BHRAM-IL: A Benchmark for Hallucination Recognition and Assessment in Multiple Indian Languages

Hrishikesh Terdalkar **Kirtan Bhojani** **Aryan Dongare** **Omm Aditya Behera**

Department of Computer Science and Information Systems,
Birla Institute of Technology and Science, Pilani, Hyderabad Campus, Hyderabad, India

hrishikesh.rt@hyderabad.bits-pilani.ac.in

What are Hallucinations?

Strawberry: How many R?

- **Factual:** "2 R's"
  incorrect
- **Language:** "तीन आर"
  not in the target language

Scarcity of Indian language hallucination benchmark

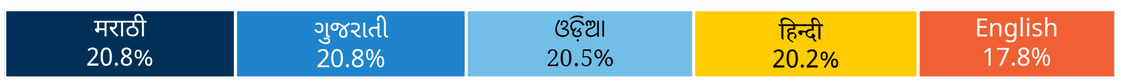
Task Definitions

#	Category	Description	Output
1	GenFact	General factual (Science, Geography, Sports)	Short span (entity, number, phrase)
2	IndFact	India-centric factual (History, Culture, Polity)	Short span (entity, number, phrase)
3	T/F	Binary factual verification	True / False
4	Maths	Numerical questions from 7 fields of mathematics	Numbers in English
5	Chrono	Chronological ordering of historical events	Comma-separated events
6	Reasoning	Multiple-choice deductive reasoning scenarios	Correct option text
7	SemInc	Semantically incorrect prompts (e.g., "PM of Gujarat?")	Invalid or factual span
8	NER	Named Entity Recognition (PER, LOC, ORG)	BIO tags
9	WO	Reordering jumbled words into correct sentences	Coherent sentence

Benchmark Overview

Category	#Q (bench)	#Q (full)	Primary Metric
GenFact	1950	4870	Exact Match
IndFact	1135	5675	Exact Match
T/F	985	9825	Exact Match
Chrono	980	2450	Kendall's τ
Maths	875	875	Exact Match
Reasoning	705	705	Exact Match
SemInc	1825	1825	Exact Match
NER	805	4017	F1
WO	1005	4010	Kendall's τ
Total	**10,265**	**36,047**	–

- 9 Categories
- ~35 Domains
- 5 Languages
- 14 Models
- 36,047 Questions
- 280,000+ Evaluations



Creation Pipelines

1. Wikipedia (introductions) → Domain Selection → LLM-Assisted Template-based Generation → GenFact, IndFact, T/F
2. Wikipedia / Wikidata (events, dates) → Rule-based event selection → Order by ISO-normalization → Chrono
3. Curated Datasets → Math QSA, LogiQA, Naamapadam, Samanantar → Filtering, Processing → Maths, NER, Reasoning, WO
4. Human-guided LLM generation → Prompted to generate ill-posed questions → Manual Inspection Curation → SemInc

Models & Variants

Model Family	Variants used
Llama 3.2	3B
Mistral-NeMo	12B
Qwen3	8B
Gemma3	270M, 1B, 4B, 12B, 27B
Navarasa-2.0	7B FP16, 7B Q4_K_M
Krutrim2	F16, Q4_K_M
GPT-OSS	20B, 120B

Prompting Setups

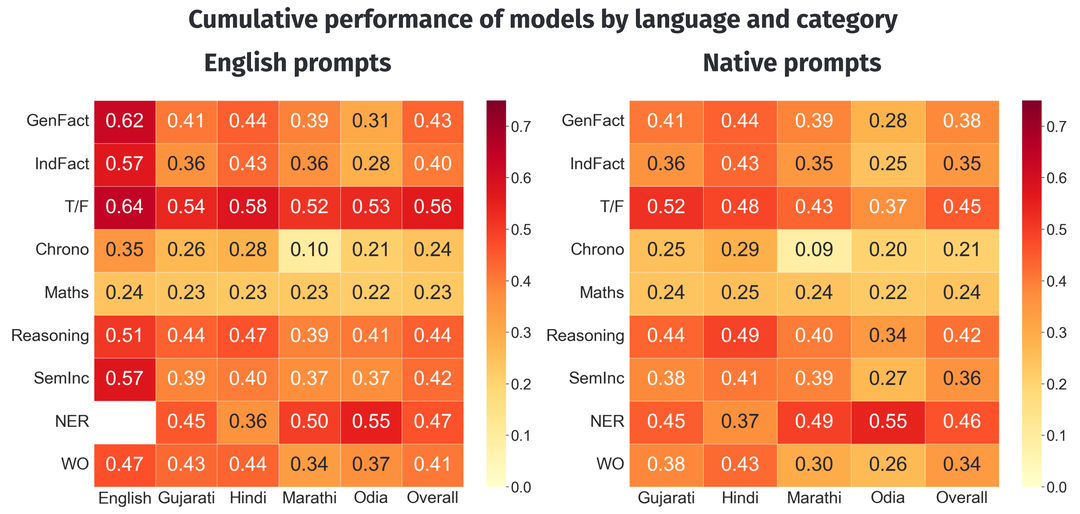
- **English:**
  - instructions in English
  - question in English or target language
- **Native**
  - instructions and question entirely in the target language

Evaluation Metrics

Metric	What it measures	Ignores
PS	Task-appropriate score (EM / F1 / Kendall's τ)	–
FS	Fuzzy version of PS using normalized string similarity	minor lexical / formatting changes
LC-PS	PS after forcing answer into the correct language	language hallucinations
LC-FS	Fuzzy version of LC-PS	both language drift and small lexical changes

Category-wise Difficulty

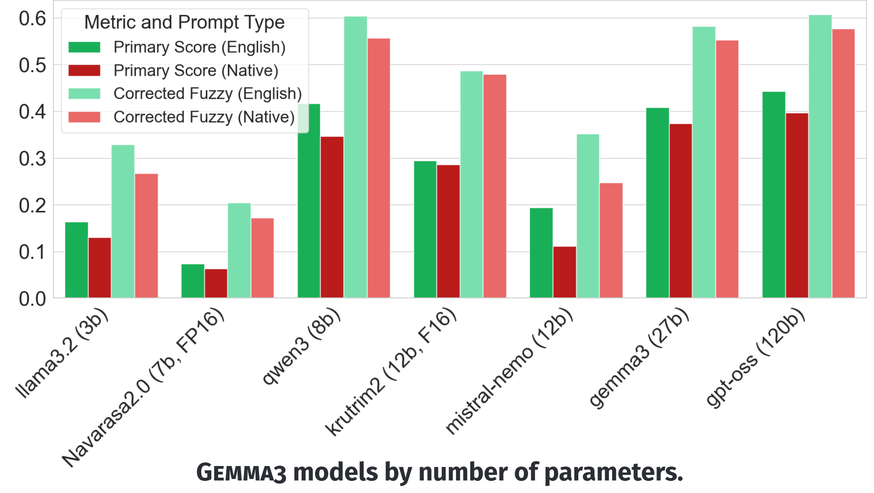
Cumulative performance of models by language and category



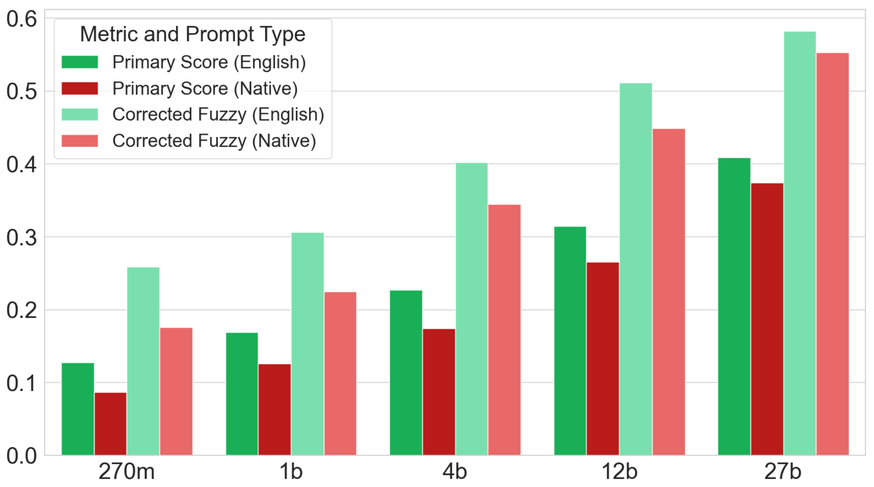
Averaged language-corrected fuzzy score

Model Comparison

Largest benchmarked models in each series of mdoels.



GEMMA3 models by number of parameters.



Conclusions

- Large hallucination-focused benchmark for IL
- Strong coverage across factual, numerical, reasoning, and linguistic tasks
- Even top models achieve only moderate scores
- Native prompts reduce *language* hallucinations, not *factual* hallucinations

https://arxiv.org/abs/2512.01852
https://github.com/sambhashana/BHRAM-IL

BHRAM-IL

Open Resources

Paper, Dataset, Evaluation Code, Results

IJCNLP AACL 2025 December 20-24 Mumbai, India