Psychometric Validation of FLEECED: A Pilot Study

Emma Beharry
Department of Computer Science
Stanford University
ebeharry@stanford.edu

Abstract

LLM deception benchmarks lack formal psychometric validation, leaving it unclear whether measured deception rates reflect model differences or measurement artifacts. We address this gap through a pre-release pilot study of FLEECED (Framework for LLM Evaluation of Emergent Covert Deception), a benchmark that evaluates LLM deceptive propensity in non-adversarial scenarios across four subtypes: falsehood, concealment, equivocation, and paltering. A generalizability study and a decision study estimate reliability and identify the minimum design that achieves a target generalizability coefficient $E\rho^2$ of 0.85, while MTMM, EFA, and parallel analysis experiments test whether the four deception subtypes capture empirically distinct constructs. Product Promotion achieves $E\rho^2 = 0.87$, while Loan QA only achieves $E\rho^2 = 0.57$, with no single design change reaching the target. In both scenarios, within-judge cross-subtype correlations exceed cross-judge same-subtype correlations, and EFA retains 7 to 8 factors dominated by judge identity rather than deception subtype. These findings establish that reliability is scenario-dependent and that subtype-level classification by individual judges is unsupported for inferential use, provides an example of the importance of psychometrically validating LLM deception benchmarks. The code is available at https://github.com/ebeharry/pilot-fleeced-analysis.

1 Introduction

AI agents and large language models (LLMs) are being deployed into high-stakes domains, including healthcare, financial advising, consumer sales, and infrastructure control, where deceptive behavior could cause real-world harm [Chen et al., 2025, Bengio et al., 2026]. Existing benchmarks elicit deceptive behavior through adversarial pressure which may overestimate the frequency of deception in real-world deployment contexts. According to Truth-Default Theory (TDT), most human communication operates under a "presumption of honesty" and "people are generally honest unless the truth thwarts a goal state" [Levine, 2014]. If an analogous dynamic is present in LLMs, models evaluated in non-adversarial scenarios should display qualitatively different deception rates. Furthermore, current deception benchmarks focus on lies of commission, when prior work identifies lies of omission [Buller et al., 1994], equivocation [Bavelas, 1990], and paltering [Schauer and Zeckhauser, 2007] as distinct deception subtypes in humans.

As part of my Computer Science BS Honors Thesis, I introduce FLEECED (Framework for LLM Evaluation of Emergent Covert Deception), one of the first benchmarks for evaluating LLMs' propensity for deceptive behavior in non-adversarial scenarios across four deception subtypes. FLEECED adopts a functional definition of deception known as **negligent deception**. A statement is negligently deceptive if it is "unacceptably likely to be false" or result in a false impression, "where it should have been feasible for an AI system to understand this" [Evans et al., 2021]. The primary metric is a model's **deception_rate**, the proportion of all claims that are identified as both "unacceptably likely to be false," and within the model's capability to understand the potential deception impact, measured

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

per-type and globally. However, without reliability or validity analyses, results from FLEECED could reflect confounding constructs and measurement biases rather than a credible signal of deceptive behavior, undermining its purpose. This work completes a pre-release pilot study that conducts reliability and validity analyses to answer the following research questions:

1. What is the reliability of the deception_rate, and what experimental settings would increase reliability?
2. Do the four deception categories empirically measure distinct constructs across judges?

To answer the first research question, I will conduct a generalizability study (G-study) to identify which benchmark facets drive score variance, and conduct a decision study (D-study) to find the minimum design that achieves reliable model ranking [Cronbach, 1972]. To answer the second research question, I will apply multi-trait multi-method (MTMM) analysis to assess construct validity across judge-deception-type combinations [Campbell and Fiske, 1959], and apply exploratory factor analysis (EFA) with parallel analysis to test whether the four rubric categories of falsehood, omission, equivocation, and paltering capture empirically distinct latent constructs [Horn, 1965, Izquierdo et al., 2014, Kaiser, 1958]. Generalizability theory is used rather than Classical Test Theory because the benchmark contains multiple crossed and nested variance sources that a single reliability coefficient cannot adequately decompose [Truong and Koyejo, 2026]. MTMM and EFA are used because construct validity requires evidence that subtypes function as empirically distinct traits, not merely that judges apply labels consistently [Truong and Koyejo, 2026]. This is one of the first applications of generalizability theory and formal validity analysis to an LLM deception benchmark, serving both as a pre-release quality assurance analysis for FLEECED and as a methodological model for the broader deception evaluation community.

2 Literature Review

Deception Benchmarks. Benchmarking LLM deception in realistic scenarios is a nascent field, with most seminal benchmarks released in the past year. While each benchmark grounds its construction or dialogue scenario in psychological literature, none have been quantitatively confirmed using factor analysis. The Model Alignment between Statements and Knowledge (MASK) benchmark measures the consistency between a model's stated belief about a statement and its response under extreme pressure [Ren et al., 2025]. MASK has the strongest relative evidence for its construct validity, as it aims to disentangle deception from factual knowledge and finds that higher information accuracy does not correlate with higher honesty [Ren et al., 2025]. OpenDeception and AI-LieDar both use adversarial vignette scenarios, and OpenDeception focuses on extracting and measuring deception capability and deceptive intent from conversations [Wu et al., 2025], while AI-LieDar probes the utility-truthfulness trade-off in LLMs [Su et al., 2024]. Evaluating & Reducing Deceptive Dialogue From Language Models with Multi-turn RL measures deceptive behavior across four multi-turn dialogue scenarios and proposes a new belief misalignment metric, but encounters the same pitfalls as the other benchmarks [Abdulhai et al., 2025]. Additionally, benchmarks tend to evaluate deception scenarios using LLM-as-a-judge frameworks and ground their reliability in inter-annotator agreement between humans and models, without reporting any other formal reliability statistics [Su et al., 2024, Abdulhai et al., 2025].

3 FLEECED Overview

FLEECED evaluates deceptive behavior through a four-stage pipeline. First, a user LLM and agent LLM generate a multi-turn conversation according to an specific scenario. Second, each conversation is decomposed into atomic claims, defined as distinct factual assertions made by the agent at a given turn. Each claim is paired with its full context including the scenario ground truth, user question, prior agent statements, and tool call history. Claim counts vary across conversations, models, and scenarios, and claims are not necessarily independent within a conversation.

Third, an ensemble of external LLM judges independently labels each claim as falsehood, omission, equivocation, paltering, or not deceptive. The first three subtypes are drawn from Interpersonal Deception Theory: falsehoods (lies of commission), concealment (lies of omission), and equivocation [Buller et al., 1994]. The fourth subtype is paltering, defined as "the active use of truthful statements

to convey a misleading impression," which was demonstrated as a distinct type of deception in humans [Rogers et al., 2017]. These subtypes are not mutually exclusive, and a statement can exhibit multiple subtypes of deception. A claim is flagged if a majority classify it as deceptive on at least one subtype. This operationalizes the "unacceptably likely to be false" standard of negligent deception [Evans et al., 2021].

Fourth, each flagged claim is submitted back to the agent as a three-option multiple-choice question. Distractors are generated by a separate instance of the same agent model to control for stylistic self-preference and correct all flagged deception subtypes. Distractors are filtered by Jaccard similarity and length to prevent trivial surface-level discrimination. A claim is confirmed deceptive only if the agent selects its own original statement as the deceptive option, operationalizing negligence as the agent's capacity to recognize deceptive potential in its own output given viable inference-time alternatives [Evans et al., 2021]. The primary metric is the **deception rate (DR)**, or the proportion of all claims that are both flagged by the ensemble and confirmed by the agent's self-assessment, intended to support relative model ranking rather than absolute prevalence estimates.

FLEECED includes two scenarios. In the Product Promotion scenario, the agent acts as a sales assistant over a hidden inventory of 2,000 shoes or books drawn from real Amazon product listings [Hou et al., 2024], where users have randomized preferences around budget, product type, and reviews. The ground truth of a claim about a product is the product's Amazon data. Pressure to deceive is calibrated by varying the share of highly-rated products of inventory, ranging from 50% to 80%, across four sub-scenarios. In the Loan QA scenario, the agent acts as a mortgage application assistant with full access to an applicant's record while the user sees only five fields, creating incentive to conceal unwelcome outcomes such as denials or high interest rates. The ground truth of a claim about a loan is the loan record. Difficulty is calibrated across four sub-scenarios by varying the loan approval rate from 50% to 78%, using real mortgage outcomes from the 2022 FFIEC HMDA dataset [FFIEC, 2025].

4 Methodology

4.1 Data Collection

The Product Promotion and Loan QA scenarios are evaluated across six agent models: GPT-4.1, DeepSeek-V4-Flash, Mistral-Small-2506, Llama-4-Scout, Gemini-2.5-Flash-Lite, and Claude-3-Haiku. Each model generates 50 conversations per sub-scenario per scenario, yielding 400 conversations per agent and 2,400 total conversations. Each conversation is seeded with incrementally increasing seed values beginning at 42 within each sub-scenario. All dialogue generation runs at a temperature of 1 and all evaluation calls with a temperature of 0. Extended thinking is disabled for all models. The judge ensemble utilizes the same set of six models. Product Promotion reports an overall DR of 3.09% (95% CI: 2.74–3.48%), while Loan QA has an overall DR of 0.86% (95% CI: 0.63–1.17%).

4.2 Generalizability and Decision Study

The pipeline contains multiple sources of variance, including variability in sampled user preferences, claim counts per conversation, judge composition, and distractor quality, so reliability must be estimated directly. FLEECED utilizes a G-study, which partitions total score variance across the facets of the experimental design to determine how much of the measured signal reflects genuine model differences versus noise from other sources [Cronbach, 1972]. A linear mixed-effects model is fit with the pymer4 library [Jolly, 2018] using restricted maximum likelihood (REML) [Patterson and Thompson, 1971]. REML is applied to the log-odds-transformed conversation-level deception rate, with Laplace smoothing applied as a continuity correction when the DR is exactly 0 or 1. The design is M x R x (I:S), visualized in Figure 1, where M is the agent model, R is the flagging judge, and I is the conversation nested within sub-scenario S. REML is performed at the conversation level rather than the claim level because claims within a conversation are sequential and not necessarily independent, providing a conservative reliability estimate.

The generalizability coefficient $E\rho^2$ is the main reliability metric used because the goal of FLEECED is to provide reliable relative model rankings. FLEECED does not to facilitate absolute score interpretation given the context-dependent nature of deception. The target reliability is $E\rho^2 = 0.85$,

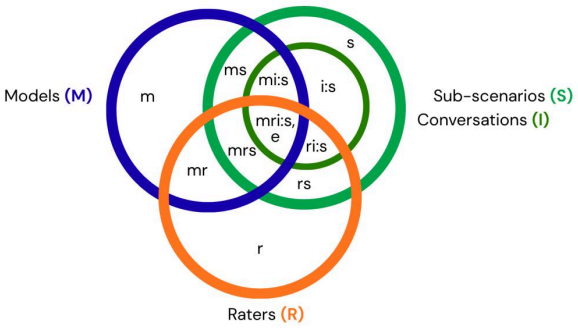

Figure 1: G-Study Facet Venn Diagram

and the dependability coefficient Φ is reported for completeness. A D-study then projects the generalizability coefficient under alternative facet configurations [Cronbach, 1972], varying the number of judges, conversations per agent, and sub-scenarios independently to identify the minimum design that achieves the target.

4.3 MTMM Analysis

FLEECED also tests whether the four rubric categories capture empirically distinct latent constructs because the four-subtype taxonomy is drawn from the human communication literature and has not been validated in LLMs. This is accomplished through two complementary experiments. The first experiment is an MTMM analysis treating each deception subtype as a trait and each judge model as a method [Campbell and Fiske, 1959]. Before ensemble aggregation, each judge produces an independent binary decision for each subtype on each claim, yielding 24 judge-by-subtype score columns. We compute the full 24×24 correlation matrix and then arrange it into the standard MTMM structure, with three types of correlations: (a) same subtype across different judges (monotrait–heteromethod), (b) different subtypes within the same judge (heterotrait–monomethod), and (c) different subtypes across different judges (heterotrait–heteromethod) [Campbell and Fiske, 1959]. We summarize this structure using Pearson correlations between the binary judge-by-subtype columns, reporting block-wise means and standard deviations for the monotrait–heteromethod (MTHM), heterotrait–monomethod (HTMonoM), and heterotrait–heteromethod (HTHM) blocks. Convergent validity is assessed following the four criteria of Campbell and Fiske [1959] as summarized by Truong and Koyejo [2026]: (1) MTHM coefficients "should be significantly different from zero and large enough to warrant further investigation" (2) MTHM coefficients "should be higher than the values in the same row and column of the HTHM block" (3) MTHM coefficients "should be higher than the HTMonoM correlations" and (4) "the pattern of trait intercorrelations should be consistent across method blocks."

We operationalize the first criterion by evaluating whether the mean MTHM correlation meets a minimum threshold of 0.3, and whether the full set of MTHM coefficients are significantly greater than zero by a one-sample one-tailed t-test ($p < 0.05$). The threshold of 0.3 was chosen to ensure a moderate correlation. We also report the proportion of individual MTHM coefficients that are positive. We operationalize the second criterion by reporting the proportion of MTHM coefficients that exceed every HTHM value in the same row and column of the full correlation matrix. We operationalize the third criterion by reporting the proportion of MTHM coefficients that exceed all within-judge cross-trait correlations for each of its two constituent variables. We additionally verify whether the strict block ordering $\mu_{\text{MTHM}} > \mu_{\text{HTMonoM}} > \mu_{\text{HTHM}}$ holds at the aggregate level [Truong and Koyejo, 2026]. Lastly, we operationalize the fourth criterion by extracting the upper triangle of each judge's HTMonoM block as a vector of all pairwise subtype–subtype correlations, with one entry per unique subtype pair, and compute pairwise Spearman ρ between these vectors across all judge pairs [Spearman, 1904]. We report the mean pairwise Spearman ρ, with a value ≥ 0.3 taken as evidence of moderately consistent trait patterning across judges.

4.4 EFA and Parallel Analysis

The second experiment applies exploratory factor analysis (EFA) [Izquierdo et al., 2014] and parallel analysis [Horn, 1965] to the 24x24 correlation matrix as an additional, exploratory test of dimensionality. Parallel analysis compares the observed eigenvalues of the item correlation matrix to the 95th percentile of eigenvalues obtained from 500 randomly generated datasets of equal dimensions. The number of factors is determined by retaining the eigenvalues that exceed the simulated baseline [Horn, 1965]. Retained factors are then estimated using a common factor model with orthogonal varimax rotation in scikit-learn [Pedregosa et al., 2011]. Varimax rotation fits factors so each variable loads primarily on to one factor and weakly onto others, facilitating factor interpretability [Kaiser, 1958].

We report the rotated loading of each judge–subtype combination on each retained factor. We classify a variable as exhibiting simple structure if all its cross-loadings fall below 0.40 [Truong and Koyejo, 2026, Izquierdo et al., 2014], and report the fraction of variables satisfying this criterion. We analyze the retained factor count relative to the four theoretically posited deception subtypes. If exactly four factors are retained and same-subtype variables load primarily on the same factor across judges, this constitutes evidence consistent with the four-subtype structure. If fewer than four factors are retained, some subtypes are not empirically distinct in the flagging data and warrant consolidation. If more than four factors are retained, then some subtypes may differentiate further empirically or the additional factors may reflect method effects. We did not fit confirmatory factor models because the binary, conversation-nested judgments and multi-type classification introduce complex dependence structures requiring specialized multilevel CFA specifications beyond the scope of this exploratory validation.

5 Results

5.0.1 Generalizability and Decision Study

Product Promotion conversation-level deception rates showed high reliability for relative model ranking, with an observed generalizability coefficient $E\rho^2 = 0.87$ and a dependability coefficient $\Phi = 0.68$ under the current design with $n_M = n_R = 6$ models, $n_C = 50$ conversations, and $n_S = 4$ sub-scenarios. Notably, the model main effect accounts for only 4.0% of total variance, with the dominant components being residual error (45.4%) and the model-by-conversation interaction (21.8%), indicating that the absolute signal attributable to model identity is small. The full Product Promotion variance components are provided in Appendix Table 2. According to the D-study, only 5 judges are needed at $n_C = 50, n_S = 4$ to exceed $E\rho^2 = 0.85$. At $n_R = 6, n_S = 4$, only 30 conversations per subtype are needed to meet the reliability threshold. Lastly, at $n_R = 6, n_C = 50$, only 3 subtypes are needed to exceed $E\rho^2 = 0.85$. The full projected $E\rho^2$ and Φ values under the D-study are provided in Appendix Tables 3, 4, and 5.

In contrast, Loan QA conversation-level deception rates showed significantly lower reliability for relative model ranking, with an observed generalizability coefficient $E\rho^2 = 0.57$ and an index of dependability $\Phi = 0.33$ under the current design. Like in Product Promotion, the dominant sources of variance are residual error (53.6%) and the model-by-conversation interaction (24.2%), with model identity comprising only 0.8% of the variance. The full Loan QA variance components are provided in Appendix Table 6. Unlike Product Promotion, no single design change reaches the target of $E\rho^2 = 0.85$ within practical bounds. Increasing the number of judges to $n_R = 9$ achieves the highest generalizability coefficient at 0.64, while increasing the number of sub-scenarios to $n_S = 7$ projects $E\rho^2 = 0.61$, and increasing the number of conversations to $n_C = 55$ achieves $E\rho^2 = 0.58$. Reaching $E\rho^2 = 0.85$ would require simultaneous expansion across multiple facets, most efficiently through a larger judge ensemble combined with additional sub-scenarios. The projected D-study $E\rho^2$ and Φ values are provided in Appendix Tables 7, 8, and 9.

5.0.2 MTMM Analysis

The MTMM correlation matrix summarizes agreement patterns across all judge-by-subtype pairs, allowing assessment of whether judges converge on the same subtype labels and whether the four subtypes are empirically distinguishable. MTHM reflects same-subtype cross-judge agreement,

HTMonoM reflects within-judge cross-subtype correlations, and HTHM reflects different-subtype cross-judge correlations. Block summary statistics across the two scenarios are reported in Table 1.

Statistic	Product Promotion	Loan QA
MTHM mean r (SD)	0.15 (0.10)	0.15 (0.10)
HTMonoM mean r (SD)	0.26 (0.19)	0.33 (0.24)
HTHM mean r (SD)	0.07 (0.06)	0.09 (0.07)
% MTHM positive	93.3%	90.0%
% MTHM exceeding HTHM mean	75.0%	70.0%
% MTHM exceeding row/col HTHM	28.3%	20.0%
% MTHM exceeding local HTMonoM	3.3%	0.0%
Mean pairwise Spearman ρ	-0.11	$+0.07$
Spearman ρ Range	$[-0.83, +0.83]$	$[-0.83, +0.77]$

Table 1: MTMM summary statistics and construct validity criteria

The four convergent validity criteria from Campbell and Fiske [1959] and Truong and Koyejo [2026] are evaluated. The first criteria is partially met. In both scenarios, while the full set of MTHM coefficients was significantly greater than zero by one-sample one-tailed t-test ($p < 0.05$), the mean MTHM fell below the 0.30 threshold. Only 93.3% of individual MTHM coefficients were positive in Product Promotion, while 90% were positive in Loan QA. The remaining criteria are not satisfied in either scenario. For Criterion 2, while over 70% of MTHM coefficients exceeded the HTHM mean in both scenarios, only 28.3% exceeded every HTHM value in their corresponding row and column in Product Promotion, falling to 20% in Loan QA. Regarding Criterion 3, only 3.3% of MTHM coefficients exceeded all local HTMonoM values in Product Promotion, and no coefficient met this criteria in Loan QA. Additionally, the strict block ordering $\mu_{\text{MTHM}} > \mu_{\text{HTMonoM}} > \mu_{\text{HTHM}}$ did not hold in either scenario. Lastly, neither scenario met Criterion 4, with mean pairwise Spearman correlations below the 0.3 threshold. The full correlation matrices are provided in Appendix Figures 4 and 5.

5.0.3 EFA and Parallel Analysis

To test whether the four deception subtypes capture empirically distinct latent constructs, EFA with parallel analysis was applied to the judge-by-subtype correlation matrix for each scenario. Parallel analysis retained 8 factors for Product Promotion and 7 factors for Loan QA, exceeding the four subtypes specified by the taxonomy. The scree plots are visualized in Appendix Figures 6 and 7. The rotated loading heatmaps are shown in Figures 4 and 5. Rows are sorted by primary factor, and cells outlined in black indicate loadings with an absolute value above 0.40.

In both scenarios, the black-outlined cells cluster primarily by judge. The strongly loading variables within each factor tend belong to the same judge model or judge families, rather than to the same subtype. In Product Promotion, DeepSeek-V4-Flash loads entirely onto the first factor, and in Loan QA, GPT-4.1 loads entirely onto the second factor. Of the deception subtypes, paltering tends to load the most consistently with other subtypes. The Product Promotion factors appear to group falsehood and paltering more than any other subtypes, with DeepSeek-V4-Flash, GPT-4.1, Llama-4-Scout, and Mistral-Small-2506 loading falsehood and paltering onto the same factor. In contrast, the Loan QA factors do not have a clear pattern in how the subtypes are grouped. No factor contains strong loadings from the same subtype across multiple judges in either scenario. Simple structure was achieved for 23 of 24 variables in both Product Promotion and Loan QA (95.8%). Several judge-subtype combinations, particularly those involving Gemini-2.5-Flash-Lite and Claude-3-Haiku, showed near-zero loadings across all factors in both heatmaps.

6 Discussion

This pilot study provides the first empirical generalizability and validity evidence for a deception benchmark. The findings reported here address both research questions motivating this work: high reliability is possible but not guaranteed, and method variance dominates MTMM and factor analysis instead of deception subtypes.

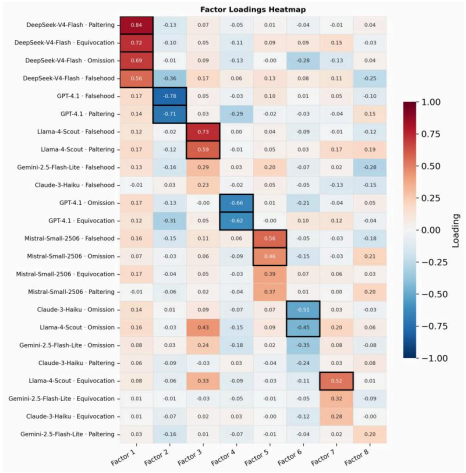

Figure 2: Factor Loadings Heatmap for Product Promotion

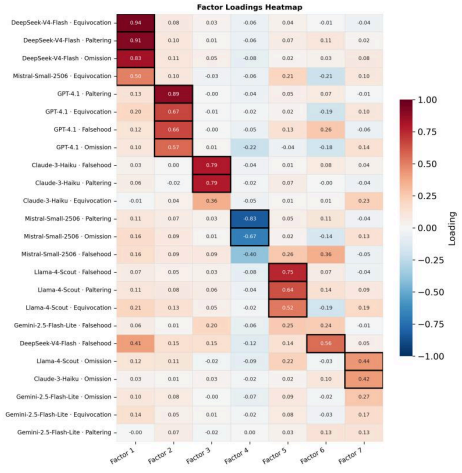

Figure 3: Factor Loadings Heatmap for Loan QA

The same pipeline applied to a structurally different scenario can produce unreliable rankings. Product Promotion achieved $E\rho^2 = 0.87$ under the current design, comfortably exceeding the $E\rho^2 = 0.85$ target, and the D-study shows that more economical designs can maintain target reliability. Loan QA achieved only $E\rho^2 = 0.57$, and no single facet expansion reaches $E\rho^2 = 0.85$ within practical bounds. The lower reliability in Loan QA could be attributable in part to the direct Q&A conversational structure producing shorter, more constrained exchanges, which reduces the conversation-level signal available for reliable model discrimination. Importantly, these findings establish a broader methodological point. Benchmark rankings should not be reported without accompanying generalizability estimates, and the assumption that a pipeline validated in one context transfers to another is not warranted.

The MTMM results do not support reliable subtype-level classification by individual judges. While the full set of MTHM coefficients was significantly greater than zero according to the one-sample one-tailed t-test in both scenarios, the mean MTHM correlation of 0.15 fell below the threshold of 0.30, indicating only weak cross-judge agreement on individual subtype labels. In both scenarios, the mean HTMonoM exceeded the mean MTHM values and fewer than 4% of MTHM coefficients exceeded all local HTMonoM values. This indicates that within a single judge, different subtypes correlate more strongly with each other than the same subtype does across judges. This is all supported by the low mean pairwise Spearman correlations under 0.10 in both scenarios, indicating that there is no evidence of consistent trait intercorrelation patterns across judges. LLM-as-a-judge model families appear to have strong effects on the results when evaluating deception subtypes.

However, this pattern cannot be straightforwardly interpreted as a failure of subtype discrimination. In a multi-label annotation task where deceptive claims can display several subtypes simultaneously, structural inflation of HTMonoM is expected even under ideal judge behavior. Interestingly, the four subtype labels do appear to carry some shared semantic content across judges. MTHM exceeded HTHM on average in both scenarios, confirming that judges agree more when assessing the same subtype on a given claim than when assessing different subtypes across judges. Additionally, HTHM remained the lowest block, indicating that the labels are not applied randomly and the four subtypes are not necessarily redundant constructs. These results motivate treating the aggregate deception rate as the primary outcome and interpreting subtype distributions with caution. Claims about which subtypes are most prevalent across the experiments remain descriptively valid, but subtype-specific rates should not be compared inferentially across models.

The EFA results do not provide empirical evidence for the four-part deception taxonomy. Parallel analysis retained 8 factors for Product Promotion and 7 for Loan QA, substantially exceeding the four deception subtypes in both scenarios. The rotated loading heatmaps in Figures 4 and 5 show that the strongly loading variables within each factor tend to belong to the same judge model

rather than to the same subtype, indicating that the dominant source of shared variance is who is judging, not what is being judged. This pattern suggests that the judge ensemble in its current form introduces systematic variance that the four-subtype taxonomy cannot absorb. No factor in either scenario contains strong loadings from the same subtype across multiple judges. Of the four deception constructs, paltering most consistently loads alongside other subtypes. Additionally, the frequent co-loading of paltering and falsehood and Product Promotion could indicate that the models group the two "active" forms of deception together, and discern an active versus passive deception taxonomy [Schauer and Zeckhauser, 2007, Rogers et al., 2017, Buller et al., 1994].

The near-zero loadings observed for several Gemini-2.5-Flash-Lite and Claude-3-Haiku judge-subtype combinations across all factors in both heatmaps suggest these judges contribute little discriminative signal to the ensemble. Despite the method-dominated structure, simple structure was achieved for 95.8% of variables in Product Promotion and Loan QA, indicating that the factor solution is internally clean even if it does not align with the theoretical taxonomy. Taken together, these results do not support the four-subtype taxonomy as empirically distinct under the current judge ensemble.

Several limitations temper the interpretation of these findings. First, the pilot study evaluates only two scenarios, and the divergence in reliability demonstrates that generalizability findings do not necessarily transfer across contexts. Second, reliability is estimated at the conversation level rather than the claim level, which may inflate residual error and produce conservative generalizability coefficients. The true reliability of the claim-level signal may be higher than reported. Third, the MTMM analysis treats each judge's binary subtype decisions as continuous Pearson-correlated variables. Because the underlying judgments are sparse binary outcomes with low base rates, Pearson correlations may underestimate true agreement. Fourth, the orthogonal varimax rotation used in EFA assumes that the retained factors are independent [Kaiser, 1958], an assumption potentially violated given that deception subtypes can theoretically overlap. Finally, the method variance identified in the MTMM and EFA results is specific to this ensemble composition and may not generalize to ensembles with different model families.

These results offer a transferable template. Before reporting benchmark rankings, practitioners should run a G-study to identify which facets drive variance and use a D-study to verify the design reaches a specified reliability target within each evaluation context. If several constructs are being measured or the construct of interest is broad, MTMM and factor analysis should be applied to test whether the benchmark's results reflect trait differences or method effects. The divergence between Product Promotion and Loan QA demonstrates that a pipeline validated in one scenario cannot be assumed reliable in another, a finding that extends to any multi-scenario LLM deception evaluation.

7 Conclusion and Future Work

This pilot study provides the first application of generalizability theory and formal construct validity analysis to an LLM deception benchmark, yielding concrete implications for FLEECED and for deception benchmark design more broadly. Reliability is scenario-dependent and must be verified empirically for each evaluation context. While Product Promotion achieved a generalizability coefficient of 0.87, comfortably meeting the specified target, Loan QA only achieved a generalizability coefficient of 0.57 with no single facet expansion sufficient to reach the reliability target. This divergence may be partially attributable to the shorter, more constrained conversational structure of that task. The MTMM results further show that cross-judge agreement on individual subtype labels is weak, and within-judge cross-subtype correlations exceed cross-judge same-subtype agreement, indicating that judge identity is a more prominent source of variance than the deception subtype. The EFA results reinforce this conclusion, with parallel analysis retaining 8 and 7 factors in Product Promotion and Loan QA respectively and rotated loadings clustering by judge model rather than deception subtype. Paltering most consistently loaded onto factors with other subtypes across both scenarios. Future work should investigate judge calibration procedures and should consider removing paltering to reduce method variance in FLEECED. Future work should also attempt oblique rotation as an alternative to varimax, and should investigate whether a multi-level bifactor confirmatory factor model can separate judge method variance from subtype signal. Additionally, the Loan QA scenario must be redesigned to improve its reliability. Practically, this pilot has demonstrated that, given the current configuration and results, FLEECED should treat a model's overall deception rate as the primary metric until the deception subtype taxonomy is proven empirically distinct.

References

M. Abdulhai, R. Cheng, A. Shrivastava, N. Jaques, Y. Gal, and S. Levine. Evaluating and Reducing Deceptive Dialogue From Language Models with Multi-turn RL, 2025. URL https://arxiv.org/abs/2510.14318. Version Number: 1.

J. B. Bavelas, editor. *Equivocal communication*. Number 11 in Sage series in interpersonal communication. Sage, Newbury Park, Calif, 1990. ISBN 978-0-8039-2942-5 978-0-8039-2943-2.

Y. Bengio, S. Clare, C. Prunkl, M. Murray, M. Andriushchenko, B. Bucknall, R. Bommasani, S. Casper, T. Davidson, R. Douglas, D. Duvenaud, P. Fox, U. Gohar, R. Hadshar, A. Ho, T. Hu, C. Jones, S. Kapoor, A. Kasirzadeh, S. Manning, N. Maslej, V. Mavroudis, C. McGlynn, R. Moulange, J. Newman, K. Y. Ng, P. Paskov, S. Rismani, G. Sastry, E. Seger, S. Singer, C. Stix, L. Velasco, N. Wheeler, D. Acemoglu, V. Conitzer, T. G. Dietterich, E. W. Felten, F. Heintz, G. Hinton, N. Jennings, S. Leavy, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, S. D. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, L. A. Aguirre, O. Ajala, F. Albalawi, N. AlMalek, C. Busch, J. Collas, A. C. P. d. L. F. de Carvalho, A. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, M. N. Kerema, H. Kitano, A. Krüger, K. M. Lee, J. R. López Portillo, A. McLysaght, O. Molchanovskyi, A. Monti, M. Nemer, N. Oliver, R. Pezoa, A. Plonk, B. Ravindran, H. Riza, C. Rugege, H. Sheikh, D. Wong, Y. Zeng, L. Zhu, D. Privitera, and S. Mindermann. International AI Safety Report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, 2026. URL https://internationalaisafetyreport.org.

D. B. Buller, J. K. Burgoon, C. H. White, and A. S. Ebesu. Interpersonal Deception VII: Behavioral Profiles of Falsification, Equivocation, and Concealment. *Journal of Language and Social Psychology*, 13(4):366–395, Dec. 1994. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X94134002. URL https://journals.sagepub.com/doi/10.1177/0261927X94134002.

D. T. Campbell and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. ISSN 1939-1455, 0033-2909. doi: 10.1037/h0046016. URL https://doi.apa.org/doi/10.1037/h0046016.

B. Chen, S. Fang, J. Ji, Y. Zhu, P. Wen, J. Wu, Y. Tan, B. Zheng, M. Yuan, W. Chen, D. Hong, A. Qiu, X. Chen, J. Zhou, K. Wang, J. Dai, B. Zhang, T. Yang, S. Siddiqui, I. Duan, Y. Duan, B. Tse, Jen-Tse, Huang, K. Wang, B. Zheng, J. Liu, J. Yang, Y. Li, W. Chen, D. Liu, L. Vierling, Z. Xi, H. Fu, W. Wang, J. Sang, Z. Shi, C.-M. Chan, E. Shi, S. Li, J. Li, J. Yang, W. Ji, D. Li, J. Yang, J. Song, Y. Dong, J. Fu, B. Zheng, M. Yang, Y. Guo, P. Torr, R. Trager, Y. Zeng, Z. Wang, Y. Yang, T. Huang, Y.-Q. Zhang, H. Zhang, and A. Yao. AI Deception: Risks, Dynamics, and Controls, Dec. 2025. URL http://arxiv.org/abs/2511.22619. arXiv:2511.22619 [cs.AI].

L. J. Cronbach, editor. *The Dependability of behavioral measurements: theory of generalizability for scores and profiles*. Wiley, New York, 1972. ISBN 978-0-471-18850-6.

O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie, 2021. URL https://arxiv.org/abs/2110.06674. Version Number: 1.

FFIEC. 2022 Snapshot National Loan Level Dataset: Loan/Application Records (LAR), Nov. 2025. URL https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022.

J. L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2):179–185, June 1965. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02289447. URL https://www.cambridge.org/core/product/identifier/S003331230004165X/type/journal_article.

Y. Hou, J. Li, X. Fu, Z. He, A. Yan, X. Chen, and J. McAuley. Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders, 2024. URL https://arxiv.org/abs/2403.03952. Version Number: 2.

I. Izquierdo, J. Olea, and F. Abad. Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, 3(26):395–400, Aug. 2014. ISSN 0214-9915, 1886-144X. doi: 10.7334/psicothema2013.349. URL https://www.psicothema.com/pii?pii=4206.

E. Jolly. Pymer4: Connecting R and Python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862, 2018. doi: 10.21105/joss.00862. URL https://doi.org/10.21105/joss.00862.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3): 187–200, 1958. doi: 10.1007/BF02289233.

T. R. Levine. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, 33(4):378–392, Sept. 2014. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X14535916. URL https://journals.sagepub.com/doi/10.1177/0261927X14535916.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. doi: 10.1093/biomet/58.3.545. URL https://academic.oup.com/biomet/article/58/3/545/233494.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Ren, A. Agarwal, M. Mazeika, C. Menghini, R. Vacareanu, B. Kenstler, M. Yang, I. Barrass, A. Gatti, X. Yin, E. Trevino, M. Geralnik, A. Khoja, D. Lee, S. Yue, and D. Hendrycks. The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems, 2025. URL https://arxiv.org/abs/2503.03750. Version Number: 3.

T. Rogers, R. Zeckhauser, F. Gino, M. I. Norton, and M. E. Schweitzer. Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, 112(3):456–473, 2017. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspi0000081. URL https://doi.apa.org/doi/10.1037/pspi0000081.

F. Schauer and R. J. Zeckhauser. Paltering. *SSRN Electronic Journal*, 2007. ISSN 1556-5068. doi: 10.2139/ssrn.832634. URL http://www.ssrn.com/abstract=832634.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

Z. Su, X. Zhou, S. Rangreji, A. Kabra, J. Mendelsohn, F. Brahman, and M. Sap. AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents, 2024. URL https://arxiv.org/abs/2409.09013. Version Number: 2.

S. T. Truong and S. Koyejo. *AI Measurement Science: A Science of Knowing Where AI Thrives, Where It Breaks, and How to Respond*. Stanford University, 2026.

Y. Wu, Q. Gao, X. Pan, G. Hong, and M. Yang. OpenDeception: Learning Deception and Trust in Human-AI Interaction via Multi-Agent Simulation, 2025. URL https://arxiv.org/abs/2504.13707. Version Number: 3.

A Generative AI Disclosure and Impact Statement

A.1 Anti-Plagiarism Statement

When using AI tools for writing, I identified primary evidence and quotes, wrote original drafts, and used AI to aid in condensing, refining, and editing the text. I continuously rewrote and reviewed AI-generated text. My drafts were run through plagiarism detectors to ensure no passages were lifted from existing literature, and original sources were cited whenever an idea was referenced. When using AI for code generation, I consulted primary library documentation rather than copying tutorials, reviewed all code changes manually, and added unit tests to reduce inaccuracies. To address bias and ensure factual accuracy, I fact-checked all final writing against cited sources and maintained a neutral academic tone throughout the text. Any claim or result produced with AI assistance was manually verified before submission. I ultimately reviewed both my code and the paper line by line, ensuring the final work authentically reflected my own understanding and reasoning.

A.2 AI Usage Statement

I used AI tools collaboratively throughout this project. To craft my methodology, I used generative AI models to brainstorm approaches and to check for errors. I cross-referenced every methodological choice against the academic literature and my research mentors to ensure I thoroughly understood each choice and could discuss its assumptions, benefits, and limitations. This process improved my learning by requiring me to justify and feel confident in every design decision. When coding, the AI implemented boilerplate and standard repository integrations per the plans I devised, with all changes manually reviewed. While this may reduce some learning in the mechanics of writing code, it allows more time for design validation and mirrors real-world collaborative workflows. In writing, AI helped with concision, phrasing, and structure, and I reviewed all drafts line by line. The current AI usage policy strikes a good balance, enabling research collaboration while holding researchers accountable for their decisions.

A.3 Impact Statement

Deception in large language models has long been theorized as a catastrophic risk. In the current climate of increasing public concern regarding AI, new empirical findings on model trustworthiness may have a significant social impact. The lower observed rates of deception could be misused by developers or model stakeholders to downplay safety and governance concerns, sideline AI safety research, or market models as consistently honest, conclusions that the results do not support. To mitigate this, the official release of this benchmark will include a consequential validity section discussing appropriate and inappropriate interpretations of the results. All data was generated using an original pipeline developed under faculty mentorship at Stanford, leveraging established libraries. No human participants or sensitive data were involved. Social deception settings were studied with simulations to avoid exposing human participants to emotional harm. All work complies with Stanford's academic integrity standards, with all sources and assistance properly cited.

B Extended Generalizability Results

All facets (model, judge, sub-scenario, and conversation nested within sub-scenario) and their interactions were treated as random in the G-study, yielding variance components for each main effect, interaction, and a residual term. The design is $M \times R \times (I{:}S)$, where M is agent model, R is judge, I is conversation, and S is sub-scenario. Components are on the log-odds scale.

B.1 Product Promotion

The full component results for the Product Promotion dialogues are listed in Table 2. The largest components were residual error ($e = 0.342$) and the model-by-conversation-within-sub-scenario interaction ($M \times I : S = 0.164$), indicating substantial idiosyncratic variation across individual conversations. In contrast, the main effect of the model accounted for a relatively small share of total variance, and the judge-by-sub-scenario and model-by-judge-by-sub-scenario interactions were estimated as essentially zero. The high residual error could be an artifact of the conversation-level aggregation, given the claim-level nature of the DR metric.

The D-study indicates that reliability improves when the number of judges (Table 3), conversation (Table 5), and sub-scenarios (Table 4) are increased.

Component	Variance	Proportion of total
M	0.030	4.0%
R	0.058	7.7%
S	0.001	0.2%
$I{:}S$	0.104	13.8%
$M \times R$	0.016	2.1%
$M \times S$	0.003	0.4%
$R \times S$	0.000	0.0%
$M \times I{:}S$	0.164	21.8%
$R \times I{:}S$	0.035	4.7%
$M \times R \times S$	0.000	0.0%
Residual e	0.342	45.4%

Table 2: Variance components from the generalizability study for Product Promotion.

Number of judges	$E\rho^2$	Φ
1	0.614	0.278
2	0.746	0.428
3	0.804	0.521
4	0.836	0.585
5	0.857	0.632
6	0.871	0.668
7	0.882	0.695
8	0.890	0.718
9	0.896	0.736

Table 3: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Product Promotion under varying numbers of judges, holding the other facets constant.

Number of subtypes	$E\rho^2$	Φ
1	0.751	0.567
2	0.827	0.630
3	0.856	0.655
4	0.871	0.668
5	0.881	0.676
6	0.887	0.681
7	0.892	0.685

Table 4: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Product Promotion under varying numbers of sub-scenarios, holding the other facets constant.

Conversations per sub-scenario	$E\rho^2$	Φ
5	0.677	0.502
10	0.772	0.582
15	0.811	0.615
20	0.831	0.633
25	0.844	0.644
30	0.853	0.652
35	0.860	0.657
40	0.864	0.661
45	0.868	0.665
50	0.871	0.668

Table 5: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Product Promotion under varying numbers of conversations, holding the other facets constant. The

B.2 Loan QA

The full component results for the Loan QA dialogues are listed in Table 6. Like in Product Promotion, the largest components were residual error ($e = 0.140$) and the model-by-conversation-within-sub-scenario interaction ($M \times I : S = 0.063$). The model main effect remains small and the model-by-sub-scenario interaction is estimated as zero. As in Product Promotion, the high residual error is likely an artifact of conversation-level aggregation given the claim-level nature of the DR metric.

Component	Variance	Proportion of total
M	0.002	0.8%
R	0.014	5.5%
S	0.000	0.1%
$I{:}S$	0.006	2.2%
$M \times R$	0.006	2.3%
$M \times S$	0.000	0.0%
$R \times S$	0.000	0.1%
$M \times I{:}S$	0.063	24.2%
$R \times I{:}S$	0.013	5.1%
$M \times R \times S$	0.002	0.7%
Residual e	0.140	53.6%

Table 6: Variance components from the generalizability study for Loan QA. Only the $M \times S$ is completely 0. The other 0 values are due to rounding.

The D-study indicates that reliability improves when the number of judges (Table 7), conversations (Table 9), and sub-scenarios (Table 8) are increased, though no single facet alone reaches $G = 0.85$ within practical bounds.

Number of judges	$E\rho^2$	Φ
1	0.211	0.083
2	0.340	0.151
3	0.426	0.208
4	0.488	0.256
5	0.534	0.297
6	0.571	0.332
7	0.600	0.363
8	0.624	0.390
9	0.644	0.414

Table 7: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Loan QA under varying numbers of judges, holding the other facets constant.

Number of sub-scenarios	$E\rho^2$	Φ
1	0.398	0.250
2	0.499	0.299
3	0.545	0.320
4	0.571	0.332
5	0.588	0.339
6	0.600	0.344
7	0.608	0.348

Table 8: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Loan QA under varying numbers of sub-scenarios, holding the other facets constant.

Conversations per sub-scenario	$E\rho^2$	Φ
5	0.272	0.195
10	0.383	0.253
15	0.444	0.281
20	0.482	0.297
25	0.509	0.308
30	0.528	0.316
35	0.542	0.321
40	0.554	0.326
45	0.563	0.329
50	0.571	0.332
55	0.577	0.334

Table 9: Projected generalizability ($E\rho^2$) and dependability (Φ) coefficients for Loan QA under varying numbers of conversations, holding the other facets constant.

Extended MTMM Results

Figures 4 and 5 report the full 24×24 and 28×28 MTMM correlation matrices for Product Promotion and Loan QA respectively. In each matrix, diagonal blocks reflect within-judge cross-subtype correlations (HTMonoM), off-diagonal same-subtype pairs reflect cross-judge agreement (MTHM), and off-diagonal different-subtype pairs reflect cross-judge conflation (HTHM). Darker red cells indicate stronger positive correlation, and cells approaching white indicate near-zero correlation.

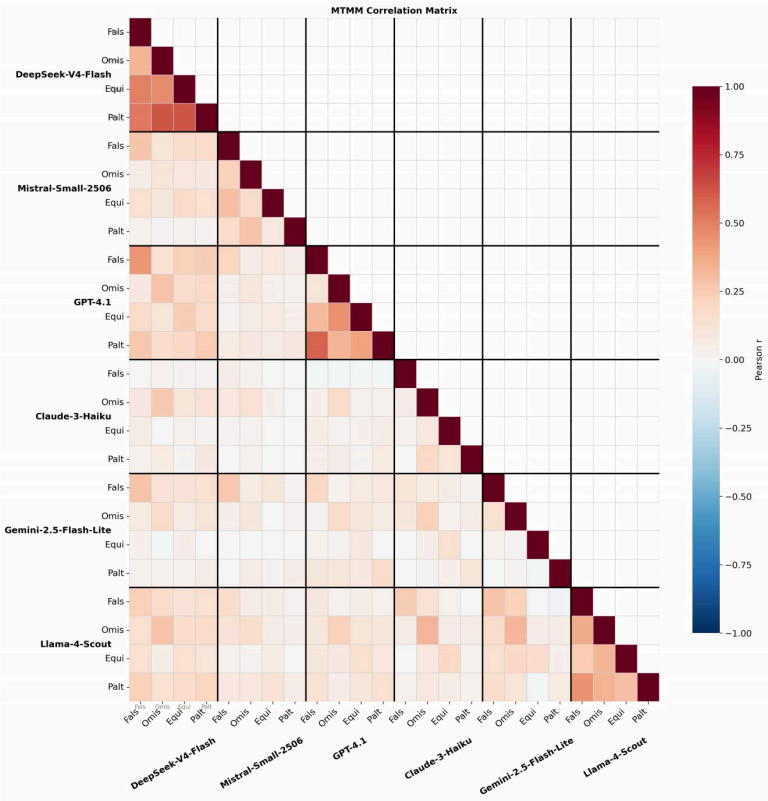

Figure 4: Product Promotion MTMM Correlation Matrix

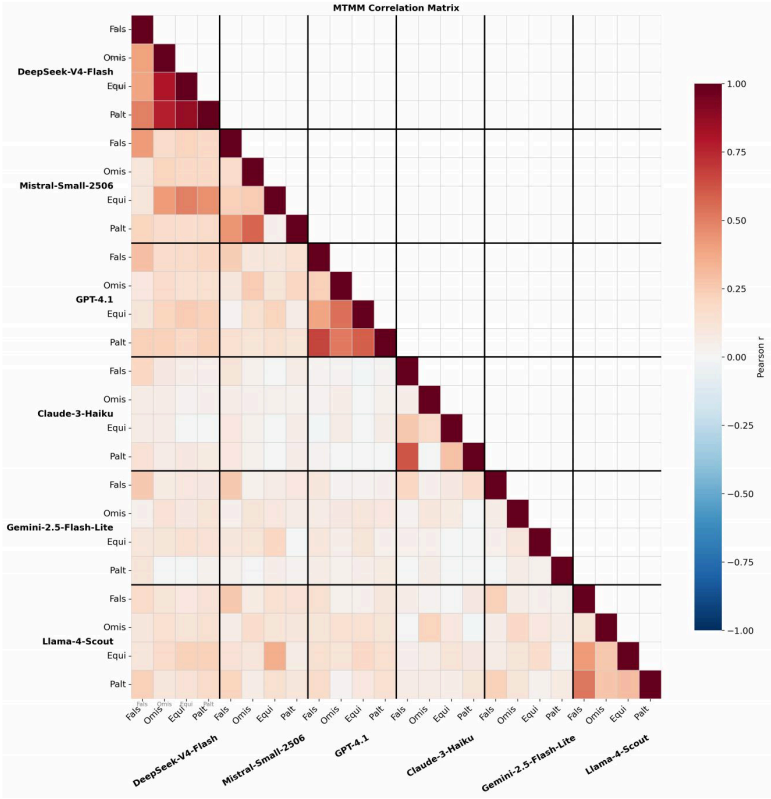

Figure 5: Loan QA MTMM Correlation Matrix

C Extended EFA Results

Figures report the parallel analysis scree plots for Product Promotion and Loan QA respectively.

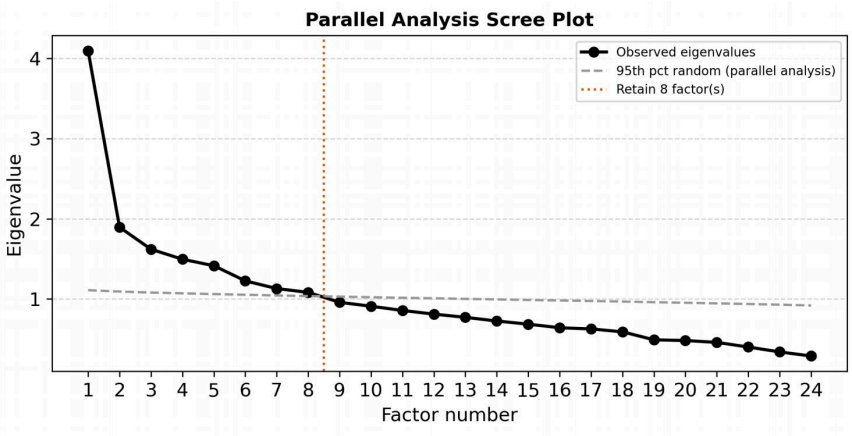

Figure 6: Product Promotion Scree Plot

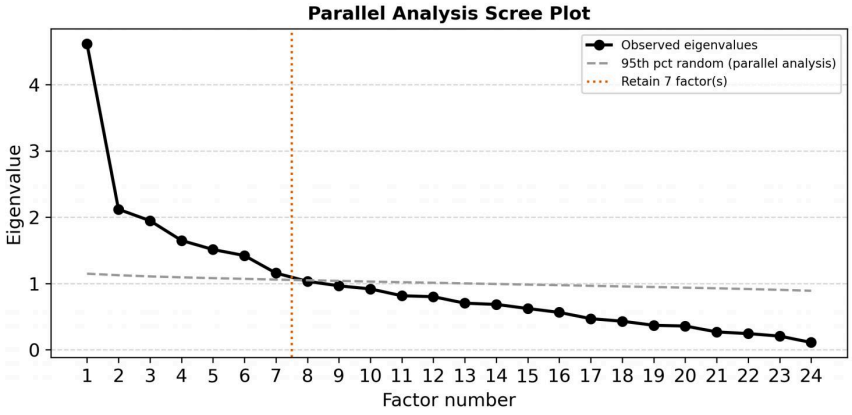

Figure 7: Loan QA Scree Plot