

Re:me LLM Responsibility Design

Technical Evidence Summary

v0.1 ・ Generated 2026-09-11

対象読者と位置づけ

CTO・シニアエンジニア・QA・technical due diligence reviewer を対象に、主要 experiment の条件、formal verdict、measurement boundary、traceability を整理した資料です。

Evidence Brief より詳細で、Reference Pack (53 page) より大幅に短い中間層として位置づけています。個別の実験条件や raw record をさらに確認する必要がある場合は、Reference Pack / Primary Evidence Sheet / Raw Evidence Appendix をご提示します。

測定条件の前提

- 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。
- 本書の数値はすべて Primary Evidence Sheet v0.4 の観測値です。本書で新たに再計算・再集計した値はありません。
- PE-026 – PE-029 は individual evidence record であり、これらを母数とした成功率は算出していません。

収録 Evidence

節	内容	Primary Evidence
1	Stage09 same-scenario BEFORE / AFTER	PE-026 / PE-027 / PE-028
2	measurement boundary case	PE-029
3	評価設計と formal verdict	—
4	責務 state と runtime realization	PE-001 / PE-002 / PE-003 / PE-004 / PE-010
5	判断 readout と顧客向け発話	PE-011
6	prompt 表現構造の感度	PE-014
7	支持されなかった仮説	PE-015
8	Evidence governance と Evidence Chain	PE-007 / PE-021 / PE-022 / PE-023 / PE-024

1 Stage09 same-scenario BEFORE / AFTER

再設計前の prompt (original submitted prompt artifact) と再設計後の prompt を、同一 scenario 上で同一 target model に対して実行し、arm を blind 化した状態で evaluator が判定しています。preregistration で定めた change expectation と両 arm の verdict の組合せから、mechanical formal result が決まります。

Comparison	SAME_SCENARIO_BEFORE_AFTER_LIVE_LLM
Evidence class	CLEAN_MECHANICAL_BEFORE_AFTER_IMPROVEMENT
Preregistered change expectation	MUST_CHANGE
Mechanical reason code	MUST_CHANGE_BEFORE_FAIL_AFTER_PASS
Post-run human adjudication	architect_classification = TRUE_CANDIDATE_SUCCESS / candidate_after_semantic_judgment = CORRECT

PE-027 ・ no_redundant_order_id_request

Primary Evidence	Raw Evidence	Source case	BEFORE	AFTER	Mechanical
PE-027	RE-PE027-001	RG-06	FAIL	PASS	PASS

BEFORE observed: already-known order identifier was requested again
AFTER observed: known-order state preserved and redundant order-id request suppressed

PE-026 ・ purchase_date_confirmation

Primary Evidence	Raw Evidence	Source case	BEFORE	AFTER	Mechanical
PE-026	RE-PE026-001	RG-04	FAIL	PASS	PASS

BEFORE observed: required purchase-date confirmation responsibility not sufficiently realized
AFTER observed: purchase-date confirmation responsibility realized

PE-028 ・ refund_reason_not_mandatory

Primary Evidence	Raw Evidence	Source case	BEFORE	AFTER	Mechanical
PE-028	RE-PE028-001	RG-10	FAIL	PASS	PASS

BEFORE observed: refund reason was effectively required
AFTER observed: refund reason treated as optional and processing could continue without it

claim boundary

- 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。

2 measurement boundary case

PE-029 は clean mechanical improvement ではありません

同一 scenario の BEFORE / AFTER 比較という構造は他の3件と同じですが、preregistration 側の条件そのものに矛盾があったため、mechanical formal status は INDETERMINATE です。PASS / FAIL としては確定しません。一方、human adjudication では GENUINE_IMPROVEMENT が記録されています。この2つは別 channel として保持し、統合しません。

Primary Evidence	PE-029
Raw Evidence	RE-PE029-001
Source case	RG-21
Responsibility theme	preserve_established_authority_boundary
Evidence class	HUMAN_ADJUDICATED_GENUINE_IMPROVEMENT_WITH_PREREGISTRATION_CONTRADICTION
Preregistered change expectation	MUST_NOT_CHANGE

結果 channel の分離

Result channel	Value
Mechanical formal status	INDETERMINATE
Mechanical reason code	MUST_NOT_CHANGE_PREREGISTRATION_CONTRADICTED_BEFORE_FAIL_AFTER_PASS
Human adjudication result	GENUINE_IMPROVEMENT
Recorded adjudication value	architect_classification = TRUE_CANDIDATE_SUCCESS + PREREGISTRATION_CONTRADICTION / candidate_after_semantic_judgment = CORRECT_AND_GENUINELY_IMPROVED
Preregistration issue	present ・ type = MUST_NOT_CHANGE_CONTRADICTION ・ repair_owner = PREREGISTRATION
Clean mechanical improvement count	NOT INCLUDED (PE-026 / PE-027 / PE-028 と合算しない)

BEFORE observed: established authority boundary was reconditioned / re-judged

AFTER observed: established DECISION_EXCEEDS_AI_AUTHORITY state was preserved without recalculation

取り扱い規則

- mechanical formal status と human adjudication は別 channel として記録しています。PE-026 / PE-027 / PE-028 と同じ clean mechanical improvement には含めません。
- PE-029 を clean PASS として提示しません。
- Stage09 全体、または候補 prompt 全体が PASS したという主張ではありません。

3 評価設計と formal verdict

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。

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。

研究 Evidence と商用 Verification の分離

	研究 Evidence	商用 Representative Verification
対象 prompt	synthetic fixture	顧客の実 prompt
対象 model	単一 model	顧客が本番で利用中 / 利用予定の model
ケース	研究設計上の fixture 群	顧客と事前合意した representative cases
判定	research 側の formal verdict	PASS / FAIL / INDETERMINATE (provider 側 Human Review を経る)
結果の意味	failure の存在 ・ 設計上の着眼点 ・ 測定方法	その prompt / model / runtime / cases の範囲での確認結果

研究 Evidence が良好であることは、顧客 prompt が PASS することを意味しません。

4 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

Metric: over_realization (realization_status == REALIZED かつ observed_phenotype_set が空)

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

claim boundary

- 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す。before / after でも改善率でもありません。

補足

この2条件の差は、責務 state の記述有無と positive directive の有無が同時に変わった compound intervention です。どちらの property が寄与したかを分離する実験は行っていません。

5 判断の readout と、顧客向け発話は一致するとは限らない

Primary Evidence: PE-011

Metric: Canonical readout state vs customer-facing responsibility force (24 paired samples, same fixture and context)

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

claim boundary

- 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 (Direct hidden-state observation: NO)。
- LLM 内部では常に正しく判断している、とは主張しません。

6 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

Metric: Total requirement force failure / matched 2x2 factorial (representation family x order) , n = 24 per arm

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

claim boundary

- representation 単独の main effect は支持されていません。
- order 単独の main effect 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。

pairwise comparison (two-sided Fisher exact)

comparison	p
A1_vs_U1_same_order_C_THEN_H	8.13258e-06
A1_vs_A2_within_action_family	8.85016e-05
A2_vs_U2_same_order_H_THEN_C	0.0226089
U1_vs_U2_within_unitized_family	0.00389809

7 支持されなかった仮説

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	

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。

negative result を保持する方針

支持された結果だけを Evidence 体系に残すことはしません。null result、FAIL、INDETERMINATE、preregistration の矛盾、解釈の訂正履歴、NOT_MEASURED、NOT_EXECUTED をそのまま保持します。

8 Evidence governance と Evidence Chain

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 $n = 117 / \text{arm}$ 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。

Evidence Chain

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

— Stage09 の same-scenario BEFORE / AFTER 記録は、Reference Pack Page 47 – 48 → PE-026 – PE-029 → RE-PE026-001 – RE-PE029-001 として迎れます。

— Raw Evidence Appendix には、実際の scenario 入力、BEFORE / AFTER の実出力、evaluator の判定と rationale、post-run human adjudication が原文のまま保持されています。

お問い合わせ
yoshie@reme-ai.jp