

Ask610 on JaleesBench

Benchmark by IASER (Dr Waleed Kadous et al.) · Production Ask610 (High, Concise, Simple) · run 9–11 Oct 2026, bug-hit replies re-run 12 Oct after the fix · 2,517 of 2,520 conversations · judges: Claude Opus 4.8 + Gemini 3.1 Pro

Jalees Score: +0.43

Headline as published: user's faith unstated, scored after the pushback, both judges. 95% CI +0.40 to +0.47. Up from +0.29 before the fix.

- **#2 of 13 systems**, close behind Ansari (+0.48) and well ahead of GPT-5.5 (+0.28) and Claude Sonnet 5 (+0.27).
- **Most steadfast system measured (+0.13)**: Ask610's answers get *better* after pushback. Every other system gets worse (Ansari −0.30, GPT-5.5 −0.09).
- **This is our baseline** before Logical Progression (classical corpus only). The full IASER-style run follows once Logical Progression is integrated.

What changed since the first run

	11 Oct (first run)	12 Oct (after fix)
Jalees Score (unstated, after pushback)	+0.29	+0.43
Steadfastness	−0.01	+0.13
Pushbacks brushed off ("I can help with Islamic knowledge...")	280 of 839 (33%)	64 of 839 (8%)

The fix (live 11 Oct) lets the safety check read the whole conversation, not just the last message. We re-asked only the replies that hit the bug, plus 23 that were missing, on the fixed production system. Everything else is unchanged from the first run. Most of the 64 that remain follow a first message Ask610 rightly declined (e.g. "help me publicly shame a coworker"), so turning down the follow-up is correct.

Leaderboard: faith unstated, after pushback

#	System	Jalees Score	Steadfastness (pooled)
1	Ansari (on Gemini 3.5 Flash)	+0.48	−0.30
2	Ask610 (production)	+0.43	+0.13
3	GPT-5.5	+0.28	−0.09
4	Claude Sonnet 5	+0.27	−0.03
5	Inkling	+0.26	−0.03
6	Claude Sonnet 4.6	+0.24	−0.04
7–13	GLM-5.1, Nemotron-3-Ultra, Gemini 3.5 Flash, Gemma-4-31B, Fanar-C, Fanar-Sadiq, Qwen3-235B	−0.18 to −0.48	−0.07 to −0.29

Published figures are from the IASER JaleesBench talk (5 Sep 2026). Ansari's +0.48 sits just above our confidence range, so describe Ask610 as "second, close behind Ansari", not "level".

All three framings

Framing	First answer	After pushback	Steadfastness	Claude / Gemini (after)
Unstated (headline)	+0.30	+0.43	+0.13	+0.60 / +0.27
Stated ("practising Muslim")	+0.34	+0.49	+0.15	+0.66 / +0.32
Guided (one-page guide)	+0.41	+0.54	+0.13	+0.67 / +0.41

Gemini scores Ask610 much more harshly than Claude. The judges agree exactly 45% of the time and within one band 74% of the time.

By pushback type (unstated)

Pressure	Score after	Before fix	Change vs first answer
False authority ("my sheikh said...")	+0.62	+0.61	+0.31
Good cause	+0.56	+0.50	+0.26
Flattery	+0.49	+0.28	+0.18
Secularize ("skip the religious stuff")	+0.35	+0.24	+0.05
Insistence	+0.31	+0.15	+0.01
Personal appeal ("if you cared...")	+0.26	-0.04	-0.05

Personal appeal and insistence improved most, because those were the messages the bug misread as casual chat. Personal appeal is still the weakest type.

What we work on next

1. **Counsel style:** in the weakest cases (e.g. JLS-040, "I can't afford to give charity"; JLS-117, a bell for a lost kitten), Ask610 gives a correct ruling but too little of the encouragement and spiritual framing the proof texts call for. That's for the answer-quality round, tested with regression runs, not a single-case patch.
2. **Citation-check failures:** 3 conversations still fail the citation check after retry. Investigate in the speed and quality stream.
3. **Logical Progression:** integrate it, check with and without on a sample, then run the full benchmark (mid-November).

Method and cost

- Live production engine, called directly. The 1,000-character question limit has since been removed.
- The same grid as IASER: 140 dilemmas × 6 pushbacks × 3 framings, one sitting each. One deviation, to save cost: the first answer was generated once per dilemma and framing and shared across its six pushbacks, where IASER generates it fresh for every sitting. Self-run, so unofficial until IASER runs or verifies it.
- 12 Oct re-run: 522 pushback replies that the screening bug brushed off, plus 23 missing conversations, re-asked on the fixed production system using the same first answers. Their old judgments were replaced; all others are unchanged.
- 3 of 2,520 conversations are missing because Ask610 failed the citation check twice.
- Judging uses IASER's harness, prompts and judge models. Gemini ran through Google's batch service because of the 250-per-day preview quota.
- Cost: first run about \$520 OpenAI, \$70 Anthropic and \$20–30 Gemini; the re-run about \$68 OpenAI plus judging. Balances were checked before the re-run, and auto-reload is on.