, expanding the summary in Section 2. This is the same vocabulary implemented in `magiclib/schema.py` and used to compute schema conformance in Section 4.1.

Dim.	Attribute	Values
D1	<code>trigger_source</code>	<code>hero_agency</code> – hero initiated through a deliberate request, search, or invocation; <code>actor_agency</code> – the magic entity acted first, unbidden by the hero; <code>external_event</code> – an outside force (villain, crisis, fate) imposed the trigger
D1	<code>trigger_mode</code>	<code>request</code> – explicit ask or invocation; <code>discovery</code> – hero encounters entity by chance or exploration; <code>autonomous</code> – entity acts on its own initiative; <code>imposed</code> – force, attack, or command pressed upon the hero; <code>crisis</code> – urgent emergency drives the contact
D2	<code>actor_name</code>	Literal identifier of the entity (string)
D2	<code>actor_nature</code>	<code>human</code> / <code>animal</code> / <code>plant</code> / <code>object</code> / <code>natural_force</code> / <code>spirit</code> / <code>other</code>
D2	<code>actor_type</code>	<code>helper</code> – exists to assist protagonists; <code>villain</code> – exists to obstruct or harm the hero; <code>object</code> – non-sentient or minimally sentient magic item
D2	<code>actor_functional_role</code>	<code>ally</code> – cooperates directly with the hero’s goal; <code>donor</code> – provides object/ability/information then steps back; <code>trickster_helper</code> – structurally antagonistic but functionally enabling; <code>tester</code> – evaluates the hero’s worthiness before granting anything; <code>obstacle</code> – actively blocks the hero’s path or goal; <code>neutral</code> – magic is incidental, no stake in hero’s goal
D2	<code>actor_autonomy</code> (ordinal)	<code>none</code> – inert without activation, purely reactive; <code>low</code> – reacts to commands or triggering conditions only; <code>medium</code> – limited independent action within a defined scope; <code>high</code> – makes autonomous decisions within situational rules; <code>full</code> – fully self-directed agency, acts on personal will
D2	<code>agency_balance</code> (ordinal)	<code>hero_dominant</code> – hero directs, actor executes; <code>balanced</code> – both parties contribute meaningfully; <code>actor_dominant</code> – actor determines what happens, hero responds or receives
D2	<code>interaction_hierarchy</code>	<code>primary</code> – directly aids or confronts the hero, no intermediary; <code>secondary</code> – aids the primary actor, not the hero directly; <code>servant_object</code> – controlled by another non-hero actor in service of the hero; <code>chain_mediated</code> – reaches the hero through one or more intermediary actors
D3	<code>magic_source</code>	<code>inherent</code> – always magical by nature (e.g. dragon, fairy); <code>bestowed</code> – enchanted externally, magic is transferred or lent; <code>environmental</code> – power drawn from a place, time, or natural condition; <code>stolen</code> – magic seized or taken from its rightful bearer; <code>transferred</code> – magic passed deliberately from one actor/object to another

Dim.	Attribute	Values
D3	activation_condition	always_active – magic operates continuously without a trigger; hero_request – activated by the hero's explicit invocation or command; crisis_triggered – activates automatically when the hero is in danger; moral_test – activates only if the hero passes a behavioral test; phrase_required – requires specific words, gesture, or ritual
D3	cost_structure	none – no cost to the hero; material – pays with a physical item (food, gold, blood); moral_behavior – must behave in a specific virtuous or prescribed way; knowledge_proof – must answer a question, solve a riddle, or demonstrate knowledge; sacrifice – gives up something of lasting personal value (years, a name, a memory); reciprocal_service – must perform a task or service in return
D3	cost_timing	none – no cost applies; upfront – cost paid before magic is granted; deferred – cost extracted after the benefit is received; ongoing – cost accumulates or repeats during use; conditional – cost only applies if a condition is later met or broken
D3	moral_condition	none / kindness / respect / courage / honesty / humility / specific_task (only meaningful when cost_structure=moral_behavior)
D3	duration	permanent – effect is irreversible and lasts indefinitely; time_limited – effect expires after a fixed period; task_limited – effect persists until a specific task is complete; single_use – magic is consumed in a single activation
D4	outcome_type	success – hero fully achieves the intended goal at no unexpected cost; qualified_success – goal achieved but with a meaningful complication or reduced result; test_passed – hero survives or passes an evaluation, gain is informational or permissive; failure – hero does not achieve the goal; unintended_consequence – outcome differs substantially from what the hero sought
D4	hero_gain (multi-value)	none / object / information / ability / rescue / alliance / permission / transformation
D4	hero_cost (multi-value)	none / material / moral_debt / constraint / transformation / delayed_consequence
D5	narrative_function (multi-value)	NF-1 Reconnaissance/Information – actor provides knowledge the hero needs; NF-2 Provision – actor gives an object, ability, or power; NF-3 Threshold Test – actor evaluates hero's virtue or skill; NF-4 Rescue/Intervention – actor saves the hero from immediate danger; NF-5 Transformation – actor changes the hero's form, nature, or status; NF-6 Alliance Formation – actor commits to an ongoing aid relationship; NF-7 Obstacle/Opposition – actor's primary function is to impede the hero
D6	story_phase	departure – hero is leaving home or entering the quest world; journey – hero is in transit, accumulating helpers and objects; threshold – hero encounters a gateway figure or test before the main challenge; climax – the decisive confrontation or challenge; resolution – events after the main challenge (return, reward, restoration)
D6	interaction_index	Sequential index (1, 2, 3, ...) in order of appearance in the text

Dim.	Attribute	Values
D6	precedes / follows	Lists of interaction_index values this interaction directly enables/is enabled by (direct causal or enabling links only, not general story order)

Table 6: Full attribute definitions for the MAGIC schema (Section 2).

B Actor Network Edge Attributes

Table 7 defines the optional actor-network extension (Section 2), which records relationships between non-hero actors – e.g., in the extraction for *Geraint and Enid*, a controls edge from a sparrow-hawk knight to the dwarf who serves him. *Baba Yaga and the Orphans* (Figure 2) has no non-hero-to-non-hero relationships and so an empty actor network; this extension is not the object of any of the three experiments in this paper.

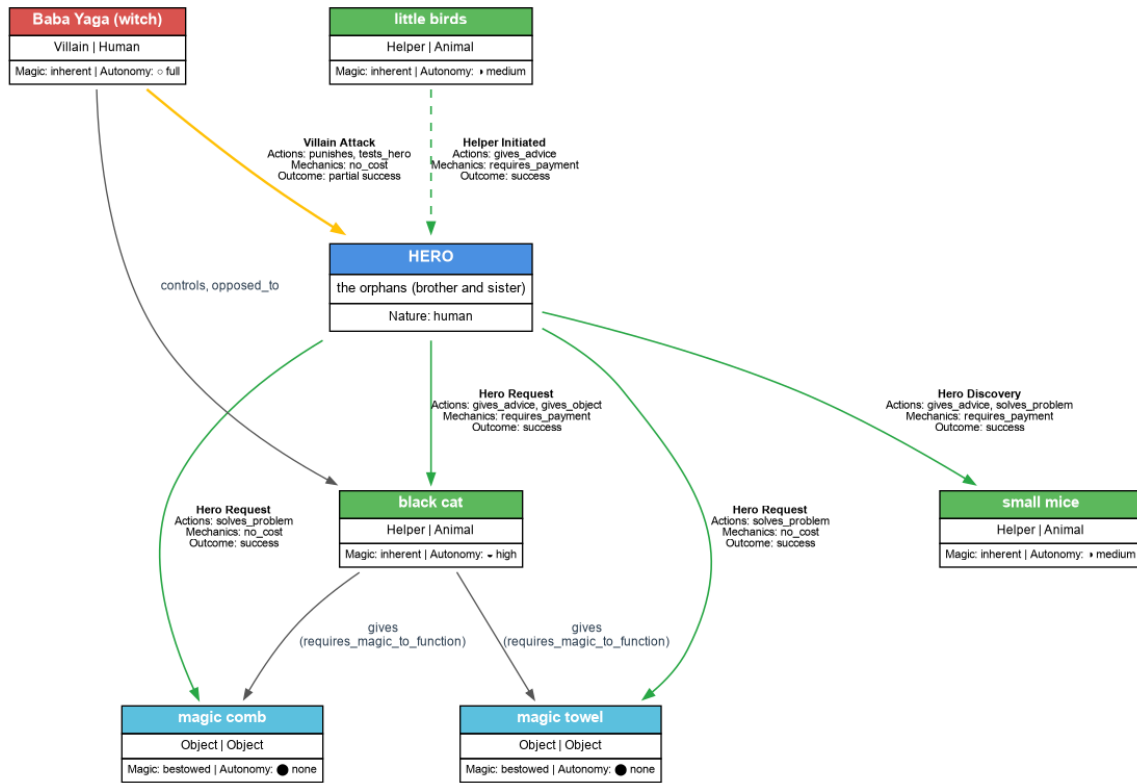


Figure 2: Actor network for *Baba Yaga and the Orphans*.

Attribute	Values
source_actor	/ Names listed in actor_name (D2)
target_actor	
direction	unidirectional / bidirectional
interaction_indices	List of interaction_index (D6) values this relationship directly enables, explains, or produces; empty if none

Attribute	Values
relationship_type (multi-value)	controls – direct command or dominance over target; created – made or summoned the target; gives – transfers an object or ability to target (or to hero via target); takes – seizes something from target; allied_with – cooperative relationship or pact; opposed_to – active conflict or antagonism; depends_on – functional reliance, cannot operate without target; contains – target is physically enclosed within source; mediates – acts as intermediary between target and hero; transforms – changes target’s form, nature, or power; deceives – misleads or entraps target
network_function	mediation – source passes something toward the hero via target; subordination – source has authority over target (master–servant or owner–tool); opposition – source and target are in active conflict affecting the hero; cooperation – source and target act jointly for a shared goal; provision_chain – source → target → hero forms a supply chain of magic objects or aid
magic_dependency	none / requires_magic_to_function / enhanced_by_magic
temporal_relation	before_hero_interaction – established before any linked interaction; concurrent – active during the linked interaction(s); after_hero_interaction – arises as a consequence of the hero’s interactions
chain_depth	0 – either source or target interacts directly with the hero; 1 – one intermediary actor; 2+ – deeper chain (exact integer)

Table 7: Actor-network edge attribute definitions (Section 2).

C Extraction Prompt

The full prompt used by `extract.py` for all five models (Section 3), reproduced verbatim from `code_draft/prompts/magic_extraction_v1.md`, is given below. The story text is appended after this prompt at extraction time.

```
# TASK
Analyze the folktale to identify all significant interactions between the HERO and MAGIC ACTORS/
OBJECTSall. Focus exclusively on:
- Magic helpers (e.g., talking animals, spirits)
- Magic objects (e.g., self-moving items, enchanted tools)
- Villains using magic against the hero

---

## PART 1: HERO INTERACTIONS (PER EVENT)

For EACH INTERACTION EVENT between hero and magic entity (not per actor), extract all fields
below.
Fields are grouped by the six MAGIC framework dimensions. Within each dimension, all fields are required.

---

### STORY-LEVEL FIELDS *(same value for every interaction)*

| # | Field | Values |
|---|---|---|
| 1 | Hero Name | Protagonist's name or identifier |
| 2 | Hero Nature | `human` / `animal` / `plant` / `object` / `natural_force` / `spirit` / `other` |

---
```

```

### DIMENSION 1 Trigger Modality
*Who and how the interaction was set in motion. Use TWO orthogonal fields instead of one.*

**3. `trigger_source`** Who caused the trigger:
- `hero_agency` The hero initiated (request, deliberate search)
- `actor_agency` The magic entity acted first, unbidden
- `external_event` An outside force (villain, crisis, fate) imposed the trigger

**4. `trigger_mode`** How the trigger was expressed (independent of source):
- `request` Explicit ask or invocation by the hero
- `discovery` Hero encounters entity by chance or exploration
- `autonomous` Entity acts on its own initiative, no external prompt
- `imposed` Force, attack, or command pressed upon the hero
- `crisis` Urgent emergency drives the contact

> **Rule:** Assign `trigger_source` and `trigger_mode` independently. Example Baba Yaga seizing
  children: `trigger_source: external_event`, `trigger_mode: imposed`. Do not let the actor's
  identity dictate `trigger_source`.

---

### DIMENSION 2 Actor Profile & Agency Distribution
*Who the actor is (structurally and functionally) and how much control each party holds.*

**5. `actor_name`** Name or description of the magic actor/object.

**6. `actor_nature`** Physical/ontological nature:
`human` / `animal` / `plant` / `object` / `natural_force` / `spirit` / `other`

**7. `actor_type`** Structural identity in the story world (stable across interactions):
- `helper` Exists to assist protagonists
- `villain` Exists to obstruct or harm the hero
- `object` Non-sentient or minimally sentient magic item

**8. `actor_functional_role`** Narrative function in THIS interaction (may differ from `actor_type`):
- `ally` Cooperates directly with the hero's goal
- `donor` Provides object, ability, or information and steps back
- `trickster_helper` Structurally antagonistic but functionally enabling (e.g., Baba Yaga)
- `tester` Evaluates the hero's worthiness before granting anything
- `obstacle` Actively blocks the hero's path or goal
- `neutral` Magic is incidental; entity has no stake in the hero's quest

> **Rule:** `actor_type` and `actor_functional_role` can and often should differ. A villain (`actor_type`) who tests then rewards the hero is a `trickster_helper` (`actor_functional_role`). Always fill both.

**9. `actor_autonomy`** Operational independence as observed in this interaction:
- `none` Inert without activation; purely reactive to physical use
- `low` Reacts to commands or triggering conditions only
- `medium` Limited independent action within a defined scope
- `high` Makes autonomous decisions but within situational rules
- `full` Fully self-directed agency; acts on personal will

> **Rule:** Base this on observed behavior in the text, not the entity's general reputation.

**10. `agency_balance`** Balance of control between hero and actor in this interaction:
- `hero_dominant` Hero directs; actor executes (e.g., commanded magic objects)
- `balanced` Both parties contribute meaningfully to the outcome
- `actor_dominant` Actor determines what happens; hero responds or receives

> **Rubric:** If the hero could substitute a different actor with the same outcome `hero_dominant`. If removing either party's choice collapses the outcome `balanced`. If the actor could have acted regardless of hero's presence `actor_dominant`.

**11. `interaction_hierarchy`** Position in the interaction dependency chain:
- `primary` Directly aids or confronts the hero (no intermediary)
- `secondary` Aids the primary actor, not the hero directly
- `servant_object` Controlled by another non-hero actor in service of hero

```

```

- `chain_mediated` Reaches hero through one or more intermediary actors

---

### DIMENSION 3 Magic Mechanics
*The operational constraints governing how the magic works.*

**12. `magic_source`** Origin of the magical property (renamed from `magic_nature`):
- `inherent` Always magical by nature (e.g., dragon, fairy)
- `bestowed` Enchanted externally; magic is transferred or lent
- `environmental` Power drawn from a place, time, or natural condition
- `stolen` Magic seized or taken from its rightful bearer
- `transferred` Magic passed deliberately from one actor/object to another

**13. `activation_condition`** What causes the magic to fire:
- `always_active` Magic operates continuously without a trigger
- `hero_request` Activated by the hero's explicit invocation or command
- `crisis_triggered` Activates automatically when the hero is in danger
- `moral_test` Activates only if the hero passes a behavioral test
- `phrase_required` Requires specific words, gesture, or ritual

**14. `cost_structure`** What the hero must give, do, or risk to use the magic (replaces binary `requires_payment`):
- `none` No cost to the hero
- `material` Pays with a physical item (food, gold, blood)
- `moral_behavior` Must behave in a specific virtuous or prescribed way
- `knowledge_proof` Must answer a question, solve a riddle, or demonstrate knowledge
- `sacrifice` Gives up something of lasting personal value (years, a name, a memory)
- `reciprocal_service` Must perform a task or service in return

**15. `cost_timing`** When the cost is extracted:
- `none` No cost applies
- `upfront` Cost paid before magic is granted
- `deferred` Cost extracted after the benefit is received
- `ongoing` Cost accumulates or repeats during use
- `conditional` Cost only applies if a condition is later met or broken

**16. `moral_condition`** Specific ethical requirement, if `cost_structure` is `moral_behavior` (otherwise `none`):
`none` / `kindness` / `respect` / `courage` / `honesty` / `humility` / `specific_task`

**17. `duration`** How long the magical effect persists:
- `permanent` Effect is irreversible and lasts indefinitely
- `time_limited` Effect expires after a fixed period
- `task_limited` Effect persists until a specific task is complete
- `single_use` Magic is consumed in a single activation

---

### DIMENSION 4 Outcome Valence
*What actually resulted for the hero. Use THREE fields instead of one.*

**18. `outcome_type`** The primary narrative result:
- `success` Hero fully achieves the intended goal at no unexpected cost
- `qualified_success` Goal achieved but with a meaningful complication or reduced result
- `test_passed` Hero survives or passes an evaluation; gain is informational or permissive
- `failure` Hero does not achieve the goal
- `unintended_consequence` Outcome differs substantially from what the hero sought

**19. `hero_gain`** What the hero concretely receives **(MULTIPLE ALLOWED)**:
`none` / `object` / `information` / `ability` / `rescue` / `alliance` / `permission` / `transformation`

**20. `hero_cost`** What the hero concretely gives up or incurs **(MULTIPLE ALLOWED)**:
`none` / `material` / `moral_debt` / `constraint` / `transformation` / `delayed_consequence`

> **Example:** Baba Yaga tests the orphans: `outcome_type: test_passed`, `hero_gain: ["information", "permission"]`, `hero_cost: ["moral_debt"]`. This is distinct from a `qualified_success`, which implies partial goal attainment.

```

```

---

### DIMENSION 5 Narrative Function *(NEW)*
*What role this interaction plays in the plot architecture.*

**21. `narrative_function`** Select the primary Proppian-grounded function code:

| Code | Name | When to use |
|-----|-----|-----|
| `NF-1` | **Reconnaissance / Information** | Actor provides knowledge the hero needs to continue |
| `NF-2` | **Provision Magical Agent** | Actor gives the hero an object, ability, or power |
| `NF-3` | **Threshold Test** | Actor evaluates hero's virtue or skill before granting anything |
| `NF-4` | **Rescue / Intervention** | Actor saves the hero from immediate physical danger |
| `NF-5` | **Transformation** | Actor changes the hero's form, nature, or social status |
| `NF-6` | **Alliance Formation** | Actor commits to an ongoing aid relationship with the hero |
| `NF-7` | **Obstacle / Opposition** | Actor's primary function is to impede the hero's progress |

> **Rule:** Choose the *dominant* function of the interaction. If an actor provides information **and**
  an object, assign the function that drives the plot forward. Multiple functions may be listed
  if genuinely co-equal: `["NF-1", "NF-2"]`.

---

### DIMENSION 6 Narrative Position *(NEW)*
*Where in the story arc the interaction occurs and how it connects to others.*

**22. `story_phase`** Macro position in the tale's narrative arc:
- `departure` Hero is leaving home or entering the quest world
- `journey` Hero is in transit, accumulating helpers and objects
- `threshold` Hero encounters a gateway figure or test before the main challenge
- `climax` The decisive confrontation or challenge
- `resolution` Events after the main challenge (return, reward, restoration)

**23. `interaction_index`** Sequential integer (1-based) indicating this interaction's position
  among all magic interactions in the tale, in order of appearance in the text.

**24. `precedes`** List of `interaction_index` values this interaction **directly enables or causes**
  **. Empty list `[]` if none.

**25. `follows`** List of `interaction_index` values that **directly enabled or caused** this
  interaction. Empty list `[]` if none.

> **Rule:** Use `precedes`/`follows` only for direct causal or enabling links (e.g., Actor A gives
  Hero an object at index 2; Hero uses that object at index 5 index 2 has `precedes`: [5],
  index 5 has `follows`: [2]). Do not record every story-order predecessor only the
  interactions this one depends on.

---

### INTERACTION ACTIONS *(unchanged, multiple allowed)*

**26. `interaction_actions`** Observable actions performed during the interaction *(MULTIPLE
  ALLOWED)*:
`gives_object` / `gives_advice` / `transforms_hero` / `transforms_object`
`performs_task` / `solves_problem` / `tests_hero` / `rescues`
`deceives` / `punishes` / `demands_payment` / `grants_passage` / `reveals_information`

---

## PART 2: ACTOR INTERCONNECTIONS (TOPOLOGICAL)

For **EACH SIGNIFICANT RELATIONSHIP** between non-hero actors (e.g., helper A controls object B):

**1. `source_actor`** Name of influencing actor (must appear in `interactions`).

**2. `target_actor`** Name of influenced actor (must appear in `interactions`).

**3. `relationship_type`** *(MULTIPLE ALLOWED)*:
- `controls` Direct command or dominance over target
- `created` Made or summoned the target

```

```

- `gives` Transfers an object or ability to target (or to hero via target)
- `takes` Seizes something from target
- `allied_with` Cooperative relationship or pact
- `opposed_to` Active conflict or antagonism
- `depends_on` Functional reliance; cannot operate without target
- `contains` Target is physically enclosed within source
- `mediates` Acts as intermediary between target and hero
- `transforms` Changes target's form, nature, or power
- `deceives` Misleads or entraps target

**4. `direction`**: `unidirectional` / `bidirectional`

**5. `network_function`** *(NEW)* Structural role of this relationship in the magic system:
- `mediation` Source passes something (object, information, effect) toward the hero via target
- `subordination` Source has authority over target (masterservant or ownertool)
- `opposition` Source and target are in active conflict that affects the hero
- `cooperation` Source and target act jointly for a shared goal
- `provision_chain` Source target hero forms a supply chain of magic objects or aid

**6. `chain_depth`** *(NEW)* Number of intermediary actors between this relationship and the hero:
- `0` Either source or target interacts directly with the hero
- `1` One actor between this relationship and direct hero contact
- `2+` Deeper chain (use exact integer)

**7. `temporal_relation`** *(NEW)* When this inter-actor relationship occurs relative to the hero's
interactions:
- `before_hero_interaction` Relationship established before any linked interaction
- `concurrent` Active during the linked interaction(s)
- `after_hero_interaction` Arises as a consequence of the hero's interactions

**8. `interaction_indices`** *(NEW)* List of `interaction_index` values from Part 1 that this
relationship directly enables, explains, or produces. Empty list `[]` if none.

**9. `magic_dependency`**: `none` / `requires_magic_to_function` / `enhanced_by_magic`

---

## RULES

### Annotation principles
- **ONE ENTRY PER INTERACTION EVENT.** If a magic sword is used three times, create three entries.
Each use is a distinct event with its own trigger, outcome, and narrative function.
- **Order by text appearance.** `interaction_index` must reflect the sequence in the original
narrative.
- **Observed behavior only.** All fields especially `actor_autonomy`, `agency_balance`, and `
actor_functional_role` must reflect what the text shows, not general knowledge about the entity
type.
- **`actor_type` vs `actor_functional_role` can diverge.** This is expected for Trickster-Tester
figures. Both fields are always required.
- **`trigger_source` is independent of `actor_type`.** A villain causing a trigger `trigger_source:
external_event`, not a villain-specific trigger value.
- **`hero_name` and `hero_nature` are story-level.** Do not vary them across interaction entries.

### Completeness rules
- Include all interactions involving the HERO and MAGIC entities. Omit mundane human interactions
.
- For hybrid actors (e.g., cursed prince), use the dominant nature during this specific interaction
.
- If no magic interactions exist, return `"interactions": []`.
- `actor_network` covers only non-hero actors. Hero relationships belong in `interactions`.
- Actor network relationships must be explicit in the text. No speculative connections.
- If no non-hero relationships exist, return `"actor_network": []`.

### Causal chain rules
- `precedes`/`follows` encode direct enabling links only, not every prior interaction.
- `chain_depth` in `actor_network` counts actors between the relationship and the hero, not
narrative steps.
- `interaction_indices` in `actor_network` entries must reference `interaction_index` values that
exist in `interactions`.

```

```

### ATU classification
- Identify the Aarne-Thompson-Uther Index (ATU) according to the predominant tale type. Use the most
  specific applicable code.

---

## FORMAT

Output **STRICT JSON** with the structure below. No text before or after. No explanations. No
  comments inside the JSON.

```json
{
 "story_title": "Short story name",
 "ATU_Index": "Identified ATU Index",
 "hero_name": "Protagonist identifier",
 "hero_nature": "human|animal|plant|object|natural_force|spirit|other",
 "interactions": [
 {
 "interaction_index": 1,
 "actor_name": "...",
 "actor_type": "helper|object|villain",
 "actor_functional_role": "ally|donor|trickster_helper|tester|obstacle|neutral",
 "actor_nature": "human|animal|plant|object|natural_force|spirit|other",
 "actor_autonomy": "none|low|medium|high|full",
 "agency_balance": "hero_dominant|balanced|actor_dominant",
 "interaction_hierarchy": "primary|secondary|servant_object|chain_mediated",
 "trigger_source": "hero_agency|actor_agency|external_event",
 "trigger_mode": "request|discovery|autonomous|imposed|crisis",
 "magic_source": "inherent|bestowed|environmental|stolen|transferred",
 "activation_condition": "always_active|hero_request|crisis_triggered|moral_test|phrase_required",
 "cost_structure": "none|material|moral_behavior|knowledge_proof|sacrifice|reciprocal_service",
 "cost_timing": "none|upfront|deferred|ongoing|conditional",
 "moral_condition": "none|kindness|respect|courage|honesty|humility|specific_task",
 "duration": "permanent|time_limited|task_limited|single_use",
 "interaction_actions": ["gives_object", "gives_advice"],
 "narrative_function": ["NF-1"],
 "story_phase": "departure|journey|threshold|climax|resolution",
 "outcome_type": "success|qualified_success|test_passed|failure|unintended_consequence",
 "hero_gain": ["object", "information"],
 "hero_cost": ["none"],
 "precedes": [],
 "follows": []
 }
],
 "actor_network": [
 {
 "source_actor": "...",
 "target_actor": "...",
 "relationship_type": ["gives", "controls"],
 "network_function": "mediation|subordination|opposition|cooperation|provision_chain",
 "direction": "unidirectional|bidirectional",
 "chain_depth": 0,
 "temporal_relation": "before_hero_interaction|concurrent|after_hero_interaction",
 "interaction_indices": [1, 3],
 "magic_dependency": "none|requires_magic_to_function|enhanced_by_magic"
 }
]
}
```

**NO TEXT BEFORE/AFTER JSON. NO EXPLANATIONS.**

---

# TEXT

```

D Pairwise Agreement, Confidence Intervals, and a Disagreement Case Study

This appendix gives the full detail behind the four checks summarized in Section 4.2.

Bootstrap confidence intervals. Table 8 gives 95% CIs (2,000 tale-resampling bootstrap iterations) on every field’s Fleiss’ κ , in both conditions, alongside the calibration field `hero_nature` (story-level: human/animal/plant/object/natural_force/spirit/other, constant across a tale and almost always unambiguous from the text). Two things stand out. First, every CI at $n=1,306$ is narrow (width ≤ 0.06 for six of the eight fields, ≤ 0.08 for the remaining two, `cost_structure` and `hero_nature`), so the full-corpus estimates in Table 4 are precise, not noisy. Second, `cost_structure`’s $n=50$ CI, $[0.040, 0.243]$, does not contain its own $n=1,306$ estimate (0.328) – the discrepancy in Table 4 is not explained away by ordinary tale-resampling uncertainty at $n=50$. To check whether this is instead a *model-set* effect (recall κ_{50} uses five models, κ_{1306} only three), we recomputed κ for `cost_structure` using only the three full-corpus models restricted to the same 50 tales: $\kappa = 0.148$, CI $[-0.057, 0.373]$ – essentially identical to the five-model $n=50$ estimate (0.151) and nothing like the $n=1,306$ estimate. Sample size, not which models are included, is responsible for the shift.

| Field | κ_{50} [95% CI] | κ_{1306} [95% CI] |
|------------------------------------|------------------------|--------------------------|
| <code>trigger_source</code> | 0.32 [0.22, 0.41] | 0.32 [0.29, 0.34] |
| <code>actor_type</code> | 0.44 [0.33, 0.56] | 0.45 [0.42, 0.48] |
| <code>actor_functional_role</code> | 0.28 [0.19, 0.37] | 0.32 [0.30, 0.35] |
| <code>cost_structure</code> | 0.15 [0.04, 0.24] | 0.33 [0.29, 0.37] |
| <code>outcome_type</code> | 0.25 [0.15, 0.34] | 0.35 [0.33, 0.38] |
| <code>narrative_function</code> | 0.29 [0.21, 0.37] | 0.31 [0.29, 0.34] |
| <code>story_phase</code> | 0.19 [0.11, 0.26] | 0.27 [0.25, 0.30] |
| <code>hero_nature</code> (calib.) | 0.64 [0.38, 0.81] | 0.79 [0.75, 0.82] |

Table 8: Fleiss’ κ with bootstrap 95% CIs, both conditions.

Pairwise agreement. Table 9 summarizes pairwise Cohen’s κ across all model pairs (10 pairs at $n=50$, 3 pairs at $n=1,306$) per field. At $n=50$, pairwise values for `cost_structure` range from -0.064 to 0.469 – three of the ten pairs are *below chance* – while `actor_type`’s ten pairs range 0.334 – 0.600 with no pair falling near zero and no single model appearing as an outlier against the rest (every model’s pairwise values with the other four are of comparable magnitude). At $n=1,306$, all pairwise ranges tighten substantially and no field shows a below-chance pair, consistent with the $n=50$ volatility reflecting sample size rather than a persistent low-agreement relationship between specific models.

| Field | $n=50$ (10 pairs) | | $n=1,306$ (3 pairs) | |
|------------------------------------|-------------------|------|---------------------|------|
| | min | max | min | max |
| <code>trigger_source</code> | 0.19 | 0.52 | 0.29 | 0.37 |
| <code>actor_type</code> | 0.33 | 0.60 | 0.41 | 0.53 |
| <code>actor_functional_role</code> | 0.05 | 0.53 | 0.29 | 0.36 |
| <code>cost_structure</code> | -0.06 | 0.47 | 0.31 | 0.37 |
| <code>outcome_type</code> | 0.16 | 0.37 | 0.33 | 0.40 |
| <code>narrative_function</code> | 0.15 | 0.38 | 0.30 | 0.35 |
| <code>story_phase</code> | 0.07 | 0.25 | 0.25 | 0.31 |
| <code>hero_nature</code> (calib.) | 0.42 | 0.92 | 0.74 | 0.81 |

Table 9: Pairwise Cohen’s κ range (min/max over all model pairs) per field, both conditions.

Disagreement case study: interaction segmentation as a confound. Table 10 shows every interaction’s `cost_structure` value, per model, for *Baba Yaga and the Orphans* (the tale used in Figure 2 and Table 1). All three models agree that the three animal-helper interactions (mice, cat, birds) cost the hero material payment – identical content, three occurrences each. Beyond these three, the models extract different numbers of additional interactions around the villain, her dwelling, and the two escape objects – DeepSeek 3 more (2 none, 1 `moral_behavior`), Qwen 4 more (all none), Gemini 6 more (5 none, 1

moral_behavior) – none of which is material. Because our tale-level metric takes the *mode* across a tale’s interactions (Section 4.2), these extra zero/non-material-cost interactions are enough to flip the tale-level dominant value from material (DeepSeek: 3 of 6) to none (Gemini: 5 of 9; Qwen: 4 of 7) – despite no disagreement about any individual interaction the three models have in common. This is a structural limitation of mode-based tale-level aggregation, not evidence that the schema’s cost-structure distinctions are themselves harder to apply consistently than the raw κ values alone would suggest.

| Actor | DeepSeek | Gemini | Qwen |
|---------------------------|----------|----------|----------|
| small mice | material | material | material |
| little birds | material | material | material |
| black cat | material | material | material |
| Other interactions | 3 | 6 | 4 |
| none | 2 | 5 | 4 |
| moral_behavior | 1 | 1 | 0 |
| Total interactions | 6 | 9 | 7 |
| Tale-level mode | material | none | none |

Table 10: cost_structure values for *Baba Yaga and the Orphans*, by model. The three animal-helper interactions are identically coded by all three models; “Other interactions” covers the villain, her dwelling, and the two escape objects (towel, comb), which each model splits into a different number of separate interactions, none of them material.

E Supplementary Taxonomy Figures

Figures 3 and 4 support the orthogonality claim in Section 4.3: each point plots one region (resp. ATU type)’s share of a cluster against its share of the full corpus, for the DeepSeek V3.2 extraction. Points falling near the diagonal are proportionally represented in that cluster relative to the whole corpus; points above/below it are over-/under-represented. No region or ATU type falls far from the diagonal for any cluster, consistent with the taxonomy cutting across both axes rather than reproducing either.

