EMNLP2026 BUDAPEST

11th Workshop on Automated Knowledge Base Construction
28 October 2026
Budapest • Hungary

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

Sergey Kovalchuk[1,2], Anastasia Ugleva[3], Oleg Kubryak[1,4]

[1]Sirius University of Science and Technology [2]ITMO University
[3]Higher School of Economics [4]National Research University "MPEI"

Email: sergey.v.kovalchuk@gmail.com

Abstract

We present **MAGIC**, a six-dimensional schema for building knowledge bases of hero–magic-entity interactions from folktales: every exchange between a protagonist and a magic helper, villain or object is decomposed into trigger, actor & agency, magic mechanics, outcome, narrative function and position. An LLM pipeline applied to **1306 tales** with five models yields an interaction KB whose clustering recovers **five nameable archetypes, stable across extractors (0.95 centroid similarity)** and orthogonal to tale type and region – even though **no single field is reliably extracted** (Fleiss' κ ⩽ 0.45). LLM-built KBs should be evaluated at the level of the structure they support.

The MAGIC Schema

Unit of analysis = one **interaction**, not a tale (ATU) or a hero arc (monomyth). 23 interaction-level attributes on six orthogonal dimensions + 2 story-level fields, closed vocabularies.

D1 Trigger Modality · 2 attrs.
who started it (hero / actor / external) vs. how (request, discovery, autonomous, imposed, crisis)

D2 Actor Profile & Agency · 7 attrs.
structural type ≠ functional role; autonomy; hero–actor control balance

D3 Magic Mechanics · 6 attrs.
source, activation, cost structure & timing, moral condition, duration

D4 Outcome Valence · 3 attrs.
result type + what the hero concretely gains and pays

D5 Narrative Function · 1 attr.
seven Proppian-derived codes, NF-1 Reconnaissance … NF-7 Obstacle

D6 Narrative Position · 4 attrs.
story phase + direct causal precedes / follows links

Worked Example

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]

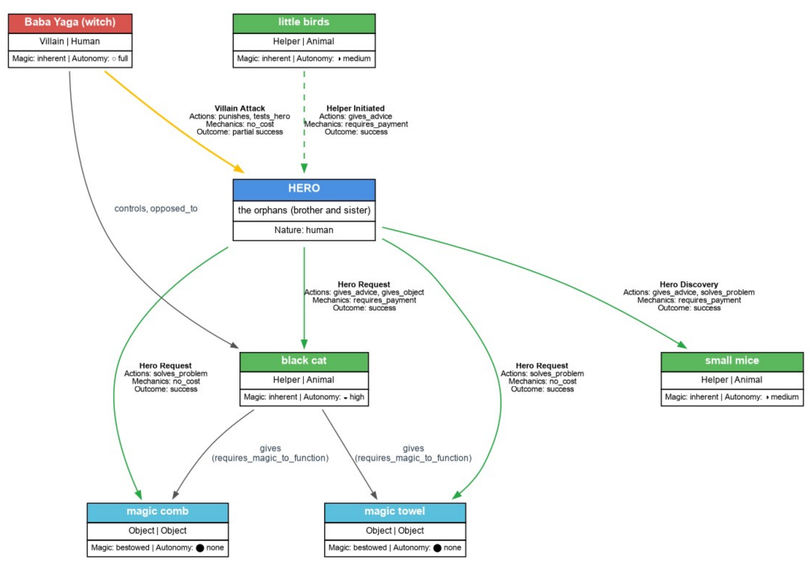


Table 1 / Figure 1. Baba Yaga and the Orphans: the black-cat interaction (left) and the tale's interaction network (right).

Extraction Pipeline

1306 folktales — worldwide Kaggle corpus, Grimm-heavy → **One prompt** — 25 fields, closed vocabularies, worked examples → **5 LLMs** — 3 full-corpus · 2 on a 50-tale subset → **Validated KB** — schema check → JSON interactions + ATU

Model	Parse	Conform.	Int./tale	Confusion share of drift
DeepSeek V3.2	96.7%	93.9%	5.10	57.5%
Gemini 2.5 Flash	98.3%	88.3%	6.54	93.3%
Qwen 3.6 35B	99.0%	90.0%	4.42	93.4%
GPT-5.2 †	98.0%	99.2%	5.43	100%
GPT-OSS-120B †	90.0%	80.6%	4.00	34.1%

Table 2. Extraction reliability. Drift = out-of-vocabulary values; confusion = value valid for another field, the rest are invented values. † 50-tale subset.

Cross-Model Field Agreement

Fleiss' κ on each tale's dominant field value, in two conditions: **κ50** – 5 models × 50 shared tales; **κ1306** – 3 models × all 1306 tales.

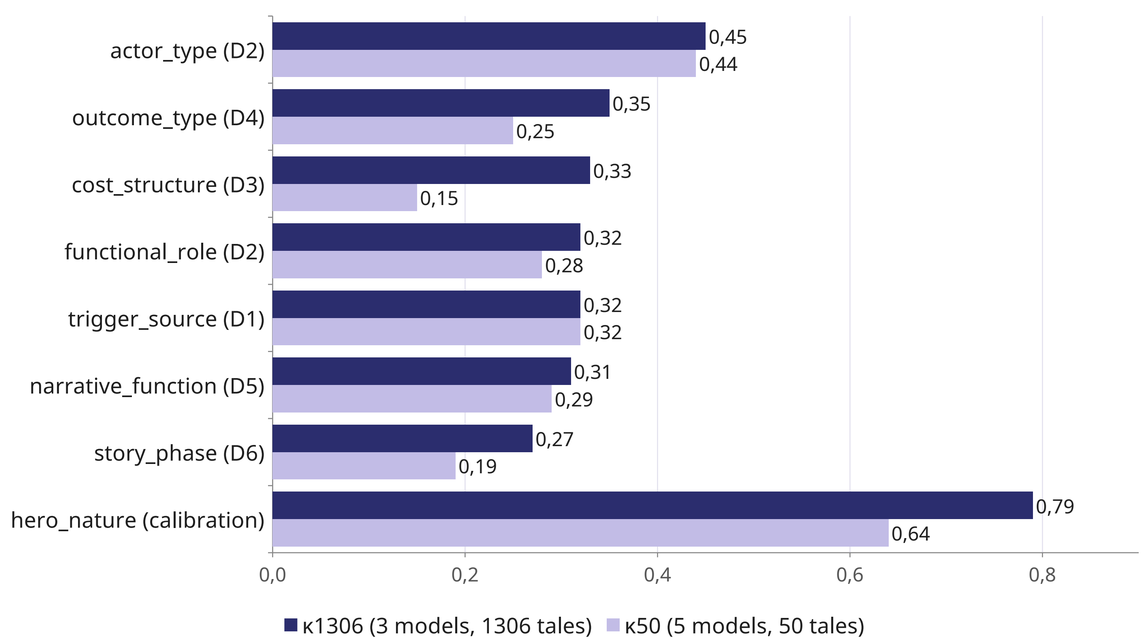


Figure 2. Agreement per MAGIC dimension vs. a near-unambiguous calibration field.

No field beyond moderate: actor_type is highest in both conditions; the calibration field reaches 0.64–0.79, so the dimensions are hard – not the pipeline.

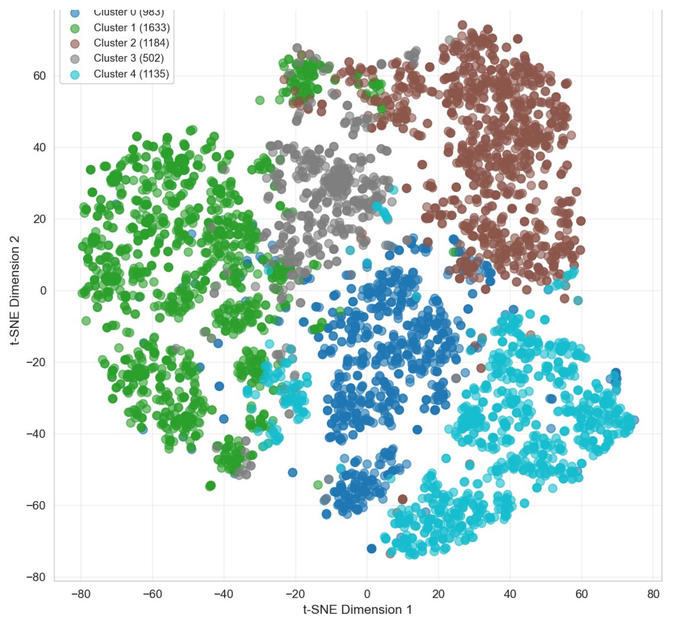
Small samples mislead: field rankings at n=50 barely match n=1306 (Spearman 0.21); cost_structure moves 0.15 → 0.33.

Raw agreement inflates: 80% of tales have ⩾4/5 models agreeing on cost_structure, driven by the 85% "none" majority class.

Drift is diagnosable: confusion clusters on field pairs (trigger_source <-> mode, actor_type <-> nature); invented values on interaction_actions & moral_condition.

Five Interaction Archetypes

k-means (k = 5, elbow) over 93-dim feature vectors, run independently per full-corpus model; villain interactions excluded.



C0 Invoked helper n=983
hero-requested 76% · helper 97%

C1 Bestowed object n=1633
hero-controlled 91% · no cost 91%

C2 Moral-test gate n=1184
moral cost 63% · mixed outcome

C3 Object obstacle n=502
actor-dominant 64% · NF-7 57%

C4 Unsolicited helper n=1135
actor-initiated 98% · no cost 99%

Figure 3. t-SNE of 5437 DeepSeek V3.2 interactions and the archetype signatures.

0.953 mean matched-centroid cosine similarity across 3 extractors

⩾ 0.858 worst of the 15 matched cluster pairs

ATU 425 top tale type in all 5 clusters – archetypes cut across type & region

Conclusion

Archetype-level structure is stable across LLM extractors even though individual fields are not. Validate LLM-built knowledge bases at the level of the structure they serve – clusters, types, aggregates – and use drift type to fix schemas: clearer boundaries for confused fields, open vocabularies for inventive ones.

Next: human-annotated gold subset, k-sensitivity, alternative encodings, and transfer of the schema to human–AI interaction.