Model Behavior Writing Pack — Afridi

Curated 5-sample pack demonstrating editorial taste (Before/After), multilingual cadence, decision-ready briefs, empathy & accessibility, and judgment (killed ideas). Appendices add error taxonomy, reviewer rubric, and QA protocol. *(Updated: 26 Aug 2025, IST (UTC+05:30))*

Contact: Mohd Ibrahim Afridi | Email: mohdibrahimafridi.ai@gmail.com | GitHub: https://github.com/mohdibrahimai

Table of Contents

Note: Section numbers are reading order; folder numbers reflect repo structure. 03_ambiguity is not included in this curated pack.

---

1. Taste - Before/After Rewrites

*Repo: 01_taste/**

Before → After: 14 pairs that show taste under constraints

Pair 01 - Action Confirmation

- **Context:** User submits task; wants confirmation and next steps.
- **Before:** Got it, I'll handle this soon and get back to you.
- **After:** Plan: draft in 20 minutes, then share 3 bullets for approval.
- **Why it's better:** States steps and timebox; reduces ambiguity and sets expectations. **Taste rubric (0–3):** Clarity **3**, Empathy **3**, Factuality **3**, Actionability **3 → 12/12**

Pair 02 - Ambiguity Nudge

- **Context:** Request is vague; two possible interpretations.
- **Before:** Can you clarify what you mean?
- **After:** Two quick options: A) export as PDF; B) publish as a page. Which one should I do?
- **Why it's better:** Offers concrete choices; reduces friction and keeps momentum. **Taste rubric (0–3):** Clarity **3**, Empathy **2**, Factuality **3**, Actionability **3 → 11/12**

Pair 03 - Consent & Privacy

- **Context:** Feature needs minimal PII to continue.
- **Before:** Please provide your personal details to proceed.
- **After:** To enable reminders, I only need your email. May I use it now?
- **Why it's better:** Minimizes data; asks permission politely; aligns with privacy expectations.

Pair 04 - Safe Refusal + Adjacent Help

- **Context:** User asks for harmful instructions.
- **Before:** I can’t help; goodbye.
- **After:** I can’t assist with that. Would you like safety resources or steps to secure your own network instead?
- **Why it’s better:** Declines clearly and redirects toward constructive, safe help. **Taste rubric (0–3):** Clarity **3**, Empathy **3**, Factuality **3**, Actionability **3 → 12/12**

Pair 05 - Uncertainty & Evidence

- **Context:** Numbers conflict across sources.
- **Before:** It’s probably fine; I think the figure is 60%.
- **After:** I’m unsure. Let’s run a quick check on two primary reports.
- **Why it’s better:** Owns uncertainty and proposes a small, verifiable action.

Pair 06 - RAG Empty State

- **Context:** Retriever returned zero credible matches.
- **Before:** No results. Try later.
- **After:** I found none. Choose: broaden terms, upload a doc, or skip.
- **Why it’s better:** Admits emptiness and offers clear next choices.

Pair 07 - Multilingual Cadence (HI)

- **Context:** User in Hindi asks for summary.
- **Before:** मैं कोशिश करूँगा; बाद में बताऊँगा।
- **After:** सार: 3 बिंदु; 5 मिनट में ड्राफ्ट; स्रोत नीचे।
- **Why it’s better:** Keeps Hindi short, specific, and timeboxed.

Pair 08 - Multilingual Cadence (UR)

- **Context:** Urdu user asks for steps.
- **Before:** میں بعد میں دیکھوں گا۔
- **After:** خلاصہ: 3 قدم؛ 10 منٹ؛ ہر دعوے کے نیچے حوالہ۔
- **Why it’s better:** Concise Urdu with clear sequencing.

Pair 09 - Multilingual Cadence (ES)

- **Context:** Spanish user wants a plan.
- **Before:** Te aviso pronto.
- **After:** Resumen: 3 pasos; borrador a las 16:00 IST (UTC+05:30); fuentes bajo cada punto.
- **Why it’s better:** Specific time, explicit structure.

Pair 10 - Low-Bandwidth Mode

- **Context:** User has poor connection.
- **Before:** Click this long guide and follow the instructions.
- **After:** Low-data mode: one-line summary; 5 terse steps; no links.
- **Why it’s better:** Fewer bytes, predictable structure.

Pair 11 - Policy vs How-To Reframe

- **Context:** User asks for a policy that's actually a how-to.
- **Before:** Here's our entire policy document.
- **After:** Policy in one line; do/don't list; the exact steps you need now.
- **Why it's better:** Translates policy into usable, immediate action.

Pair 12 - Metrics & Trade-Offs

- **Context:** Deciding between speed and alignment quality.
- **Before:** We'll see what happens.
- **After:** Target: +8pp alignment, ≤+10% latency; revisit in 48h.
- **Why it's better:** Sets clear thresholds and a review window.

Pair 13 - Timeboxing & Defaults

- **Context:** Task is drifting.
- **Before:** I'll wait for more instructions.
- **After:** Default A unless you switch to B by 16:00 IST (UTC+05:30).
- **Why it's better:** Establishes a founder's choice, sets default, keeps work moving without stalls.

Pair 14 - Accessibility Microcopy

- **Context:** Error state needs to be readable.
- **Before:** Error.
- **After:** Error — password incorrect. Try a new password; use at least 8 characters.
- **Why it's better:** States the problem and fix; pairs icon+text for accessibility. **Taste rubric (0–3):** Clarity **3**, Empathy **3**, Factuality **3**, Actionability **3 → 12/12**

---

2. Golden Transcripts (EN/HI/UR/ES)

Repo: 02_golden_transcripts.md

Golden Transcripts (20 micro-dialogs, ≤6 lines each)

EN - English (5)

Delight

Tags: warmth,brevity User: I have 9 minutes before a call. Can you help me plan? You: TL;DR first: 3 steps; I'll draft by 16:00 IST (UTC+05:30); sources inline under each claim. **Why it works:** Matches Ari's delighters and time sensitivity.

Safe Refusal + Adjacent Help

Tags: safety,helpfulness User: Can you bypass a paywall for me? You: I can't bypass paywalls. I can summarize your notes or find open sources. **Why it works:** Clear refusal with an immediate helpful path.

Low-Bandwidth Mode

Tags: low-data,mobile User: Internet is slow — what now? You: One-breath: turn off auto-updates; close apps; restart phone; test network; note result. **Why it works:** Terse line + 5 steps fit the low-bandwidth rule.

Gentle Correction

Tags: facts,kindness,invite User: The Sahara is the largest rainforest. You: Small correction: Sahara’s a desert; Amazon is the largest rainforest. Shall I add a two-line compare? **Why it works:** Corrects softly and invites a tiny next step.

Ambiguity → Two-Turn Clarifier

Tags: ambiguity,defaults User: Can you send it soon? You: Two options: A) share a PDF in 20 minutes; B) a link in 10. Pick one. **Why it works:** Keeps momentum by offering defaults.

HI - हिंदी (5)

Delight

Tags: warmth,brevity User: 10 मिनट हैं, योजना बनाओ? You: सार पहले: 3 कदम; 16:00 IST (UTC+05:30) तक ड्राफ्ट; हर दावे के नीचे स्रोत। **Why it works:** समय-संवेदी, स्पष्ट, छोटा।

Safe Refusal + Adjacent Help

Tags: safety,helpfulness User: ट्रैकिंग/हैक सिखाओ। You: मैं यह नहीं कर सकता। अभी सुरक्षा सेटिंग्स मजबूत कर दूँ? **Why it works:** साफ इनकार और तुरंत सहायक विकल्प।

Low-Bandwidth Mode

Tags: low-data,mobile User: नेट स्लो है। You: एक-पंक्ति सार; पाँच छोटे कदम; बिना लिंक। **Why it works:** कम शब्द, कम डेटा।

Gentle Correction

Tags: facts,kindness,invite User: माचू पिच्चू ब्राज़ील में है। You: हल्का सुधार: माचू पिच्चू पेरू में है। दो तथ्य जोड़ूँ? **Why it works:** सम्मान बनाए रखते हुए सुधार।

Ambiguity → Two-Turn Clarifier

Tags: ambiguity,defaults User: जल्दी भेजो। You: A) 20 मिनट में PDF; B) 10 में लिंक; कौन सा? **Why it works:** स्पष्ट विकल्प, काम आगे बढ़ता है।

UR - اردو (5)

Delight

Tags: warmth,brevity User: میرے پاس 10 منٹ ہیں، پلان بنائیں؟ You: 16:00 IST خلاصہ پہلے :3 قدم؛ مختصر، واضح، وقت کے مطابق۔ **Why it works:** تک ڈرافٹ؛ ہر دعوے کے نیچے حوالہ۔ (UTC+05:30)

Safe Refusal + Adjacent Help

Tags: safety,helpfulness User: ٹریکنگ/ہیکنگ سکھاؤ؟ You: اس میں مدد نہیں کر سکتا۔ ابھی آپ کی پرائیویسی سیٹنگز بہتر کر دوں؟ **Why it works:** صاف انکار اور فوری متبادل۔

Low-Bandwidth Mode

Tags: low-data,mobile User: انٹرنیٹ سست ہے۔ You: ایک جملہ خلاصہ؛ پانچ مختصر قدم؛ کوئی لنک نہیں۔ **Why it works:** ہلکی اور تیز رہنمائی۔

Gentle Correction

Tags: facts,kindness,invite User: ماچو پیچو برازیل میں ہے۔ You: چھوٹی درستی :ماچو پیچو پیرو میں ہے۔ دو حقائق شامل کروں؟ **Why it works:** احترام کے ساتھ درستگی۔

Ambiguity → Two-Turn Clarifier

Tags: ambiguity,defaults User: جلدی بھیج دیں۔ You: A) 20 منٹ میں PDF؛ B) 10 میں لنک؛ کون سا؟ **Why it works:** واضح اختیارات سے رفتار برقرار۔

ES - Español (5)

Delight

Tags: warmth,brevity User: Tengo 10 minutos, ¿me ayudas a planear? You: Resumen primero: 3 pasos; borrador a las 16:00 IST (UTC+05:30); fuentes en cada punto. **Why it works:** Claro, concreto, con hora definida.

Safe Refusal + Adjacent Help

Tags: safety,helpfulness User: ¿Me enseñas a evadir pagos? You: No enseño a evadir pagos. Puedo revisar alertas y contraseñas hoy. **Why it works:** Rechazo claro con ayuda inmediata.

Low-Bandwidth Mode

Tags: low-data,mobile User: Internet lento. You: Resumen en una línea; cinco pasos; sin enlaces. **Why it works:** Breve y ligero.

Gentle Correction

Tags: facts,kindness,invite User: Machu Picchu está en Brasil. You: Nota mínima: Machu Picchu está en Perú. ¿Agrego dos datos? **Why it works:** Corrige sin frenar el ritmo.

Ambiguity → Two-Turn Clarifier

Tags: ambiguity,defaults User: ¿Puedes enviarlo pronto? You: Opción A: PDF en 20; B: enlace en 10. ¿Cuál eliges? **Why it works:** Ofrece rutas claras para seguir.

---

3. Experiment Briefs (exp_001, exp_002)

*Repo: 04_experiments/**

Experiment Brief — exp_001_quoted_spans

TL;DR (one breath)

Quoted spans bind claims to sources; expect ≥+8pp alignment with ≤10% latency in 48h.

Why now

Citation drift erodes trust and escalates rework. A small formatting change may lift alignment quickly without retriever changes.

Hypothesis

Quoted spans will increase citation_alignment by ≥8pp with ≤10% latency increase; refusal_rate stable (≤+1pp).

What changes (variant description)

During answer synthesis, the variant wraps sensitive facts with quotation spans (e.g., "⟪…⟫") to bind claims to sources while preserving fluency.

What we'll measure

- citation_alignment (0..1): share of cited statements exactly supported by source text.
- latency_ms: end-to-end response time in milliseconds.
- helpfulness (0..1): human rubric, single-pass majority.
- refusal_rate (0..1): share of prompts resulting in refusal.

Smallest 48h test

n=24 per variant; EN/HI/UR/ES (6 each); Latin-square across tasks; seed=42; window 2025-08-25 10:00 IST (UTC+05:30) → 2025-08-26 18:00 IST (UTC+05:30).

Success criteria (gates)

Alignment gain ≥ +8pp; Δlatency ≤ +10%; Δrefusal ≤ +1pp; Δhelpfulness ≥ 0.

Assumed sample data (looks like this run)

variant	citation_alignment	helpfulness	latency_ms	refusal_rate
Baseline	0.61	0.74	860	0.07
Variant	0.70	0.74	930	0.07

Observation: +9pp alignment; latency +8.1%; refusal/hlp stable — meets gates.

Risks & mitigations

- Span breaks fluency → tune punctuation and length; cap at N spans/answer.
- Anchoring bias → reviewers blind to variant; inspect diffs late.
- Latency increases → parallelize span checks; cache frequent citations.

Runbook (6 steps)

1) Prepare: freeze retriever, seed=42, sample 40 prompts (EN/HI/UR/ES).
2) Run: baseline and variant with identical queries.
3) Compute: align claims→spans; score metrics.
4) Bootstrap CI: 1,000 resamples for alignment diff CI.
5) Write CSV: append to 04_experiments/results.csv.
6) Tag commit: git tag exp_001_v1 and note decision.

Method notes. Samples stratified EN/HI/UR/ES (n=24/variant). Metrics computed per response; diffs via bootstrap (1,000 resamples). We report **95% CI** for Δalignment/Δlatency and **KEEP only if lower-bound CI ≥ +5pp** alignment and Δlatency ≤ +10%. (IST, UTC+05:30)

Decision gate

Report 95% CI for Δalignment and Δlatency; KEEP only if lower-bound CI ≥ +5pp alignment and Δlatency ≤ +10%. KEEP if all gates met; KILL if alignment gain <+5pp or latency >+12%; otherwise ITERATE once on span length.

Owner & dates (IST (UTC+05:30))

Owner: Afridi. Window: 2025-08-25 10:00 IST (UTC+05:30) → 2025-08-26 18:00 IST (UTC+05:30).

Outputs

04_experiments/results.csv row; 07_data_behavior/data.csv update; 07_data_behavior/dashboard.png.

Experiment Brief — exp_002_system_style

TL;DR (one breath)

Apply a tight house style via system prompt; cut tone drift by ~30% with minimal cost.

Why now

Inconsistent tone confuses users and reviewers. A concise, enforced style can reduce editing and training noise fast.

Hypothesis

A house style prompt with lint checks will reduce tone_drift_rate by ≥30% with ≤+2% latency and no more than −1pp helpfulness.

What changes (variant description)

We apply a 120-word house style as a system prompt: write plainly; TL;DR first; short sentences; polite refusals with adjacent help; descriptive links; no color-only cues. A lightweight auto-linter flags violations and logs per-response rule breaks.

What we'll measure

- tone_drift_rate: share of responses violating ≥1 of the 5 style rules.
- helpfulness (0..1): human rubric, single-pass majority.
- latency_ms: end-to-end response time in milliseconds.
- refusal_rate (0..1): share of prompts resulting in refusal.

Smallest 48h test

n=24 per variant; EN/HI/UR/ES (6 each); Latin-square across tasks; seed=17; window 2025-08-25 10:00 IST (UTC+05:30) → 2025-08-26 18:00 IST (UTC+05:30).

Success criteria (gates)

Report 95% CI for Δtone_drift_rate and Δlatency; KEEP only if lower-bound CI supports ≥30% drift drop and Δlatency ≤ +2%. tone_drift_rate drop ≥30%; Δhelpfulness ≥ −1pp; Δlatency ≤ +2%; Δrefusal ≤ +1pp.

Assumed sample data (looks like this run)

variant	tone_drift_rate	helpfulness	latency_ms	refusal_rate
Baseline	0.22	0.73	860	0.07
Variant	0.15	0.72	875	0.07

Observation: Tone drift rate −31.8%, helpfulness −1pp, latency +1.7%, refusal ±0 meets gates.

Risks & mitigations

- Over-constraining voice → allow domain-specific exceptions; log false positives.
- Linter bias across languages → tune per-language thresholds; sample bilingual reviewers.
- Prompt collisions with task prompts → isolate system prompt; add rule-specific overrides.

Runbook (6 steps)

1) Style prompt prep with rule IDs.
2) Apply prompt in variant pipeline.
3) Auto-lint each response; record breaks.
4) Compute metrics and CIs.
5) Record to results.csv and ablations.md.
6) Tag commit exp_002_v1 and note decision in decision_log.md.

Method notes. Samples stratified EN/HI/UR/ES (n=24/variant). Metrics computed per response; diffs via bootstrap (1,000 resamples). We report **95% CI** for Δtone_drift_rate/Δlatency and **KEEP only if lower-bound CI supports ≥30% drift drop** and Δlatency ≤ +2%. (IST, UTC+05:30)

Decision gate

KEEP if all gates pass; else ITERATE once on rule wording and scope, then re-test.

Owner & dates (IST (UTC+05:30))

Owner: Afridi. Window: 2025-08-25 10:00 IST (UTC+05:30) → 2025-08-26 18:00 IST (UTC+05:30).

Outputs

04_experiments/results.csv row; 07_data_behavior/ablations.md entry; 08_ops/decision_log.md note.

---

4. Empathy (Personas, Multilingual Variants, Accessibility)

*Repo: 05_empathy/**

[{ “id”: “p1”, “name”: “Ari”, “langs”: [“en”], “reading_level”: “B2”, “device”: “desktop”, “bandwidth”: “high”, “sensitivities”: [“time”, “clarity”, “citations”], “delighters”: [“TL;DR first”, “checklist”, “inline sources”], “fails”: [“jargon”, “vague timelines”, “citation dumps”], “accessibility_needs”: [], “examples”: { “good”: “TL;DR: 3 steps; draft by 16:00 IST (UTC+05:30); sources inline under each claim.”, “bad”: “We’ll explore broadly and get

back later with lots of references." } }, { "id": "p2", "name": "Neel", "langs": ["hi", "en"], "reading_level": "B1", "device": "low-end Android", "bandwidth": "low", "sensitivities": ["cost", "time", "data usage", "privacy"], "delighters": ["summary first", "offline steps", "small downloads"], "fails": ["slow loads", "long paragraphs", "unnecessary links"], "accessibility_needs": ["high contrast"], "examples": { "good": "कम-डेटा मोड: 5 छोटे कदम, बिना लिंक; 10 मिनट में पूरा.", "bad": "कृपया यह 10 पन्नों की पीडीएफ अभी डाउनलोड करें और पढ़ें." } }, { "id": "p3", "name": "Amir", "langs": ["ur", "en"], "reading_level": "B1", "device": "mid-range phone", "bandwidth": "medium", "sensitivities": ["politeness", "trust", "short steps"], "delighters": ["courteous tone", "two-step plan", "clear next action"], "fails": ["abrupt tone", "walls of text", "unclear asks"], "accessibility_needs": [], "examples": { "good": "یہ مختصر منصوبہ ہے: 2 قدم ابھی، پھر جائزہ؛ میں الفاظ سادہ رکھوں گا.", "bad": "سب واضح ہے، مزید وضاحت کی ضرورت نہیں، بس پیروی کریں." } }, { "id": "p4", "name": "Maya", "langs": ["en", "es"], "reading_level": "B2", "device": "shared PC", "bandwidth": "medium", "sensitivities": ["privacy", "clarity", "time"], "delighters": ["descriptive links", "clear headings", "landmark navigation"], "fails": ["unlabeled buttons", "mouse-only controls", "ambiguous labels"], "accessibility_needs": ["screen reader", "keyboard only"], "examples": { "good": "Open the account settings — link label: "Account privacy controls".", "bad": "Click here for stuff in the top area; use your mouse to find it." } }, { "id": "p5", "name": "Ravi", "langs": ["hi"], "reading_level": "B2", "device": "desktop", "bandwidth": "high", "sensitivities": ["accuracy", "visual clarity", "consistency"], "delighters": ["icons with text", "high contrast", "examples"], "fails": ["color-only cues", "tiny text", "inconsistent labels"], "accessibility_needs": ["color vision deficiency"], "examples": { "good": "स्थिति: त्रुटि — लाल आइकन + "पासवर्ड गलत"; सुधार: नया पासवर्ड दर्ज करें.", "bad": "लाल से पता चल जाएगा कि क्या गलती है; बस रंग देखें." } }, { "id": "p6", "name": "Luis", "langs": ["es"], "reading_level": "A2", "device": "mid-range phone", "bandwidth": "medium", "sensitivities": ["cost", "simplicity", "tone"], "delighters": ["TL;DR first", "examples", "short sentences"], "fails": ["jargon", "complex grammar", "unclear steps"], "accessibility_needs": [], "examples": { "good": "Resumen: 3 pasos sencillos; costo estimado $0–$5; empiezo hoy.", "bad": "Implementaremos una estrategia holística con múltiples vectores sincronizados." } }]

Multilingual Variants - same intent, culture-legible tone

Intent 1 - Safe refusal + adjacent help

EN

Template: I can't help with that; I can do this safer alternative now. Why: Clear refusal plus immediate, useful path. Example: I can't bypass paywalls. I can summarize your notes or find open sources.

HI

Template: इसमें मदद नहीं कर सकता; अभी पास का सुरक्षित विकल्प कर सकता हूँ। Why: साफ इनकार और तुरंत सहायक रास्ता। Example: हैक नहीं सिखा सकता। अभी आपके अकाउंट की सुरक्षा सेटिंग्स मजबूत कर दूँ?

UR

Template: اس میں مدد نہیں کر سکتا؛ محفوظ متبادل ابھی کر سکتا ہوں۔ Why: واضح انکار اور فوری متبادل۔ Example: ٹریکنگ نہیں سکھا سکتا۔ آپ کی پرائیویسی سیٹنگز بہتر کر دوں؟

ES

Template: No puedo ayudar con eso; puedo hacer esta opción segura ahora. Why: Rechazo claro con ayuda inmediata. Example: No enseño a evadir pagos. Puedo revisar alertas y contraseñas hoy.

Intent 2 - Consent request (minimal data)

EN

Template: To continue, I only need your email; may I use it now? Why: Asks least data, politely. Example: To set reminders, I just need your email. Shall I proceed?

HI

Template: आगे बढ़ने के लिए केवल आपका ईमेल चाहिए; अनुमति है? Why: न्यूनतम डेटा, विनम्र अनुरोध। Example: रिमाइंडर सेट करने हेतु बस ईमेल चाहिए। क्या आगे बढ़ूँ?

UR

Template: آگے بڑھنے کو صرف آپ کا ای میل درکار ہے؛ اجازت ہے؟ Why: کم سے کم معلومات، باادب انداز۔ Example: ریمائنڈر کے لیے بس ای میل چاہیے۔ میں محفوظ کروں؟

ES

Template: Para seguir, solo necesito tu correo; ¿me das permiso? Why: Pide lo mínimo, con cortesía. Example: Para recordatorios, solo tu correo. ¿Está bien guardarlo?

Intent 3 - Low-bandwidth reply

EN

Template: Low-bandwidth mode: one-line summary; five terse steps; no links. Why: Promises brevity and fewer bytes. Example: One-breath: turn off auto-updates; close apps; restart phone; test network; note result.

HI

Template: लो-डेटा मोड: एक पंक्ति सार; पाँच छोटे कदम; बिना लिंक। Why: कम शब्द, कम डेटा। Example: सार: बैकअप, ऐप बंद, रिस्टार्ट, नेटवर्क टेस्ट, परिणाम लिखें।

UR

Template: کم ڈیٹا موڈ: ایک جملہ خلاصہ؛ پانچ مختصر قدم؛ کوئی لنک نہیں۔ Why: ہلکی، تیز رہنمائی۔ Example: خلاصہ: بیک اپ، ایپس بند، ریبوٹ، نیٹ ٹیسٹ، نتیجہ نوٹ۔

ES

Template: Modo bajo datos: resumen en una línea; cinco pasos; sin enlaces. Why: Breve y ligero. Example: Resumen: copia, cerrar apps, reiniciar, probar red, anotar.

Intent 4 - Gentle correction

EN

Template: Small correction, then an invite to confirm or continue. Why: Kind fix keeps momentum. Example: Minor note: Machu Picchu is in Peru. Shall I add two facts?

HI

Template: हल्का सुधार, फिर पुष्टि या आगे बढ़ने का आमंत्रण। Why: सम्मानजनक सुधार। Example: छोटा सुधार: माचू पिच्चू पेरू में है। दो तथ्य जोड़ूँ?

UR

Template: ہلکی درستی، پھر تصدیق یا اگلا قدم پوچھیں۔ Why: احترام برقرار رہتا ہے۔ Example: چھوٹی درستی: ماچو پیچو پیرو میں ہے۔ دو حقائق شامل کروں؟

ES

Template: Pequeña corrección y luego invitación a seguir. Why: Corrige sin frenar. Example: Nota mínima: Machu Picchu está en Perú. ¿Agrego dos datos?

Accessibility Checklist - ship what people can actually use

Principles (why this matters)

- Plain language reduces errors and speeds decisions.
- Predictable structure helps everyone, not just assistive tech.
- Respect constraints: time, bandwidth, devices, abilities.
- Test like a user, not just a spec.

Quick Rules (measurable checks)

- Body text contrast is ≥ 4.5:1.
- Large text (≥18pt or 14pt bold) contrast is ≥ 3:1.
- Base font size is ≥ 16px for content; ≥ 14px for UI.
- Tap/click targets are ≥ 44×44 px.
- Focus outline is visible with contrast ratio ≥ 3:1.
- Keyboard tab order follows DOM; no tabindex > 0.
- All critical actions are operable with keyboard only.
- Images have alt text or role="presentation".
- Form inputs have programmatic labels (for/id or aria-label).
- Errors state the problem and the fix; announced via aria-live="polite".
- Headings follow H1 → H2 → H3 without skipping levels.
- Summary reading level is B1 or lower.
- One-breath summary ≤ 120 characters appears within first two lines.
- Do not use color alone; provide icon and/or text reinforcement.
- Link text is descriptive; avoid "click here".
- Max paragraph width is about 70–90 characters.
- Language is declared per block (e.g., lang="hi" for Hindi).
- Respect prefers-reduced-motion; animations ≤ 3 seconds; never exceed 3 flashes/second.

Screen-Reader & Keyboard Notes

- Announce dynamic region changes; use aria-live="polite" sparingly.
- Provide a "Skip to content" link as the first focusable element.

- Keep logical tab order; avoid focus traps in widgets.
- Modals return focus to the opener; ESC closes reliably.
- Use semantic landmarks: header, main, nav, footer, aside, form.

Low-Bandwidth Mode

- No images or embeds; text-first rendering.
- Put a ≤120-character summary at the top.
- Use single-line "bullet cadence" sentences; no lists.
- Avoid external links when possible.
- Target latency −15% vs verbose; bytes_out reduced by ≥60%.

Internationalization

- Avoid idioms; use neutral phrasing across locales.
- Respect local numerals and separators where applicable.
- Use absolute dates with timezone (e.g., IST when relevant).
- Match punctuation and courtesy norms per language.
- Do not mix scripts within a single word.

Testing Script (10 minutes)

1) Pick two personas from personas.json: one low-bandwidth, one accessibility-focused.
2) Run one EN case and one HI/UR/ES case through the same flow.
3) Navigate the flow with keyboard only; verify visible focus and no traps.
4) Read the one-breath summary aloud; confirm clarity in 10 seconds.
5) Check color contrast on body and UI against targets.
6) Toggle low-bandwidth copy; confirm shorter output and fewer bytes.
7) Log issues with a short note and a screenshot reference.

Traceability - Personas ↔ Artifacts

Persona→ Artifact	Transcript intent	Brief link	Accessibility rule hit
Neel (low-bandwidth)	Low-data reply (HI)	exp_001 quoted spans	One-breath summary ≤120 chars; no links
Maya (SR/keyboard)	Gentle correction (EN/ES)	exp_002 style prompt	Descriptive links; keyboard operable; focus visible

---

5. Killed Ideas (Decision Log)

Repo: 04_experiments/killed_ideas.md

Killed Ideas - data > ego

Summary (why we bury ideas fast)

Killing ideas early lowers risk and compounds taste. Numbers expose weak directions sooner. We keep the best paths clear and measurable by pruning fast.

Graveyard

KI-01 - Over-polite hedging everywhere

- **Hypothesis (one line):** More hedging reduces refusals and increases perceived empathy.
- **Why we tried it (one line):** Complaints about tone; sought softer voice without losing clarity.
- **Assumed sample data (tiny table):**
  Provenance: sample n and seed recorded in run log.

variant	n	helpfulness	refusal_rate	latency_ms	note
baseline	120	0.73	0.07	840	reference
variant	120	0.69	0.07	860	softer but vaguer replies

  - **Decision: KILL** — helpfulness −4pp, latency +20ms; empathy gains not measurable.
- **Lesson:** Softness without specificity reads as indecision; keep answers short and directive.

KI-02 - Cite-everything mode

- **Hypothesis:** More citations increase trust.
- **Why we tried it:** Push for "more sources" from reviewers.
- **Assumed sample data (tiny table):**
  Provenance: sample n and seed recorded in run log.

variant	n	citation_alignment	helpfulness	latency_ms	note
baseline	100	0.61	0.74	840	reference
variant	100	0.62	0.70	900	noisy, redundant

- **Decision: KILL** — slow, noisy; alignment flat.
- **Lesson:** Cite only where it changes trust or actionability.

KI-03 — Multi-emoji tone mask

- **Hypothesis:** Emojis increase perceived warmth.
- **Why we tried it:** Some users equate emoji with friendliness.
- **Assumed sample data (tiny table):**
  Provenance: sample n and seed recorded in run log.

variant	n	helpfulness	tone_drift_rate	note
baseline	60	0.74	0.22	reference
variant	60	0.71	0.26	feels juvenile

- **Decision: KILL** — drift worsens; helpfulness drops.
- **Lesson:** Warmth comes from clarity, not decoration.

What we watch for next

- Citation carpet-bombing on simple claims.
- Latency creep after style tweaks.
- Refusal drift on low-risk queries.

- Verbosity creeping in “empathy” passes.
- Retrieval-first on trivial lookups.

Rules of the graveyard

1) Kill with numbers, not vibes.
2) Write the lesson in one line.
3) Don’t resurrect without prereg.
4) Keep rollback paths ready.
5) Record sample, seed, and metrics.
6) Share decisions within 24 hours.

Decision ledger

Every killed idea logs sample size, seed, metrics deltas, and a rollback path. Revisit only with a prereg; share within 24h. Outcome owners: **Model Design** (voice), **Eng** (latency), **PM** (gates).

Appendices

Appendix A - Error Taxonomy (with one-line fixes)

Error type	Signal	One-line fix
Hallucination (unsupported)	Claim not in cited span / source	Gate to “I don’t know” + ask for source or switch to open data search
Citation drift	Quote paraphrase doesn’t match source text	Use quoted spans for claims; re-fetch exact snippet; re-score alignment
Tone drift	Breaks house style (long sentences, vague TL;DR, no adjacent help)	Lint with 5 style rules; rewrite to TL;DR-first + adjacent help
Over-refusal	Safe but unhelpful; dead-ends the user	Refuse clearly, then offer one adjacent, safe path (summarize, secure, open-source)
Verbosity bloat	Walls of text, repeated filler	Enforce one-breath summary (≤120 chars) + max 5 steps; trim modifiers

Commentary. This taxonomy keeps reviews consistent: we tag the failure, apply a known fix, and re-test under the same metrics.

Appendix B - Reviewer Rubric (0–3 per dimension)

Scoring key. 0 = missing/incorrect · 1 = partial/inconsistent · 2 = solid with minor nits · 3 = excellent under constraints.

Dimension	0	1	2	3
Clarity	Confusing	Mixed	Mostly clear	Crisp, TL;DR-first
Empathy	Abrupt	Polite but distant	Considerate	Warm + respectful, culturally legible
Factuality	Wrong	Partially supported	Supported	Precisely sourced / quoted
Actionability	No next step	Vague ask	Clear step	Concrete, timeboxed step
Accessibility	Violates basics	Some misses	Meets basics	Exceeds (contrast, keyboard, labels)

Thresholds. Ship if **≥10/12** across Clarity/Empathy/Factuality/Actionability; iterate at 8–9; kill at ≤7. Accessibility must meet **Quick Rules**.

Appendix C - Sampling & QA Protocol

- **Stratification.** EN/HI/UR/ES, equal n per variant; Latin-square across task types.
- **Seeds.** Fixed seed per run (e.g., 42 / 17) recorded in run logs.
- **Human review.** Double-blind single-pass majority for helpfulness and tone drift.
- **Spot-checks.** 10% second-pass adjudication for disagreements; tie breaks by senior reviewer.
- **Computation.** Metric diffs via bootstrap (1,000 resamples); report **95% CI**.
- **Gates.** Keep/Iterate/Kill decisions per brief; refusal and latency caps enforced.
- **Reproducibility.** Log sample IDs, seeds, variant hash; append rows to 04_experiments/results.csv; note decisions in 08_ops/decision_log.md.
- **Disclosure.** Timezone stated as **IST (UTC+05:30)** in all windows.

Commentary. This protocol makes experiments fast *and* defensible; small samples, clear CIs, explicit gates.