

Are Aliases Neutral?

Company-Name Form in LLM Equity Valuation

Author

John Garcia

School of Management, California Lutheran University, 60 West Olsen Road, Thousand Oaks, CA 91360, USA

Email: jgarcia@callutheran.edu; ORCID: 0000-0002-7062-0032

Acknowledgements: None

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Registration and materials (for the editor)

- Registration of record: <https://osf.io/ewcmf> (submitted September 23, 2026, after data collection and before confirmatory analysis; embargoed to April 1, 2027). Editor's anonymized view-only link: https://osf.io/ewcmf/overview?view_only=6a4b32bc1ae04b3fb25d9f588df301c2 (the ordinary identifier does not open an embargoed registration).
- OSF project with the design plan, stimuli, prompts, and hash manifest uploaded before confirmatory collection, and the pre-analysis amendment: <https://osf.io/jxegm>.
- Please do not forward the project link to reviewers; the amendment and code identify the author.
- Reviewers receive the same anonymized link through the data availability statement and Supplement S1. It hides contributors, and the frozen plan, materials, and hash manifest carry no name or email.
- Supplement S1 gives the OSF registration timeline without identifiers.

Manuscript facts

- Main text (Sections 1–4, with headings): 2,208 words; abstract: 198 words; total: 2,406 words, excluding references, Table 1, figure captions, and declarations. One table, two figures, and an anonymized online supplement.

1. Introduction

Experiments that present financial information to LLMs can mask firm identity because a model may recall what happened to a real firm after the information date (look-ahead bias; Didisheim et al., 2025; Glasserman & Lin, 2024; Sarkar & Vafa, 2024). Researchers delete identifiers, supply unlabeled numeric data (Chen et al., 2024), replace names in real text with placeholders that sacrifice economic information (Wu et al., 2025), or substitute invented company names. Entity neutering asks whether a firm can still be recognized after identifying information is stripped (Engelberg et al., 2025), and this study asks what the replacement text does.

Substitution keeps a prompt readable but adds an input: an alias may resemble a word, suggest an industry or country, or sound more credible than another spelling. LLM outputs can shift with prompt features that should be irrelevant (Sclar et al., 2024). A label that is arbitrary to the researcher need not be neutral to the model.

Human evidence motivates this concern. Alter and Oppenheimer (2006) document higher return expectations for fabricated names that are easier to pronounce, and Fenneman et al. (2023) show that the instructions change the size of this gap. Archival name effects may reflect information about firms (Montone et al., 2023), which an experimental alias lacks. For language models, Vitevitch (2025) documented ChatGPT ratings of non-word word likeness, and Glasserman and Lin (2024) reported similar forecasts across three replacement strings. None of these studies measured whether the invented name itself moves an LLM's fair value, which is the quantity that an alias policy must leave undisturbed.

This study directly audited the alias policy. I crossed 60 matched name pairs, split by model-rated pronunciation ease, with nine fictional company profiles and three commercial models,

collecting 27,762 responses under a plan frozen before collection and registered before the confirmatory analysis. The registered fluent-minus-disfluent contrast was 0.5% and inconclusive, whereas the secondary estimates carried the practical message that the generated labels lowered every model's valuations relative to an unnamed profile, and one sentence calling the label arbitrary shifted both valuation levels and the name gap. The results concern label sensitivity in these settings rather than the effectiveness of anonymization or distinct psychological mechanisms. For researchers who mask firm identity with invented names, the alias is therefore a design input whose assignment rule and balance across treatment arms belong to the preregistration.

2. Experimental Design

2.1. Names and Profiles

Using a fixed-seed generator, I produced 5,600 seven- and eight-letter candidate strings across 1,867 letter families. Two open-weight models outside the subject families, Llama 3.3 70B and Qwen 2.5 72B, rated each candidate's pronunciation ease and English word likeness on a 0–100 scale in three independent calls. A candidate's ease score averages the two raters' medians; fluent names score at least 65 and disfluent names at most 35. These labels are model-rated proxies, not validated human pronunciations.

I matched pairs on equal length and identical contextual token counts across the three subject models, with at most one pair per letter family. An integer program then minimized the differences in word likeness and dictionary-based bigram scores. The dictionary and entity screens preceded any financial outcome; the entity screen removed five pairs, each matched to its fluent member (Supplement S2). The mean rated ease was 69.9 for fluent and 32.4 for disfluent members (e.g., Toyewis and Feoccls); word likeness, which varies with ease, still differed after

matching (54.4 versus 24.9), so each contrast pairs selected spelling bundles, of which pronunciation ease is one component.

Each alias was crossed with nine fixed, profitable, fictional company profiles. Prompts retain the operating description and fundamentals but omit current prices and price-linked multiples. The alias appears only once, in the heading, and all other references read "The company." Because the prompt is otherwise identical for a given profile and condition, crossing isolates the substitution of one selected string for another.

2.2. Models and Conditions

The subject panel comprises Claude Haiku 4.5, GPT-5.6 Luna, and Gemini 3.8 Flash, accessed via OpenRouter with pinned providers and no fallback. Claude ran on its default tier; GPT and Gemini ran on the discounted flex tier with low reasoning effort. Temperature was 0.7 where accepted, and each request ran in a fresh conversation with a strict JSON schema.

The primary condition presented the alias without comments, with four repetitions per name–firm–model cell (12,960 requests). The disclosed condition added one sentence: "The company's displayed label was assigned arbitrarily for this study and conveys no information about the business," with two repetitions (6,480 requests). Comparators used the same undisclosed task with the heading "The company," numeric identifiers, or the 30 fictitious company names of Fenneman et al. (2023), whose fluent names are mostly real surnames or place names and whose disfluent names are invented strings. A name-only task elicited a one-year return in nine categories from –40% to +40%. Two variants on 30 pairs added a random-assignment statement or an assertion that pronounceable names signaled quality.

A prespecified rule, calibrated on a held-out pilot of 30 pairs (5,400 requests), set $J = 60$, the smallest sample size with at least 90% simulated zero-effect equivalence power for both

estimands at $\alpha = .0125$. Collection ran for seven days from September 16, 2026; substantive requests were complete by September 18, with daily sentinels thereafter. All 27,762 substantive requests returned a valid response. The entire study cost \$23.80 in API fees. Because the inferential unit is the matched pair, the primary tests have an effective sample of 60 pairs.

The design plan, stimuli, prompts, and hash manifest binding the code and configuration were uploaded to OSF storage before the first confirmatory request. The record registration was filed on September 23, after data collection and before confirmatory analysis, with the design unchanged. Supplement S1 documents the registration history.

2.3. Outcomes and Inference

For each valid positive valuation V , I define $Z = 100 \ln(V)$, so that differences in Z are log points (approximately percentages), and average Z over valid repetitions in each cell. The registered analysis did not use trimming or winsorization. For each pair, the fluent-minus-disfluent cell differences are averaged equally across firms and models; the primary estimand θ_2 is their mean over the 60 pairs, and $\exp(\theta_2/100)$ is the geometric valuation ratio. The valuation test asks whether the ratio lies between 0.99 and 1.01, a log-point interval of $[-1.005, 0.995]$. The 1% tolerance is one-tenth of the 10% decision hurdle used below, and it neither measures pricing accuracy nor bounds each alias-specific effect. The companion estimand θ_1 is the analogous return gap in percentage points, with bounds ± 2 .

For each outcome, a two-sided zero-null test and two one-sided equivalence tests were evaluated (Lakens, 2017), and a Holm adjustment was applied to the four statements. Both nonzero and equivalent can hold. Establishing neither is inconclusive. Inference uses t intervals with $J - 1$ degrees of freedom and a paired bootstrap as a sensitivity check. Secondary quantities (disclosure contrasts, label-policy ratios, instruction contrasts) carry pointwise intervals without

the confirmatory family's error control and no verdict. Analyses outside the registration are labeled exploratory.

3. Results

3.1. Valuation Test

Table 1, Panel A, reports the confirmatory statements. The estimated gap favors fluent names ($\theta_2 = 0.49$ log points, $SE = 0.24$; Holm-adjusted $p = .094$ for a nonzero gap and $.066$ for equivalence), corresponding to a geometric ratio of 1.0050 (95% CI [1.0001, 1.0099]), or approximately \$0.25 on a \$50 valuation. The ordinary 90% equivalence interval, [0.09, 0.90], lies within the tolerance, and the ordinary 95% interval excludes zero. However, neither conclusion survives the registered Holm adjustment. For scale, single-query noise (pilot within-cell standard deviations of 4 to 22 log points) and reference-price controls (1.1 to 34.8 log points) were much larger (Supplements S6 and S7.5). The valuation test is inconclusive; the gap is small but cannot be certified below 1%.

Figure 1 shows the model-specific estimates, which are descriptive; the registered verdict applies only to the pooled estimand (Supplement Table S7). The design achieved its planned precision (pair-difference standard deviation 1.89 log points, projected 1.91), but the estimate lies near half the upper bound, a pre-specified planning scenario with 41% simulated power to establish equivalence (Supplement Table S5). In exploratory checks, excluding 47 Claude responses more than a factor of two from their cell medians, or using cell medians, leaves θ_2 between 0.48 and 0.56 (Supplement S7.7).

3.2. Label Policy and Disclosure

Figure 2 presents secondary comparisons with an unnamed profile. The generated names lowered valuations by 2.6% for Claude, 0.9% for GPT, and 0.8% for Gemini, with pointwise

ratio intervals that exclude 1 (Table 1, Panel B). Because the registered 1% bound applies to a different estimand, these estimates show sensitivity to the labeling policy without testing it against that bound. Exploratory direct contrasts, in which the unnamed reference cancels, place the generated names 1.5 log points below the numeric identifiers $[-2.2, -1.0]$ and 1.3 below the published fluent names $[-2.1, -0.2]$, pooled over models, but only 0.6 below the published disfluent names $[-1.3, 0.1]$. With cell medians, only the first contrast excludes zero, although the discount maintains its sign for every model (Supplement S7.7). The discount mainly attaches to the generated names and possibly to invented strings more generally.

The disclosure sentence raised pooled valuation levels by 0.58 log points $[0.23, 0.93]$, the largest for GPT (1.62 $[1.25, 1.99]$), and reduced the pooled fluent-minus-disfluent gap by 0.96 log points $[0.19, 1.73]$, leaving a disclosed gap of $-0.46 [-1.05, 0.12]$. Model-specific changes in the gap (Supplement Table S8) are descriptive: the design tests no contrast between them, and differing interval coverage of zero does not indicate model heterogeneity.

In an offline buy/hold/sell replay against each firm's stored price with a $\pm 10\%$ hurdle, the two members of a pair are classified differently in 28.8% of the comparisons $[26.8, 30.9]$. Much of this disagreement reflects query noise near the thresholds. On a two-repetition basis, exploratory between-name disagreement exceeded within-name disagreement by only 2.6 pp $[1.0, 4.1]$ (Supplement S7.7).

3.3. Name-Only Returns

On the string-coded response scale, 1,436 of the 1,440 returned guesses for the generated names were exactly zero, leaving almost no variation to be tested. The pooled gap, $\theta_1 = 0.07$ percentage points $[-0.07, 0.21]$, passes the registered equivalence test (Holm-adjusted $p < .001$), and the zero-null test is not rejected ($p = .321$). The pattern is degenerate and encoding-

dependent (Supplement S5): equivalence by the absence of variation, as the registration anticipated, rather than evidence against name bias.

The instruction variants serve only as an instruction-response diagnostic. Across the 30 diagnostic pairs, the random-assignment statement did not detectably change the gap (-0.10 points $[-0.40, 0.19]$), whereas the assertion that pronounceable names signal quality widened it by 8.79 points $[6.65, 10.92]$: the models separated the two groups once told which feature mattered. Return guesses fell for both groups under that assertion, and further for disfluent names. However, without an unnamed-instruction control, this does not identify a general discount for invented names. Human benchmarks (Fenneman et al., 2023) appear in Supplement Table S10 as context only because the populations and protocols differ.

4. Conclusion

This study audits an alias policy that LLM valuation experiments treat as neutral. For these profiles and model configurations, the registered fluent-minus-disfluent valuation contrast remains unresolved: the estimated gap of approximately 0.5% is neither certified below the 1% tolerance nor distinguished from zero. Secondary estimates suggest that the labeling policy itself can shift valuation levels by 0.8% to 2.6% relative to an unnamed profile, and that one sentence about the label can move levels and the name gap; these point estimates exceed the registered gap, but the design tests do not rank the two.

Practical protection depends on the assignment of aliases and the treatment estimand. A level shift common to all aliases cancels in a treatment contrast only under that contrast's additive and scale assumptions; a common log-point shift need not cancel in dollars. Aligning alias policies with arms creates confounding, whereas independently randomized aliases add sampling variability, and a single realized assignment can be unbalanced.

Researchers should (1) report the assignment rule, (2) balance arms with a common alias bank, and (3) model dependence among repeated uses of an alias, for example, with a random effect. A shared alias makes the contrast conditional on that label; interactions with other experimental treatments remain untested. An arbitrary-label sentence is no substitute, because it changes the estimates. Any tolerance should match the outcome and treatment effect of interest.

This study has several limitations. First, the treatment is a selected bundle of spelling features; therefore, learned associations, inferred meaning, and tokenization remain possible channels (Supplement S7.5). Second, the scope is limited to English-like seven- and eight-letter strings shown once in a heading, nine profitable fictional profiles, and three low-cost model configurations in one collection window; repeated mentions and corporate-style labels such as "Company A" were not tested. Third, the fluent group is limited to pronounceable names that survived the specified company, product, and brand screens. Fourth, the name-only task is degenerate, and a pilot change in response encoding between runs coincided with changed answers, mostly for one model.

The most useful extension is a direct test of the alias-by-treatment interactions. Other extensions include non-English names, corporate-style labels, loss-making firms, and new model versions.

Data Availability

The registered design plan, pre-analysis amendment, prompts, stimuli, firm profiles, parsed response ledgers (values, request identifiers, and timestamps), analysis code, registered outputs, and exploratory scripts with their outputs will be deposited in the study's OSF project, which will be made public once the registration embargo is lifted and no later than publication. The registration of records, including the frozen design plan, materials archive, and dated hash manifest, can be inspected during the review via an anonymized view-only link:

https://osf.io/ewcmf/overview?view_only=6a4b32bc1ae04b3fb25d9f588df301c2. The project files, including the pre-analysis amendment, plan, and hash manifest uploaded before the confirmatory collection, identify the author; a view-only link to them is supplied to the editors on the title page. Verbatim model responses will be shared, subject to the providers' terms. This study did not involve human participants; its data were model outputs, and its stimuli were generated for this study or obtained from published research (Fenneman et al., 2023).

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this study, the author used Claude (Anthropic; Claude Opus 5 model, accessed through the Claude desktop application) for data engineering support, drafting analysis code, auditing reported numbers, and manuscript editing assistance. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

References

- Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences*, 103(24), 9369–9372.
<https://doi.org/10.1073/pnas.0601071103>
- Chen, S., Green, T. C., Gulen, H., & Zhou, D. (2024). *What does ChatGPT make of historical stock returns? Extrapolation and miscalibration in LLM stock return forecasts* [Working paper]. arXiv.
<https://arxiv.org/abs/2409.11540>
- Didisheim, A., Fraschini, M., & Somoza, L. (2025). AI's predictable memory in financial analysis. *Economics Letters*, 256, Article 112602. <https://doi.org/10.1016/j.econlet.2025.112602>
- Engelberg, J., Manela, A., Mullins, W., & Vulicevic, L. (2025). *Entity neutering* [Working paper]. SSRN.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5182756
- Fenneman, A., Janssen, D.-J., Nolte, S., & Zeisberger, S. (2023). Nomen est omen? How and when company name fluency affects return expectations. *PLOS ONE*, 18(8), Article e0287995.
<https://doi.org/10.1371/journal.pone.0287995>
- Glasserman, P., & Lin, C. (2024). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *Journal of Financial Data Science*, 6(1), 25–42.
<https://arxiv.org/abs/2309.17322>
- Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
<https://doi.org/10.1177/1948550617697177>
- Montone, M., van den Assem, M. J., & Zwinkels, R. C. J. (2023). Company name fluency and stock returns. *Journal of Behavioral and Experimental Finance*, 39, Article 100819.
<https://doi.org/10.1016/j.jbef.2023.100819>
- Sarkar, S. K., & Vafa, K. (2024). *Lookahead bias in pretrained language models* [Working paper]. SSRN. <https://doi.org/10.2139/ssrn.4754678>

- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2310.11324>
- Vitevitch, M. S. (2025). Examining Chat GPT with nonwords and machine psycholinguistic techniques. *PLOS ONE*, 20(6), Article e0325612. <https://doi.org/10.1371/journal.pone.0325612>
- Wu, K., Yang, B., Ying, Z., & Zhou, D. (2025). *Anonymization and information loss* [Working paper]. arXiv. <https://arxiv.org/abs/2511.15364>

Table 1*Confirmatory Statements and Secondary Estimates***Panel A. Confirmatory statements (Holm adjustment across the four statements)**

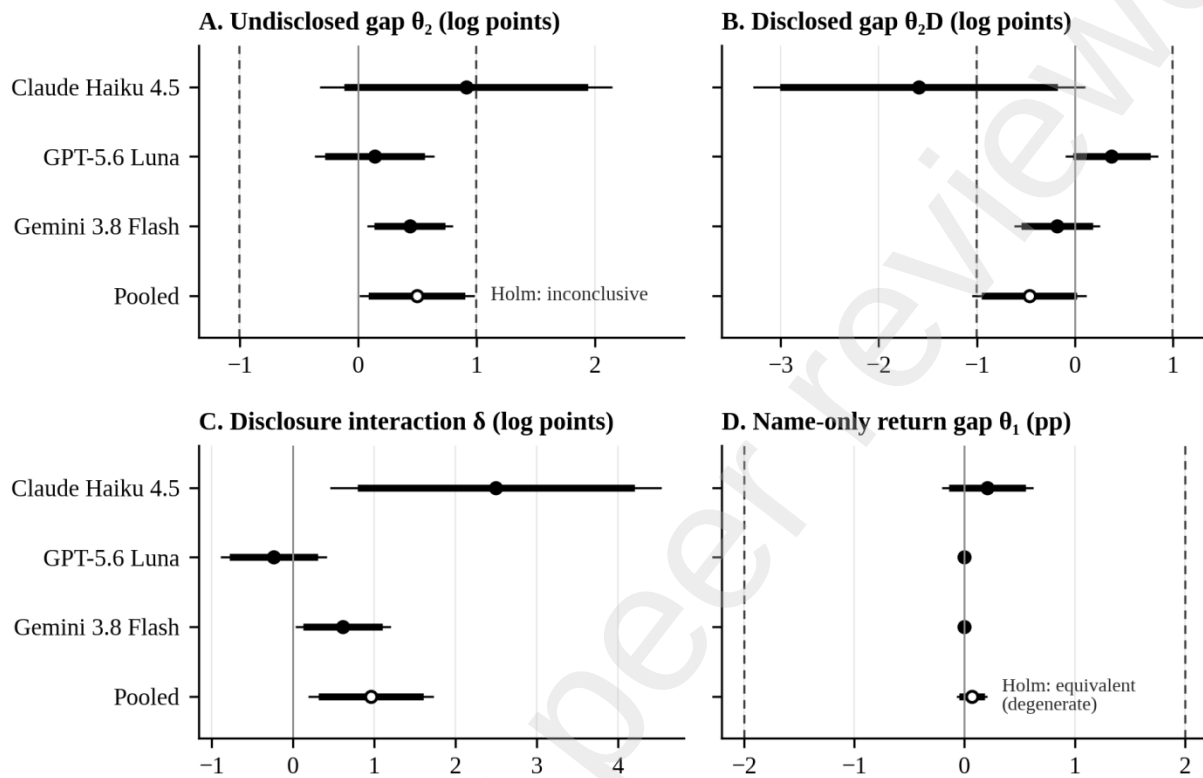
Quantity	Estimate	95% interval	Holm p , nonzero	Holm p , equivalence	Registered verdict
Valuation gap θ_2 , log points	0.49	[0.01, 0.98]	.094	.066	Inconclusive
Name-only return gap θ_1 , pp	0.07	[-0.07, 0.21]	.321	< .001	Equivalent, degenerate

Panel B. Secondary estimates (pointwise intervals, no verdict)

Quantity	Estimate	95% interval
Disclosed gap θ_2D , log points	-0.46	[-1.05, 0.12]
Disclosure interaction δ (undisclosed gap minus disclosed gap), log points	0.96	[0.19, 1.73]
Valuation level, disclosed minus undisclosed, log points	0.58	[0.23, 0.93]
Generated names relative to unnamed profile, ratio: Claude	0.974	[0.954, 0.994]
Generated names relative to unnamed profile, ratio: GPT	0.991	[0.983, 0.998]
Generated names relative to unnamed profile, ratio: Gemini	0.992	[0.987, 0.998]
Validity gap, primary and disclosed, pp	0.0	—

Note. $J = 60$ matched pairs, nine firms, and three models weighted equally. The valuation gap is conditional on valid responses; all 12,960 primary and 6,480 disclosed requests were valid, so no cell was empty and both full-panel estimates were available. θ_2 corresponds to a geometric valuation ratio of $\exp(\theta_2/100) = 1.0050$ [1.0001, 1.0099]. The equivalence bounds are [-1.005, 0.995] log points, a ratio of [0.99, 1.01], for θ_2 and [-2, 2] pp for θ_1 . Ordinary 90% equivalence intervals are [0.09, 0.90] for θ_2 and [-0.05, 0.19] for θ_1 ; conservative 97.5% equivalence intervals are [-0.07, 1.05] and [-0.09, 0.23], respectively. The unadjusted p values are presented in Supplement Table S6. θ_1 is degenerate: 1,436 of the 1,440 responses were zero. The intervals in Panel B are pointwise and carry none of the confirmatory family error controls. $\delta = \theta_2 - \theta_2D$, so a positive δ means that the sentence reduced the gap; the level row is the within-name disclosed-minus-undisclosed difference, whose registered sign convention (η) is the opposite. The θ_2D , δ , and level intervals are t intervals on 59 degrees of freedom, and the ratio intervals come from a nested pair, named and unnamed-repetition bootstrap. The model, firm, dispersion, decision replay, and sensitivity estimates are presented in the Supplement.

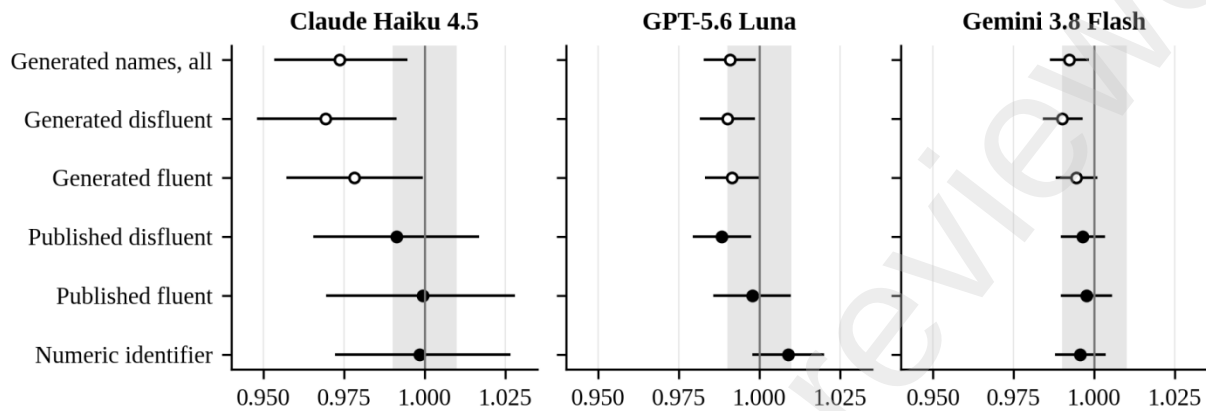
Figure 1
Fluent-Minus-Disfluent Contrasts by Model



Note. The panels show (A) the undisclosed valuation gap θ_2 , (B) the disclosed gap θ_{2D} in log points, (C) the disclosure interaction δ in log points, and (D) the name-only return gap θ_1 in percentage points. Points are estimates, with open circles for the pooled estimates. Thick bars are unadjusted 90% confidence intervals, and thin bars are unadjusted 95% confidence intervals, displayed without the Holm adjustment. The registered verdicts printed beside the pooled estimates in Panels A and D follow the Holm-adjusted tests in Table 1, not the displayed intervals. Dashed lines mark the registered practical bounds for θ_2 in panel A and for θ_1 in panel D; in panel B, they repeat θ_2 's bounds for reference only because θ_{2D} carries no equivalence verdict. The models appear in the order Claude Haiku 4.5, GPT-5.6 Luna, and Gemini 3.8 Flash, and the pooled row weighs them equally. Without instruction, every GPT and Gemini name-only answer for the generated names was zero; therefore, panel D shows no interval for GPT or Gemini. The figure redraws the registered output figure1_primary.png from the same result files.

Figure 2

Valuation Under Alternative Label Policies Relative to an Unnamed Profile



Geometric-mean valuation ratio to the unnamed profile (shaded: $\pm 1\%$ reference band, not an equivalence test)

Note. Geometric mean valuation ratios for each label policy relative to the unnamed profile ("The company"), by model, in the undisclosed condition, with nested pair, name, and unnamed-repetition bootstrap 95% intervals, which are pointwise. The shaded band marks ratios within 1% of the unnamed profile; no label policy receives an equivalence verdict. Open circles mark the study's 120 generated names, overall and by group. Published names are the 30 fictitious company names from Fenneman et al. (2023), 15 fluent and 15 disfluent. Numeric identifiers run from Entity 001 to Entity 009. Exploratory direct contrasts between these policies appear in Supplement Table S12. The figure redraws the registered output figure2_comparators.png from the same results file.