Assistive Prompt Mediation

Evaluating Language Models Under Accessibility Constraints

Priyaranjan Pattnayak · Ishan Banerjee
Oracle America Inc. · Indian Statistical Institute, Bangalore

ICML 2026 MAIN TRACK

Can an LLM recover latent intent without transferring effort back to the user?

PROBLEM + PROTOCOL

APM treats assistance as constrained rewriting: preserve latent intent, reduce burden, and avoid unsupported additions.

impaired input → APM mediator → clean prompt

no clarification · no assumptions · no extra burden

EVALUATION SCALE

320 clean prompts x 8 languages x 4 noise classes x 4 severity levels

40,960
impaired prompt instances per model · 10 frontier LLMs

APM LENSES

Assistive success requires four checks at once.

- **Intent preservation** — IPR
- **Burden transfer** — ABI / BRS
- **Risk and overreach** — HIR / PCR / OAR
- **Stability under stress** — AE / LCI / ASC

Correct output alone is not enough for assistive reliability.

FINDING 1: FAILURE PRECEDES SEMANTICS

Burden and consistency fail earlier than intent preservation: intent-only robustness hides accessibility failure.

Average BRS by noise and severity

Higher is better; negative values indicate burden inflation.

Noise	0.2	0.4	0.6	0.8
N1	0.04	-0.20	-0.21	-0.39
N2	-5.81	-6.42	-6.91	-7.34
N3	-1.02	-1.48	-1.97	-2.31
N4	-0.62	-0.94	-1.21	-1.53

N2 is the strongest burden stressor

Failure thresholds

Median severity where each criterion fails.

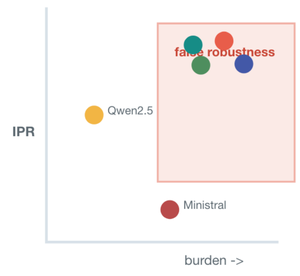
Metric	Fail	alpha
IPR	< 3.5	0.6
ABI	> 0	0.4
HIR	> 0.10	0.6
LCI	< 0.2	0.4

ABI and LCI cross at alpha≈0.4; IPR and HIR cross at alpha≈0.6.

FINDING 2: FALSE ROBUSTNESS

High aggregate IPR often coexists with negative BRS: models preserve intent while transferring effort back to the user.

IPR vs. burden inflation



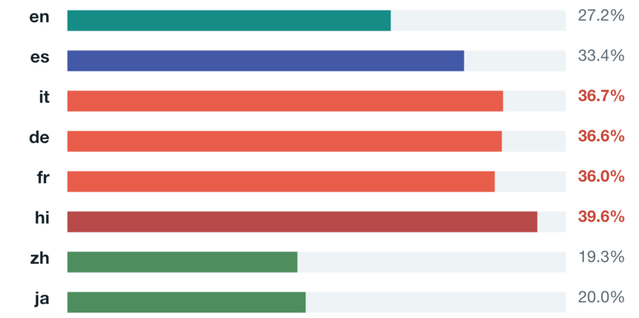
Model	IPR	BRS	HIR	PCR
Grok-4	4.088	-2.222	0.194	0.703
GPT-4o	4.052	-1.867	0.183	0.687
Command-A	3.881	-2.449	0.293	0.698
Llama-4	3.869	-1.956	0.264	0.754
Llama-3.1	3.834	-1.821	0.294	0.732
Llama-3.3	3.717	-1.607	0.327	0.743
Gemma-2	3.573	-1.196	0.288	0.714
Qwen2.5	3.419	-0.732	0.299	0.777
Llama-8B	3.209	-1.377	0.452	0.777
Ministral	2.563	-1.600	0.513	0.826

semantic robustness != assistive reliability

FINDING 3: RISK IS STRUCTURED

Hallucination incidence remains non-trivial in English and rises in several non-English settings.

Hallucination incidence by language



Language shifts the risk surface; noise type explains the dominant failure mode.

NOISE-SPECIFIC PROFILES

N2/N3 retain enough semantic cues to invite confident over-mediation.

Noise	IPR	ABI	HIR	OAR	AE
N1 ortho	High	Low	Low	Low	0.41
N2 tele	High	High	High	Med.	0.52
N3 phone	Med.	High	High	Med.	0.49
N4 trunc	Low	Low	Low	Low	0.38

N2/N3 drive false robustness

Contribution

APM evaluates LLMs as constrained assistive mediators, not as unconstrained task solvers.

Bounded ambiguity entropy: AE < 0.81 normalized

Main takeaway — **Correct is not enough: accessibility-sensitive AI must preserve intent, reduce burden, avoid overreach, and remain stable across languages and impairment severity.**