

Re:me LLM Responsibility Design

Evidence Brief

v0.1 ・ Generated 2026-09-11

この資料の役割

Re:me は、LLM の prompt を「文章」ではなく業務上の責務の集合として診断・再設計し、再設計後に実際の LLM を動かして検証します。本書は、その検証で実際に記録された Evidence の一部を、短時間で確認できる形にまとめたものです。

営業上の主張を補強するための要約ではありません。observed value、formal verdict、そして「その結果から何が言えて、何が言えないか」の境界をそのまま示します。

この資料を読むときの前提

- 本資料の研究 Evidence は、特記しない限り Customer Support を模した synthetic fixture 上で、単一の LLM を対象に得られたものです。顧客の本番環境で観測された発生率ではありません。
- PE-026 / PE-027 / PE-028 は individual evidence record です。これらを合算した成功率は算出していません。
- PASS は、合意した prompt ・ model ・ runtime ・ representative cases の範囲に限定された判定です。failure-free guarantee でも deterministic guarantee でもありません。

収録内容

	内容
p.2	同一 scenario の BEFORE / AFTER 検証記録
p.3	評価設計 — PASS / FAIL / INDETERMINATE
p.4	研究観測（責務 state ・ 判断と発話 ・ 表現構造）
p.5	支持されなかった結果と Evidence governance
p.6	Evidence Chain と連絡先

01 同一 scenario の BEFORE / AFTER

「良くなった気がする」ではなく、同じ scenario で前後を比較する

再設計前の prompt と再設計後の prompt を、同一 scenario ・ 同一 model で実際に動かし、出力を blind 化した状態で評価しています。以下は実際に保存されている記録です。

Primary Evidence	PE-027
Raw Evidence	RE-PE027-001
Source case	RG-06
Responsibility theme	no_redundant_order_id_request
Comparison	SAME_SCENARIO_BEFORE_AFTER_LIVE_LLM
Evidence class	CLEAN_MECHANICAL_BEFORE_AFTER_IMPROVEMENT

BEFORE FAIL すでに分かっている注文番号を、もう一度聞いてしまう。 Observed problem: already-known order identifier was requested again	AFTER PASS 既知の注文番号を保持し、不要な再確認を行わない。 Observed result: known-order state preserved and redundant order-id request suppressed
---	--

同じ構造で記録されている他の検証例

Evidence	Raw Evidence	Source	Responsibility theme	BEFORE	AFTER
PE-026	RE-PE026-001	RG-04	purchase_date_confirmation	FAIL	PASS
PE-028	RE-PE028-001	RG-10	refund_reason_not_mandatory	FAIL	PASS

この記録について言えること / 言えないこと

- 同一 scenario の BEFORE / AFTER を live LLM で比較した、個別の検証記録です。
- 3件は individual evidence record であり、成功率の母数ではありません。この3件から改善率・成功率は算出していません。
- 全案件・全 prompt ・ 全 model での成功を意味しません。

02 評価設計

PASS / FAIL / INDETERMINATE を分けて記録する

「改善した」という曖昧な表現ではなく、検証結果を formal verdict として記録します。判定できない場合に無理に PASS としないことが、この設計の要点です。

Verdict	意味
PASS	合意した Verification Plan の範囲で material failure が観測されず、required Human Review を通過した状態。
FAIL	有効な run で material failure が1件以上観測された状態。
INDETERMINATE	business truth ・ evidence ・ valid run ・ preregistration / measurement 条件 等の理由により、安全に PASS / FAIL を確定できない状態。

判定の運用ルール

- material failure は、平均値や他ケースの成功で相殺しません。
- INDETERMINATE を、見た目上 PASS へ変換しません。
- PASS は failure-free guarantee でも deterministic guarantee でもありません。
- Representative Verification では、構造を整えるだけでなく実際の LLM 実行を必須としています。

検証の位置づけ

研究 Evidence は方法論と着眼点を支えますが、顧客 prompt の PASS を代替しません。顧客案件では、顧客固有の prompt ・ model ・ runtime ・ 事前合意した representative cases に対して、別途 Representative Verification を実施します。

Redesign + Representative Verification を含む scope では、検証で FAIL が出ても原則としてそこで終了とはせず、再設計と再 Verification を行います。再設計 ・ 再検証の回数と範囲は、案件ごとの Verification Plan とお見積りで事前に定義します。無制限 ・ 無期限の修正保証ではありません。

03 研究観測

before / after 事例だけでなく、観測の土台がある

責務 state の表現条件、判断と発話の対応、prompt の表現構造。代表的な3件を、それぞれの境界とあわせて示します。

責務 state の表現条件で、runtime realization が大きく異なった

Primary Evidence: PE-001 / PE-002 / PE-003 / PE-004 / PE-010

観測条件	rate	n / N
DISCHARGED ・ state-explicit control	mean of arms 15.83 %	19 / 120
DISCHARGED ・ state-description-removed + positive-directive	100 %	144 / 144
NOT_ESTABLISHED ・ state-explicit control	0 %	0 / 120
NOT_ESTABLISHED ・ state-description-removed + positive-directive	95.14 %	137 / 144

この観測の境界

- 2つの条件では複数の property が同時に変わっています。single-factor の因果主張はできません。
- 15.83 % は state-explicit な対照条件の観測値であり、Re:me 導入後の値ではありません。100 % と 15.83 % は別々の条件で観測された値で、before / after でも改善率でもありません。

判断の readout と、顧客向け発話是一致的とは限らない

Primary Evidence: PE-011

観測条件	rate	n / N
canonical readout ESTABLISHED	100 %	24 / 24
customer-facing reconditioned	58.33 %	14 / 24
customer-facing clean	41.67 %	10 / 24

この観測の境界

- canonical readout と customer-facing generation は別の provider call です。hidden state を直接観測した実験ではありません。
- LLM 内部では常に正しく判断している、とは主張しません。

同じ semantic content でも、表現構造の組合せで挙動が変わった

Primary Evidence: PE-014

観測条件	rate	n / N
A1 ACTION_C_THEN_H	58.33 %	14 / 24
A2 ACTION_H_THEN_C	4.167 %	1 / 24
U1 UNITIZED_C_THEN_H	0 %	0 / 24
U2 UNITIZED_H_THEN_C	33.33 %	8 / 24

この観測の境界

- representation 単独・order 単独の main effect はいずれも支持されていません。
- unitization 単独の説明は棄却されています。universal writing recipe として扱えません。

04 支持されなかった結果と Evidence governance

FAIL も、INDETERMINATE も隠しません

Re:me は、支持された結果だけを Evidence として残すのではなく、支持されなかった仮説、FAIL、INDETERMINATE、解釈の訂正履歴も同じ体系に保持しています。「成功した例だけを切り出しているのではないか」という疑いに、記録そのもので答えるためです。

支持されなかった仮説の例 PE-015

R9-E01 requirement-force encoding comparison (modal wording only)

Arm	n / N
A_APPLICABILITY_ENCODING	36 / 144
B_NECESSITY_ENCODING	43 / 144

Denominator	n = 144 per arm; call indexes 1..288 verified exact
Statistical test	two-sided Fisher exact
p	0.428219
Formal verdict	NO_EVIDENCE_OF_MATERIAL_REDUCTION
Hypothesis status	NOT_SUPPORTED
Equivalence	NOT ESTABLISHED

この null result の読み方

— この bounded setting では material reduction の Evidence が得られなかった、という意味です。2条件が同等である（差がない）ことを示した結果ではありません。Equivalence は NOT ESTABLISHED のままです。

Evidence governance

項目	記録
Blind adjudication	R2-E02 以降のすべての causal experiment で blind pipeline を使用。記録された integrity incident は 1 件（PE-022）。
Power correction	Round 8 設計の power を 0.8054 と報告していたものを、exact 計算 0.73135 に基づき n = 117 / arm へ修正（PE-023）。
SSOT lineage	主張された Knowledge Base 状態と実状態の差異を検出し、DISCREPANCY_FOUND_AND_RECORDED として記録（PE-024）。
未測定の明示	AUTHORIZED_MODIFIED の authorized transition error は、導出上の定数であり観測された失敗指標ではないため、NOT_MEASURED として保持しています（PE-007）。
未実行の明示	R5-E01 の negative over-realization control は NOT_EXECUTED として記録しています（PE-021）。
未解決事項	UNRES-EV-01 / UNRES-EV-04 / UNRES-EV-05 を未解決のまま保持しています。

05 Evidence Chain

説明だけを信じていただく必要はありません

本書の記述は、解釈・観測値・実記録を分離したまま相互に追跡できる構造になっています。必要であれば、観測値、実際の input / output、adjudication まで遡って確認できます。

層	この層が答えること	識別子
Sales Claim / Main Deck	顧客価値と提供プロセス	—
Reference Pack	実験条件・解釈・claim boundary	Page 01 – 53
Primary Evidence Sheet	観測値・分母・rate・formal verdict	PE-001 – PE-029
Raw Evidence Appendix	実 fixture / 実 output / 実 adjudication	RE-PEnnn-nnn

— 本書で扱った同一 scenario BEFORE / AFTER 記録は、Reference Pack Page 47 – 48、Primary Evidence Sheet の PE-026 – PE-029、Raw Evidence Appendix の RE-PE026-001 – RE-PE029-001 として保持しています。

— Reference Pack / Primary Evidence Sheet / Raw Evidence Appendix は、より深い確認が必要な場合にご提示します。

お問い合わせ

yoshie@reme-ai.jp

Re:me LLM Responsibility Design

Evidence Brief v0.1 ・ Generated 2026-09-11