Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Shared Option-Order Errors in LLM-Based Social Measurement: Evidence and Design Rules

Tamas Olah

olah.tamas@econ.unideb.hu

University of Debrecen

Levente Nadasi

University of Debrecen

Marianna Abuczki

University of Debrecen

Tibor Tokes

University of Debrecen

Research Article

Keywords: large language models; social measurement; response-order effects; measurement error; research design; survey methodology

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11720269/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Shared Option-Order Errors in LLM-Based Social Measurement: Evidence and Design Rules

Abstract

Large language models (LLMs) serve as measurement instruments when researchers apply a fixed prompt to many texts or questions. A fixed prompt standardizes presentation, but incidental presentation features can push many responses in the same direction. We study one such feature, response-option order: when does its error survive aggregation, and which remedies remain effective when the instrument changes? Using responsibility and solution judgments on 150 European Union policy issues, we presented all 24 orders of four options twice, in four languages and 31 configurations of 21 open-weight models (1,771,200 responses). Comparing the two replications separates reproducible order effects from generation noise and isolates the part shared across items, which, like an interviewer effect in surveys, does not average out. In English, averaging 150 issues under one fixed order cancels order error about as well as averaging three unrelated issues, and conventional 95 percent intervals cover the order-averaged reference in only 49.0 and 27.1 percent of cases (medians). Drawing the order independently for each item requires no additional calls and reduces estimated mean squared error in all 246 replicated conditions. This design rule holds across languages, instructions, labels and temperatures. Fitted corrections depend on language, policy domain and task: out-of-domain correction often helps, but cross-task position corrections usually underperform local fits. Reasoning can increase residual error, and prespecified calibration gains do not repay pilot costs. We translate these results into a design and reporting workflow for LLM-based social measurement.

Keywords: large language models; social measurement; response-order effects; measurement error; research design; survey methodology

1 Introduction

Researchers use large language models (LLMs) to turn texts and questions into social measurements. Researchers classify documents, place political actors on scales and elicit numerical judgments by applying a common prompt to many items (Argyle et al. 2023; Gilardi et al. 2023; Le Mens and Gallego 2025; Halterman and Keith 2026). Fixing the prompt makes a procedure easy to describe and repeat. It does not establish that incidental features of its presentation leave the resulting measures unaffected. Guidance for LLM annotation already stresses validation and prompt-stability checks (Törnberg 2024), and a recent review of computational political measurement emphasizes comparing the assumptions and context of different instruments (Parschan and Jakob 2026). An LLM prompt, like any instrument, should be judged by the error it leaves in the quantities researchers report, not only by whether it returns the same answer twice.

Consider an instrument that asks who bears responsibility for 150 European Union (EU) policy problems and which approach should solve them. Merely changing the order of four response options makes `granite4:3b` hold EU institutions most responsible for anywhere from 0 to 80 issues. The issue texts, model and substantive question are unchanged. A researcher could nevertheless report a different account of responsibility, accompanied by a small standard error calculated across the same issues. The problem is not just that a response changes when its prompt changes. It is that one presentation can move many responses in the same direction.

Survey research has long studied both parts of this problem. Answers depend on the order in which response options are offered, with primacy effects typical of visual presentation and recency effects of oral presentation (Krosnick and Alwin 1987; Krosnick 1991; Galesic et al. 2008). Questionnaire designers therefore rotate or randomize option order across respondents, so that order effects vary rather than accumulate in one direction. More generally, errors such as interviewer effects and common method variance can have a shared component that is not reduced simply by collecting more observations (Kish 1962; Podsakoff et al. 2003; Groves and Lyberg 2010). LLM measurement can reproduce this problem in a stark form. A study that applies one prompt to every item uses the same

respondent and the same option order throughout, so any position preference is repeated across the entire corpus.

LLMs are known to be sensitive to option order, labels and prompt formatting (Pezeshkpour and Hruschka 2024; Zheng et al. 2024; Sclar et al. 2024), including in survey-style questionnaires and political opinion probes (Tjuatja et al. 2024; Dominguez-Olmedo et al. 2024; Röttger et al. 2024); on World Values Survey items, almost all tested models favor the last-listed option (Rupprecht et al. 2026). Randomizing or fully crossing option order is already used to remove such bias: Dominguez-Olmedo et al. (2024) randomize answer order before comparing models with survey populations, Brucks and Toubia (2025) show that aggregating over a full factorial of prompt features eliminates order and label bias, and evaluator studies and political-bias audits present both orders or average over perturbations (Wang et al. 2024; Debevc et al. 2026). Design choices also change conclusions: prompt choice alters annotation accuracy (Abraham et al. 2025), plausible configurations yield incorrect conclusions for a third to a half of tested hypotheses (Baumann et al. 2025), and labels that are reliable under one task design shift under another (Reiter et al. 2026). Reliability procedures for LLM coding accordingly assess consistency across repeated generations (Hsu and Hsu 2026) and semantically similar prompts (Barrie et al. 2024). We ask what an unrandomized order costs an aggregate and which design recovers it at a fixed budget. Our comparison follows order error from individual responses to corpus averages, then evaluates common and item-level assignment, position balance and fitted corrections across changes in model, language and instructions. A model can be highly repeatable under each order and still yield different aggregate conclusions across orders.

Recent statistical work directly addresses uncertainty that survives aggregation. Messing (2026) decomposes LLM pipeline variance, demonstrates undercoverage as item counts grow and uses design studies to allocate measurement effort; the study also examines position counterbalancing in pairwise judgments. Leng (2026) defines prompt-averaged estimands over governed prompt families and carries their uncertainty into downstream inference, while Helm and Priebe (2026) model correlated perturbation effects across

personas and derive valid tests. For LLM judges, Feng et al. (2026) derive unbiased estimators that separate prompt, answer-order and call-level variation. Our study develops an empirical design comparison for a fully enumerated nuisance factor: the 24 permutations of a four-option instrument. A uniform average over the permutations defines a transparent target within each model and instrument. Enumerating every permutation twice supports direct replicated diagnosis of shared order error and matched-call comparisons of common versus item-level assignment, position balance and fitted correction. The language, domain and task transfer tests and full-budget evaluation ask which remedies remain useful after instrument changes and pilot costs.

The paper makes three contributions. First, it provides a diagnostic for aggregate measurement. With two replications of the full order set, standard repeated-measurement moments separate reproducible order effects from generation noise and estimate the component shared across items, which we summarize as an effective number of items analogous to Kish's design effect. Second, it supplies broad evidence on the consequences and on the simplest remedy. Across 246 replicated language–task–configuration cells (1,771,200 responses), one common order leaves an error that conventional standard errors miss, whereas assigning orders independently across items reduces mean squared error in every cell at no additional cost. Third, it tests correction portability across languages, policy domains and tasks, the reasoning-related balance of order and generation error, and whether calibrated designs repay their pilot costs. Section 7 translates these results into a workflow, and the replication materials provide software for assignment and diagnosis.

The EU judgments provide a common instrument for studying this methodological problem. They are not estimates of citizens' opinions or verified assessments of political responsibility. Congressional Bills coding and a labeled stance check supply secondary extensions with external reference labels (Supplementary Information, Online Resource 1, Sections A3.2 and A7.6). Throughout, we keep three questions separate: whether outputs repeat, whether estimates are stable to incidental presentation, and whether the resulting target measures a substantively valid construct. Section 2 describes the instrument and evidence, Section 3 develops the shared-error diagnostic, Sections 4–6

report the assignment comparisons, instrument changes and limits to fitted repairs, and Sections 7–8 give recommendations and discuss their scope.

2 An instrument audit using EU policy judgments

2.1 Questions and option assignment

An LLM call sends a text prompt to a model and receives generated text, which the researcher parses into numbers or labels. Fig. 1(a) shows two calls from our EU instrument. The prompt describes an issue, lists four candidate actors labeled A to D and asks for a responsibility score from 0 to 100 for each. Because models generate answers by sampling, two calls with an identical prompt can return different scores even at the low temperature (0.2) we use; we call this generation noise. A model can also respond systematically to where an option appears. In the example, the score for external global factors falls from 85 to 35 when the options are listed in reverse order, while the score for member-state governments rises from 40 to 75. An independent second replicate returned exactly the same scores, reproducing the order difference across blocks. Because the labels A to D follow the displayed position, a position effect also includes any preference for particular letters, which Zheng et al. (2024) identify as a main source of option-order sensitivity in multiple-choice questions.

Four options can be listed in 24 orders. Fig. 1(b) shows four ways to assign them to items. A common design uses one fixed order for all items. Drawing an order at random for each item uses the same number of calls. With four calls per item, a researcher can present a balanced set: four orders in which each option appears once in every position, as in a Latin square. A balanced set cancels effects that depend only on an option's position, such as a tendency to favor the first-listed option, but not effects that depend on which options surround it. One balanced set can again be used for every item, or a balanced set can be drawn at random for each item. Independent assignment breaks the connection between one presentation and the whole corpus; Section 3 formalizes the resulting risk.

(a) Same issue, original and reversed option order

Inflation in the eurozone	
A Member-state governments	**40**
B EU institutions	**20**
C Economic actors	**65**
D External global factors	**85**

Inflation in the eurozone	
A External global factors	**35**
B Economic actors	**65**
C EU institutions	**20**
D Member-state governments	**75**

(b) Assigning four-option orders to six items

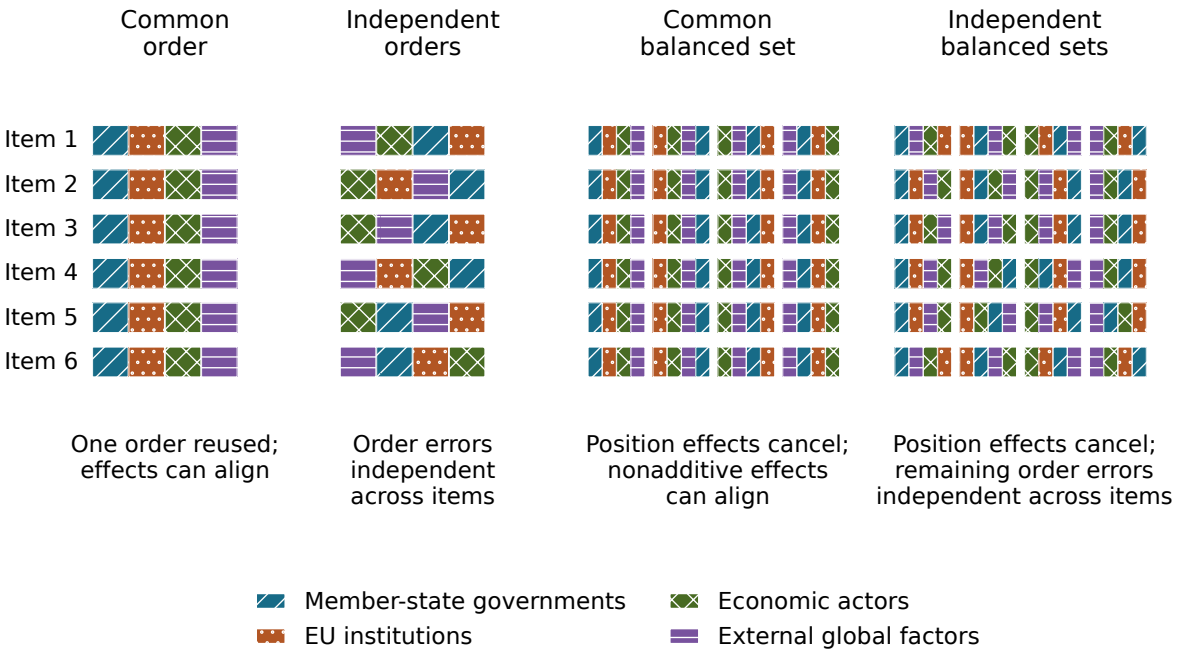


Fig. 1 Option order in LLM measurement (2026 EU corpus). (a) Responsibility scores from `gemma4:e4b` (thinking off) for the first issue in the EU corpus under the original and the reversed option order; a second replicate block returned identical scores. This issue's order variance lies at the 77th percentile for this model. (b) Four designs for assigning the 24 orders of four options to items, shown for six items. Each bar is one call; colors and hatch patterns identify the substantive option in each displayed position

2.2 A crossed EU corpus

The main EU corpus contains 150 issues in 11 policy domains, including eurozone inflation, housing and disinformation; the calibration panels contain subsequently collected responses. Responsibility scores concern member-state governments, EU institutions, economic actors and external global factors. Solution scores concern national sovereign action, closer EU integration, market self-regulation and social responsibility. The instructions require a clear score spread and forbid abstention; variants relax these restrictions (Supplement A1).

The frozen source corpus contains 2,649,600 saved responses in 488 files, spanning 25 model tags and 35 tag/setting combinations (Table 1). A model tag identifies the loaded model and size; a configuration combines that tag with its recorded thinking setting. Off and on denote explicit settings, low and medium denote effort levels, and unset means no explicit control was recorded; it does not establish an absence of reasoning. Language, prompt wording and temperature are separate instrument conditions. Fig. 2 lists every tag/setting combination and shows which conditions were collected.

The replicated analysis uses 31 configurations of 21 tags, about 2–20 billion parameters. Each issue receives all 24 orders twice at temperature 0.2. German, French and Spanish retain the issue identities and task structure. Excluding the two incomplete Spanish cells leaves 246 complete cells and 1,771,200 responses. The four larger tags have one block in every language; the other instrument arms use 13 configurations of ten tags. The conditions are not fully crossed, and token budgets need not match across reasoning modes.

The EU corpus was collected as a unified study of how task framing (responsibility versus solution), language, prompt variants and model configuration affect LLM political judgments. These dimensions were part of the original collection plan. The shared-error diagnostics and fitted-repair analyses were subsequently developed using the saved responses. Early order analyses also informed calibrated designs whose requests, parsers, policies and endpoints were frozen before 62,400 fresh evaluation calls were collected.

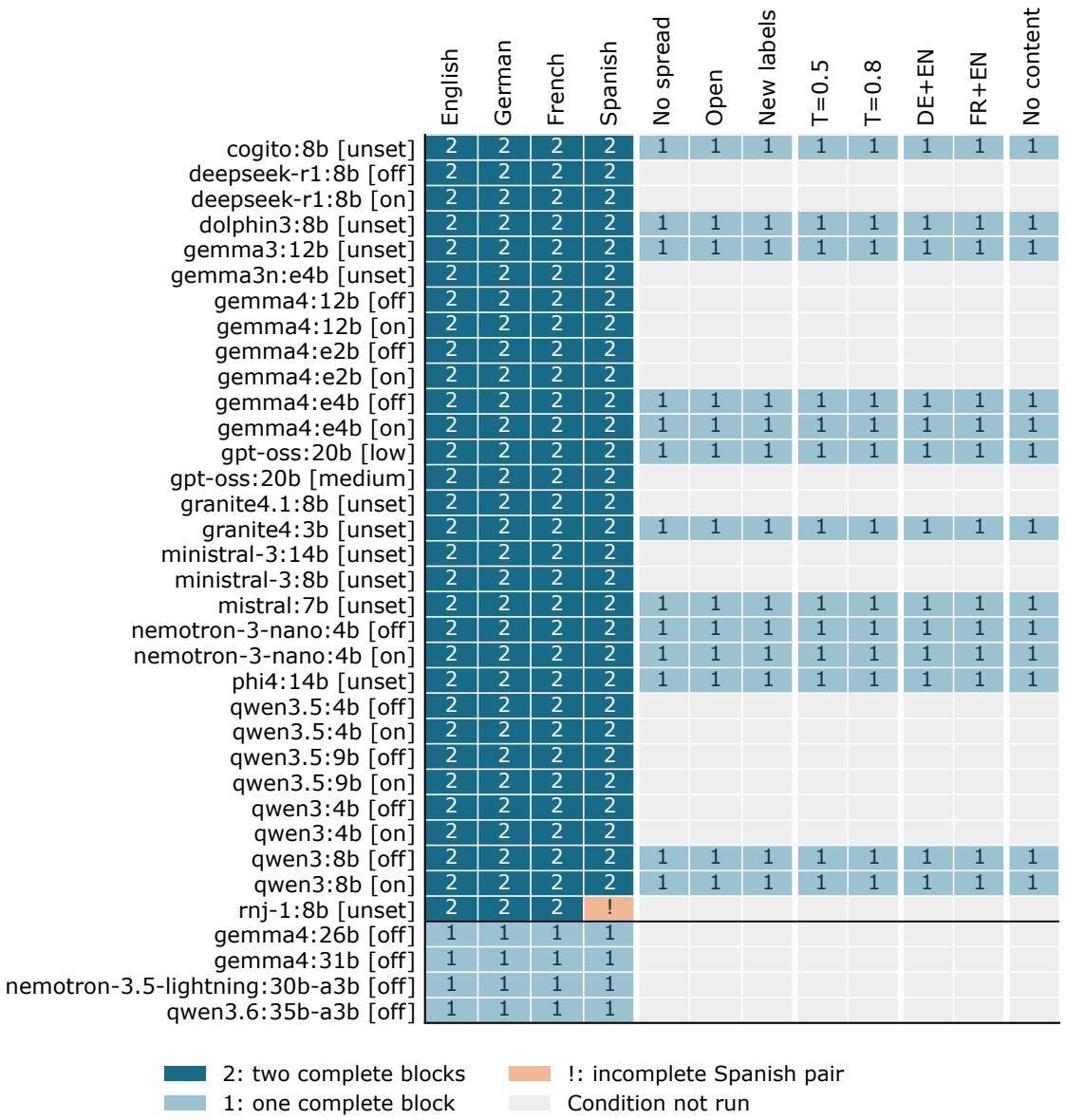


Fig. 2 The full EU collection: every model tag, recorded thinking setting and available instrument condition. Numbers show complete blocks of all 24 orders in both tasks on 150 issue identities; gray cells were not run. The Spanish `rnj-1:8b` pair is incomplete. Off/on are explicit controls, low/medium effort levels and unset no explicit control. No spread relaxes score-spread instructions; open additionally relaxes anti-hedging instructions. DE+EN and FR+EN use translated issues with English instructions. The bottom four rows are the larger-model arm

Table 1 Scope of the EU evidence (2026): the frozen source corpus and fresh calibration

EU evidence	Tags/settings	Languages and blocks	Saved rows
Replicated English	21/31	English; two blocks	446,400
Replicated translations	21/31	German, French, Spanish; two blocks	1,324,800
Incomplete Spanish	1/1	Spanish; excluded from replicated analysis	14,400
Larger-model extension	4/4	All four languages; one block	115,200
Instrument variants	10/13	Seven language–prompt–temperature arms; one block	655,200
Content-free prompt	10/13	English; one block	93,600
Frozen EU source total	**25/35**	Four languages; two tasks	**2,649,600**
Fresh EU calibration	4/4	English and German; two fresh blocks	23,040

Note: A setting is a recorded thinking control; tasks, languages, instructions, labels and temperature define additional conditions. Source totals include 10 malformed rows among the 14,400 incomplete Spanish rows; both affected cells are excluded from the replicated analysis. The other 2,649,590 source rows are accepted. Section 5 reports the larger-model row's 28,800 English calls. Fresh EU calibration uses 60 solution issues per language; the prespecified collection also includes 960 pilot and 38,400 Bills calls (Supplement A6).

2.3 Targets and uncertainty

We evaluate conventional per-option 95 percent intervals for the average score profile, using standard errors (SEs) computed across issues. Their reference is the order-balanced mean of the independent opposite block. Fixed designs average all orders or balanced sets; random designs average 1,000 assignments. Summaries are medians across fixed configurations.

Bracketed 95 percent ranges describe issue-composition sensitivity with models and recorded generation blocks fixed. Monte Carlo (MC) standard errors instead quantify the numerical precision of sampled assignments or pilot splits, conditional on those responses. We identify them separately in each display; neither measures uncertainty from future model generations or a population of models (Supplement A3). Replication

materials (Online Resources 2 and 3) reproduce the results from saved responses and cached embeddings without calling a model. AI tools assisted with literature search, text editing and analysis code; the full disclosure appears in the AI-use statement.

3 From repeatability to shared measurement error

The following diagnostic applies standard repeated-measurement and sampling identities to the finite order set (Searle et al. 1992; Brennan 2001; Kish 1962). Its purpose is to identify the error relevant to an aggregate and make the assignment comparison explicit.

3.1 Target and components

Let X_{qpr} be the output for item $q = 1, \ldots, Q$ under order $p = 1, \ldots, P$ in replicate r. For EU issues it is the vector of four scores, arranged by option rather than by displayed position and centered within the call so that it records the relative profile; for coding tasks it is a one-hot vector of categories. Squared norms $\|x\|^2$ and inner products $\langle x, y \rangle$ average over coordinates. Because the order is incidental to the substantive question, we define the target as the order-averaged output,

$$\theta_q = \frac{1}{P} \sum_{p=1}^{P} \mathbb{E}(X_{qpr} \mid p), \tag{1}$$

the model's expected answer when the order is chosen uniformly at random. This is a property of the model and instrument, not a claim about human opinion. Write

$$X_{qpr} = \theta_q + \alpha_{qp} + \varepsilon_{qpr}, \qquad \sum_{p} \alpha_{qp} = 0, \quad \mathbb{E}(\varepsilon_{qpr} \mid p) = 0, \tag{2}$$

where α_{qp} is the reproducible order effect and ε_{qpr} is generation noise, independent across calls. Order variance is $\tau_q^2 = P^{-1} \sum_p \|\alpha_{qp}\|^2$ and generation variance is $\sigma_q^2 = P^{-1} \sum_p \mathbb{E} \|\varepsilon_{qpr}\|^2$.

With two complete replicate blocks, both components are identified:

$$\widehat{\sigma_q^2} = \frac{1}{2P}\sum_p \|X_{qp1} - X_{qp2}\|^2, \qquad \widehat{\tau_q^2} = \frac{1}{P}\sum_p \left\langle X_{qp1} - \bar{X}_{q\cdot 1}, X_{qp2} - \bar{X}_{q\cdot 2}\right\rangle, \tag{3}$$

where $\bar{X}_{q\cdot r}$ averages replicate r over orders. Paired differences remove the order effect, and the cross-product of independent replicates removes generation noise in expectation; both estimators are unbiased (Supplement A2). With only one call per item and order, the two components are confounded.

3.2 Shared order effects

Researchers rarely care about a single item. For the item average $\bar{\theta} = Q^{-1}\sum_q \theta_q$, what matters is the order effect averaged over items, $\bar{\alpha}_p = Q^{-1}\sum_q \alpha_{qp}$. Define the shared order variance and its estimator,

$$\tau^2_{\text{sh}} = \frac{1}{P}\sum_p \|\bar{\alpha}_p\|^2, \qquad \widehat{\tau^2_{\text{sh}}} = \frac{1}{P}\sum_p \left\langle \bar{X}_{\cdot p1} - \bar{X}_{\cdot\cdot 1}, \bar{X}_{\cdot p2} - \bar{X}_{\cdot\cdot 2}\right\rangle, \tag{4}$$

where dots denote averages over items or orders. If order effects were unrelated across items, τ^2_{sh} would be about $\bar{\tau}^2/Q$, with $\bar{\tau}^2 = Q^{-1}\sum_q \tau_q^2$; if every item had the same order effect, it would equal $\bar{\tau}^2$. We summarize the alignment of the item effects by the effective number of items,

$$n_{\text{eff}} = \frac{\bar{\tau}^2}{\tau^2_{\text{sh}}}, \qquad \frac{Q}{n_{\text{eff}}} = 1 + (Q-1)\bar{\rho}, \tag{5}$$

where $\bar{\rho}$ is the average between-item order covariance divided by the average item-level order variance (Supplement A2). It equals the average pairwise correlation when item variances are equal. At positive population variances, $n_{\text{eff}} \geq 1$, but it can exceed Q if item-order effects offset; it is undefined when the shared variance is zero. These are population properties, not finite-sample bounds for the estimated ratio. The second expression has the form of Kish's interviewer design effect (Kish 1962), with option order in the role of the interviewer.

3.3 Designs and inference

Consider one call per item. If one order, chosen uniformly at random, is used for every item, the item average has mean squared error (MSE) $\tau_{\text{sh}}^2 + \bar{\sigma}^2/Q$, where $\bar{\sigma}^2 = Q^{-1}\sum_q \sigma_q^2$. If the order is drawn independently for each item, the MSE is $(\bar{\tau}^2 + \bar{\sigma}^2)/Q$ with the same calls. The ratio of the two is

$$1 + s\left(\frac{Q}{n_{\text{eff}}} - 1\right), \qquad s = \frac{\bar{\tau}^2}{\bar{\tau}^2 + \bar{\sigma}^2}. \tag{6}$$

Independent assignment improves MSE when $\tau_{\text{sh}}^2 > \bar{\tau}^2/Q$; negatively associated item-order effects can instead offset under a common order. The improvement holds in every replicated cell evaluated here. For example, with an order share of $s = 0.6$ and $n_{\text{eff}} = 3$ among $Q = 150$ items, one fixed order has about 30 times the MSE of randomized orders. Reducing generation noise alone leaves order effects intact: for fixed order effects and $n_{\text{eff}} < Q$, it raises s and with it the ratio, which reaches Q/n_{eff}, about 50 here, under deterministic decoding. Outputs then become reproducible without becoming robust to order. A persistent common order component need not shrink as items are added. It is also invisible to the usual standard error, which is computed from the spread of outputs across items: a shift common to all items leaves that spread unchanged. Under a fixed order, the conventional interval therefore narrows around a displaced estimate as Q grows, much as in the "big data paradox" of Meng (2018). An interval for behavior under that particular order answers a different question. Conclusions intended to average over order require the corresponding presentation uncertainty. Under per-item randomization, order effects become independent across items and enter the spread that the standard error measures.

A balanced set cancels the part of each α_{qp} that is additive in option and position. One balanced set used for all items therefore removes the additive shared component but leaves the nonadditive part shared, whereas a balanced set drawn independently for each item makes all remaining order error independent across items. All four designs include

every order equally often in expectation, so they estimate the same target. Supplement A2 derives the corresponding MSEs.

4 Consequences for aggregate measurement

4.1 Why more issues do not remove the error

Order effects contribute substantially to individual answers in both tasks (Table 2). More importantly, much of that error is common across issues. For the median English configuration in each task, averaging 150 issues under one fixed order cancels order error about as well as averaging three unrelated issues. The effective item count is far below 150 in every replicated cell, including the translated instruments.

Table 2 Order and generation components in the English EU instrument (2026)

	Responsibility	Solution
Order share of single-call variance (%)	59.7 [56.8, 60.9]	64.3 [61.5, 65.6]
Order SD (points)	8.3 [8.0, 8.5]	7.9 [7.4, 8.1]
Generation SD (points)	8.2 [7.6, 8.5]	6.3 [6.2, 6.5]
Effective number of issues	3.3 [3.0, 3.6]	2.6 [2.3, 2.7]
Normalized order covariance	0.30 [0.28, 0.33]	0.38 [0.37, 0.43]
MSE ratio, fixed/random order	25 [23, 27]	40 [38, 43]

Note: Rows are separate medians over 31 configurations, each with 150 issues, 24 orders and two blocks. Brackets are central 95 percent ranges from 3,000 whole-issue resamples with configurations and blocks fixed. The last two rows use Equations 5 and 6. SD denotes standard deviation, measured on the centered relative profile.

Fig. 3 shows all 31 English configurations, including the recorded off/on pairs and both effort levels, rather than selecting a few examples. The size of the conclusion change varies substantially across configurations. The substantive example in the introduction is therefore not an isolated unstable answer. Under fixed orders, `granite4:3b`'s count of issues attributing the greatest responsibility to EU institutions ranges from none to more than half of the corpus; the exact blockwise central 95 percent assignment ranges under

independent assignment span roughly 15–30 issues. Across configurations, order can also change which option is rated highest on the most issues (Supplement A3.1).

The component is largely additive in option and position, so four-order position balance removes much of it. Its direction differs across tasks and configurations, however; a universal primacy correction is inappropriate. Order sensitivity is also greatest on close calls: the gap between the two highest order-averaged scores is negatively rank-correlated with order variance. Supplement A3.4 reports these patterns and their uncertainty.

4.2 The inferential consequence

Table 3 compares the four assignment designs. One fixed order leaves a shared shift that the conventional standard error misses. Independent per-issue assignment reduces the error at the same call count and raises coverage above the nominal level. Four calls with independent position balance further reduce error; reusing the same balanced set still leaves nonadditive shared effects.

Table 3 Assignment design determines precision for English EU issue averages (2026)

	Responsibility	Solution
A. Coverage of conventional 95% intervals (%)		
One fixed order	49.0 [45.3, 52.6]	27.1 [25.5, 30.7]
Random order per issue	99.6	98.5
One fixed balanced set	97.4 [93.2, 97.9]	74.5 [70.8, 79.7]
Random balanced set per issue	100.0	100.0
B. Error of the issue average (score points)		
RMS error, one fixed order	3.9 [3.8, 4.2]	4.6 [4.3, 5.0]
RMS conventional SE, one fixed order	1.4 [1.4, 1.5]	1.0 [1.0, 1.1]
RMS error, random order	0.9	0.8
RMS error, random balanced set	0.4	0.3

Note: Medians over the same 31 configurations and 150 issues. RMS denotes root mean square. One order uses one call per issue; a balanced set uses four. Fixed designs enumerate orders or sets; random designs average 1,000 assignments. Coverage and error use the opposite block's balanced mean. Brackets are 95 percent issue-resampling sensitivities; assignment MC precision is reported in Table 4.

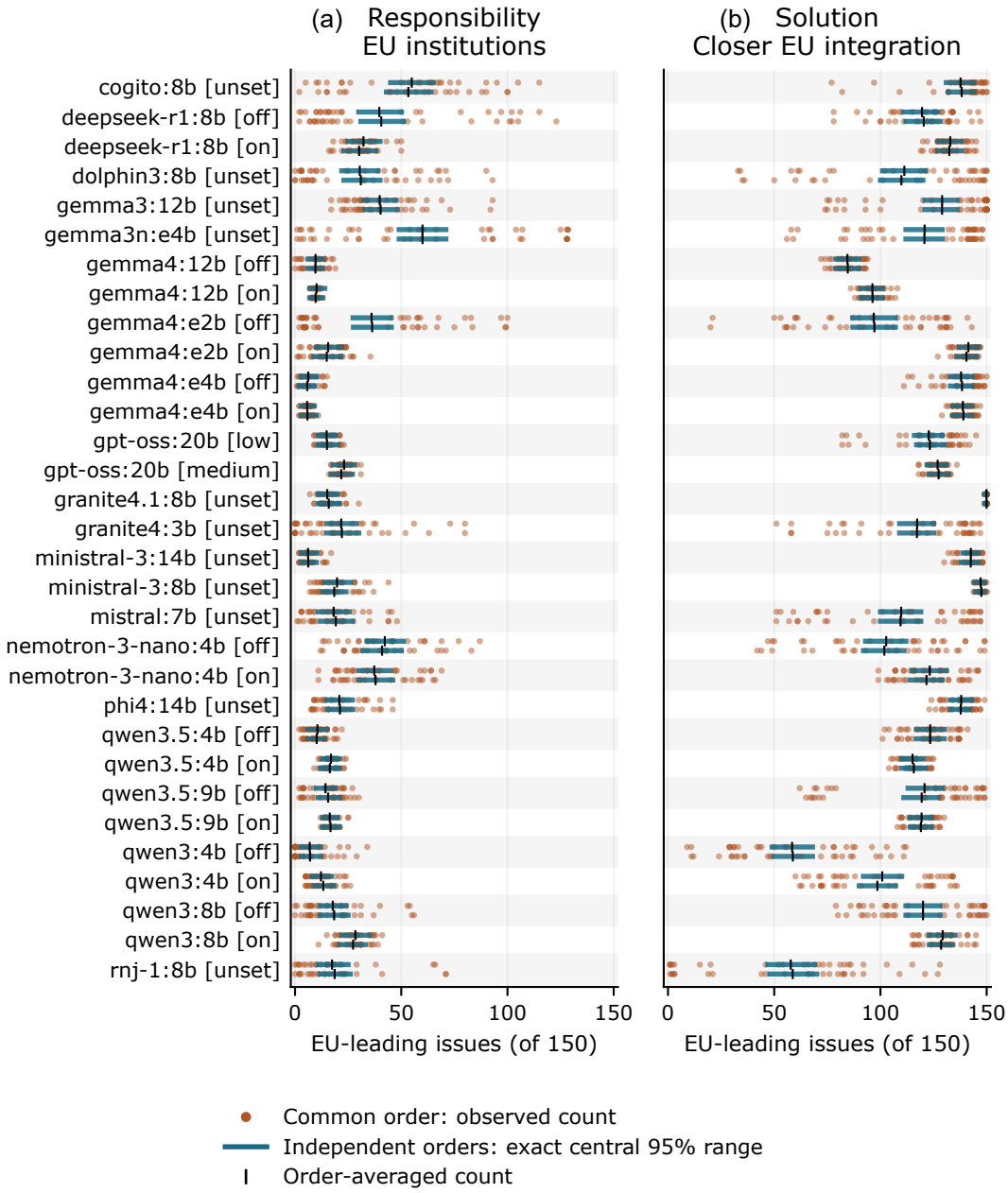


Fig. 3 EU-leading issue counts across all 31 English configurations (150 issues per task; 2026 corpus). Panels (a) and (b) show responsibility and solution. Labels identify the model tag and recorded thinking setting; unset denotes no explicit setting. Dots show counts under each common order, separately for the two blocks. Bars show exact central 95 percent ranges under independent per-issue assignment; ticks mark order averages. Tied maxima share credit

Coverage above 95 percent under randomization is expected for this fixed corpus: variation across issues contains both measurement noise and differences in the issues' true profiles, although the latter are not error for their finite average. For independently sampled issues, randomized intervals are approximately nominal under a central limit approximation.

Fig. 4 shows the diagnostic signature. Adding issues makes fixed-order coverage worse: the standard error shrinks while the shared shift persists. Randomized designs remain above nominal coverage throughout. A fixed balanced set performs better than one fixed order but still under-covers solution averages at the full corpus size.

4.3 Comparisons between policy domains

Comparisons within one instrument are less exposed than levels because the common part of the shift cancels. Across the 55 pairs of EU policy domains, fixed-order Welch intervals cover in a median of 97.3 percent of responsibility comparisons and 90.8 percent of solution comparisons. They are still vulnerable when the order effect differs between domains. Regression coefficients behave similarly: the intercept absorbs a constant shift, but associations with item characteristics remain exposed to shifts that vary with those characteristics (Supplement A2.2 and A3.5). Comparisons across models, prompts or studies measured under different orders lack this cancellation.

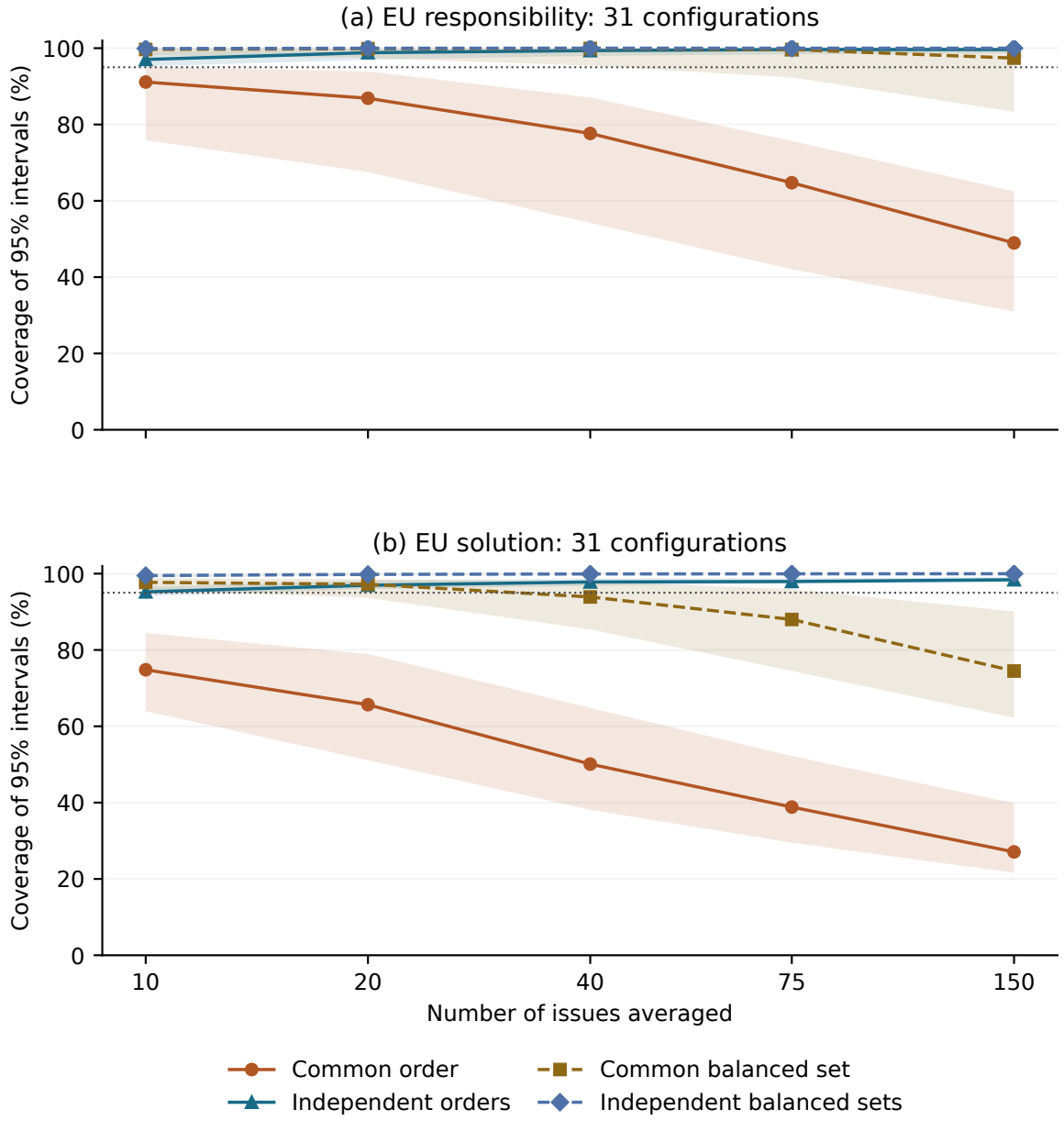


Fig. 4 Coverage of conventional 95 percent intervals for the average score profile of randomly drawn subsets of the 150 EU issues in the 2026 corpus, by design. Panels (a) and (b) show responsibility and solution judgments. Lines are medians and shading interquartile ranges across 31 configurations, using 50 random subsets for each size below 150. Fixed designs average over all orders or balanced sets, random designs over random assignments. Targets are order-balanced means of the other replicate block. The dotted line marks 95 percent

5 What survives changes to the instrument?

5.1 Languages and task framing

Responsibility and solution questions have different targets and different fixed-order coverage, but the same design remedy works in both. Table 4 reports the parallel comparison for each language. Every language–task median shows low fixed-order coverage and high coverage under independent assignment. Across all 246 replicated cells, independent assignment reduces estimated MSE and its lowest observed coverage is 95.4 percent. These comparisons hold each language's own target fixed; they do not show that translations elicit identical judgments.

Table 4 Coverage across the four language versions of the EU instrument (2026, %)

Question	Language	Fixed order	Random order	Random set
Responsibility	English	49.0	99.6 (0.04)	100.0 (0.00)
	German	49.0	99.5 (0.05)	100.0 (0.00)
	French	45.3	99.3 (0.09)	100.0 (0.00)
	Spanish	45.6	99.4 (0.06)	100.0 (0.00)
Solution	English	27.1	98.5 (0.08)	100.0 (0.01)
	German	28.1	97.8 (0.08)	99.9 (0.02)
	French	29.2	97.8 (0.11)	99.9 (0.02)
	Spanish	26.6	97.7 (0.08)	100.0 (0.01)

Note: Medians over 31 configurations per task and language (30 in Spanish). Fixed-order coverage enumerates all orders. Parentheses are assignment MC SEs in percentage points from 5,000 conditional resamples. A random set uses four calls; the other designs use one. Recorded blocks and configurations are fixed. Supplement A3.8 gives the full uncertainty procedure.

5.2 Instructions, labels and temperature

The same qualitative result survives the prompt treatments (Table 5). It does not depend on requiring a score spread, suppressing hedging, using the original labels or maintaining the low base temperature. Combining translated issues with English instructions also retains the gap between fixed and independent assignment. Larger models show the same

pattern. These single-block arms support the assignment comparison, not a separate decomposition of order effects and generation noise.

Table 5 Coverage (%) of conventional 95 percent intervals in EU instrument variants (2026)

	One fixed order		Random order per issue	
	Resp.	Sol.	Resp.	Sol.
A. Instrument variants (13 configurations, one block each)				
Base prompt	44.8	26.0	99.6 (0.11)	98.3 (0.20)
No forced score spread	38.5	20.8	99.6 (0.12)	98.8 (0.16)
Also no instruction against hedging	39.6	26.0	99.5 (0.12)	98.8 (0.16)
Paraphrased option labels	41.7	22.9	99.8 (0.09)	97.8 (0.23)
Temperature 0.5	37.5	29.2	99.5 (0.14)	97.9 (0.25)
Temperature 0.8	45.8	32.3	99.6 (0.12)	97.8 (0.23)
French and German prompts	42.7	27.6	99.0 (0.11)	97.9 (0.14)
French and German issues, English instructions	34.4	27.1	99.1 (0.14)	98.1 (0.13)
B. Larger models, 26–35B (4 configurations, one block)				
Base prompt	59.9	38.5	99.9 (0.04)	99.5 (0.12)

Note: Medians over configurations of per-option coverage for averages over 150 issues. Each arm has one block of 24 orders per issue, and its target is the arm's own order-balanced mean. French and German rows pool both languages. Random-order values average 400 assignments; parentheses are conditional MC SEs from 5,000 resamples of their recorded outcomes.

5.3 Reasoning changes the allocation of error

Reasoning is a different kind of treatment: it changes both the order effect and generation noise. Fig. 5 separates those changes rather than treating lower order sensitivity as sufficient evidence of greater precision. The strict subset compares the two Gemma4 tags and `nemotron-3-nano:4b` in every language. The estimated shared order variance is lower with thinking in 23 of 24 comparisons but the random-balanced-set risk estimate is higher in all 24. Once position balance removes much of the order effect, additional generation noise can dominate.

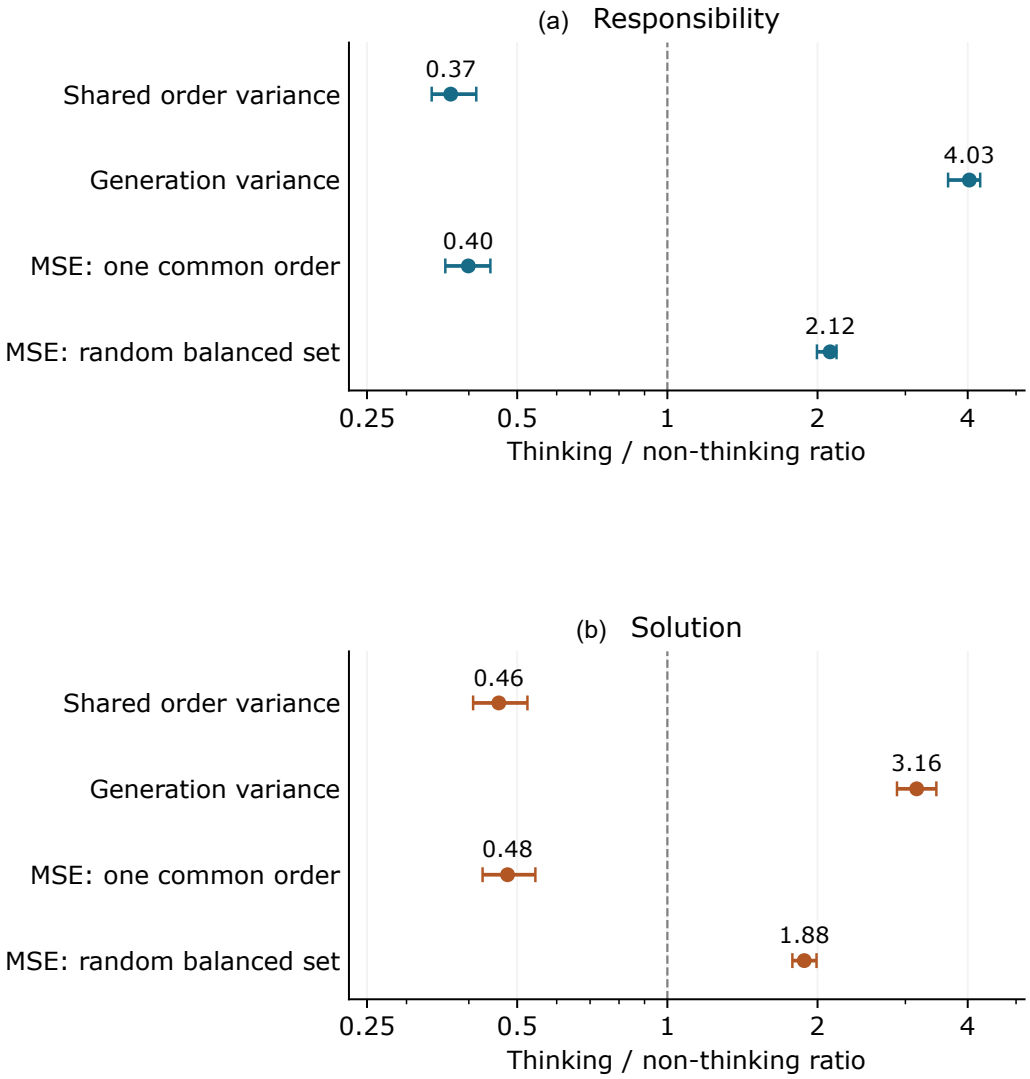


Fig. 5 Reasoning comparisons for (a) responsibility and (b) solution judgments with equal token caps and strict prompt-path restrictions (12 pairs per task across four languages; 2026 EU corpus). The included tags are `gemma4:e2b`, `gemma4:e4b` and `nemotron-3-nano:4b`. Points are medians; whiskers are central 95 percent ranges from 3,000 paired whole-issue resamples, conditional on fixed models and recorded blocks. The one-order and balanced-set designs use one and four calls per issue. Supplement A3.9 gives the exclusions and broader pairs

The strict subset contains three tags from two families, rather than a representative model sample. Across the broader 72 recorded comparisons, balanced-set risk rises in 48; the result is therefore not universal. The full table and exclusion rules are in Supplement A3.9. Matched caps do not imply equal token use, and each reasoning mode is evaluated against its own target.

6 When fitted repairs fail to travel

6.1 Repairing existing fixed-order measurements

For existing outputs, changing the original assignment is no longer possible. A pilot of real issues can estimate the average order profile for subtraction, a score-level analogue of the probability-based debiasing studied by Zheng et al. (2024). Table 6 shows that this lowers RMS error, but treating the estimated correction as known still under-covers. Widening intervals for pilot uncertainty is essential. The issue-free baseline adapts the content-free-input idea of Zhao et al. (2021) to score-profile subtraction. It lowers median coverage here, showing the limits of this adaptation for a content-dependent order profile.

6.2 Correction transfer across languages, domains and tasks

Fig. 6 compares an English pilot with an otherwise identical pilot in the destination language, using the same 30 issue identities and the opposite block for evaluation. The English correction has higher held-out MSE in 182 of 184 destination cells; it even increases error relative to no correction in 80. The local advantage remains in 123 of 124 comparisons after conservative cap and prompt-path restrictions.

This is the distinction between a portable design rule and a fitted correction. Independent assignment requires no transported profile or pilot. Even local correction has higher held-out MSE than randomization in all 184 destination cells before charging its pilot calls. The evidence concerns these language changes, not every type of transfer. A 30-issue correction fitted outside each destination domain has a median corrected/uncorrected MSE ratio of 0.51 but raises error in 311 of 2,706 cells. Size-matched random pseudo-domains

Table 6 Correcting EU issue averages measured under one fixed order (2026, exploratory)

	Coverage (%)		RMS error (points)	
Correction of fixed-order averages	Resp.	Sol.	Resp.	Sol.
A. Pilot issues under all 24 orders (31 configurations)				
None, 140 held-out issues	50.5	29.3	3.9	4.4
Subtract profile, 10 pilot issues (240 calls)	60.6	50.3	3.3	2.9
and widen interval for pilot error	94.9	93.4	3.3	2.9
None, 120 held-out issues	55.2	31.3	3.9	4.5
Subtract profile, 30 pilot issues (720 calls)	85.2	75.6	2.0	1.8
and widen interval for pilot error	97.5	96.0	2.0	1.8
B. Content-free prompt under all 24 orders (13 configurations)				
None	47.9	25.0	–	–
Subtract content-free profile	30.7	23.4	–	–

Note: Medians over configurations. Panel A uses 50 random splits: 10/30 pilot issues in one block and 140/120 held-out issues in the other. Each pilot size has its own uncorrected baseline; widening adds the estimated profile's sampling variance. Panel B subtracts the profile fitted to a placeholder prompt.

also yield 205–263 adverse cells (median ratio 0.40). Real domains are less favorable than all 100 groupings, indicating reproducible heterogeneity associated with the observed grouping (Supplement A7.8).

Responsibility and solution change both the question and option menu, so their substantive score profiles cannot simply be equated. We instead transfer a generic four-position correction, fitted on 30 issues, between tasks and compare it with a destination-task fit on the same pilot. On 120 held-out issues, transport has higher MSE in 244 of 246 destination cells (median ratio 1.51), and 46 of 123 configuration–language pairs reverse first-versus-last preference in both blocks. These saved-data diagnostics support validating correction for the intended domain and task, rather than assuming a universal model-specific position bias (Online Resource 4).

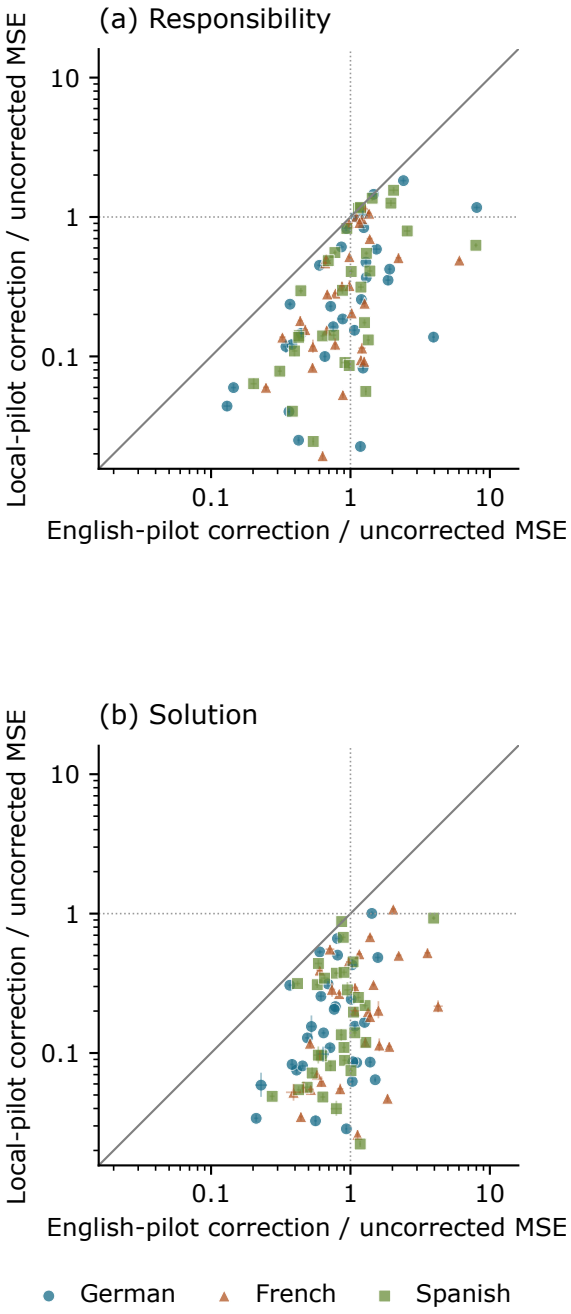


Fig. 6 English-to-local-language correction transfer (2026 EU corpus). Panels (a) and (b) show responsibility and solution. Each point is one of 184 destination cells; below the diagonal, local correction has lower held-out MSE. Both pilots use the same 30 issue identities, evaluating 120 other issues. Losses average 50 splits and both block directions. Whiskers are pointwise 95 percent MC ranges from 5,000 paired split resamples, conditional on fixed models and blocks. Axes are logarithmic; marker shapes identify languages. Supplement A3.8 details the procedure

6.3 Does a calibrated prospective design repay its pilot?

A prospective pilot can do more than subtract an order profile: it can select probabilities over balanced sets while keeping every order's marginal inclusion probability uniform. We developed such target-preserving designs before the shared-error analysis, capped their departure from uniform sampling and froze six estimators before evaluating fresh panels. Supplement A5–A6 gives the construction, reference-free risk contrasts and full evaluation.

Table 7 reports the English and German EU results for the four prespecified calibration configurations: `gemma4:e4b` and `qwen3:8b` with thinking off, and `granite4:3b` and `mistral:7b` with no explicit thinking setting. These are fresh score-only responses to previously studied issues, rather than unseen political content. Combined correction and calibration adds modest gains over simple balance or a shrunken pilot correction (Owen 2013). In English these gains are 5.6 and 2.5 percent, so the prespecified criterion of at least five percent against both comparators was not met, and they depend strongly on one model (Supplement A6.5).

Table 7 Prespecified EU evaluation (2026): gains from combined correction and calibration

Instrument	Comparator	Δ (SE)	MSE reduction (%)
English scores	Uniform balance	-0.91 (0.10)	5.6 [3.2, 7.8]
	Shrunk control	-0.40 (0.09)	2.5 [0.5, 4.5]
German scores	Uniform balance	-1.47 (0.13)	6.5 [4.6, 8.3]
	Shrunk control	-1.03 (0.12)	4.6 [3.1, 6.2]

Note: Equal-weight averages over the four named calibration configurations, 60 solution-task issues per language with fresh responses and two blocks. Δ is the reference-free risk difference in squared score points, averaged over coordinates (generation SE); negative values favor calibration. Positive reference-MSE reductions favor calibration; brackets are pointwise 95 percent paired issue-resampling sensitivities. Comparators use four calls per issue and omit pilot cost. Every full-cost comparison favors spending those calls on more orders (Supplement A6).

The full-budget comparison charges the pilot: spending the same total calls on more orders gives lower reference MSE for every model in both EU panels. The separate

Bills panel reaches the same full-cost conclusion. Thus the prespecified evaluation supports the simpler design choice established by the crossed EU evidence, rather than a recommendation to fit elaborate sampling policies.

7 Reporting and design recommendations

The practical decision depends on whether measurements are still to be collected, already fixed, or being moved to another instrument (Table 8). Independent assignment is the supported default for the evaluated instruments. Legacy correction and instrument diagnosis require additional pilot calls, whose costs and uncertainty must be retained. The replication module supplies order/set assignment, answer remapping, component diagnostics and widened pilot correction.

8 Discussion and conclusions

The results distinguish repeatability under a fixed prompt from precision under incidental presentation changes. A repeated output can be stable within an order while its systematic order effect is shared across many items. Averaging those items then produces an apparently precise aggregate around a displaced value. In survey terms, a fixed option order behaves like a single interviewer who conducts every interview; drawing the order independently for each item is the analogue of rotating response options across respondents. The standard variance identities explain the mechanism; the crossed corpus establishes its empirical importance for this instrument and tests the decisions it suggests.

Independent item-level assignment reduces estimated MSE in every replicated cell evaluated here. Its advantage is operational as well as statistical: it requires neither an estimated correction profile nor a pilot whose costs must be recovered. This is an application of randomization to a clearly defined nuisance set, not a universal dominance result. When item-order effects are sufficiently negatively associated, a common order can offset errors across items; Section 3 gives the condition. The empirical question is therefore how the order effects align within the intended set of measurements.

Table 8 A workflow for LLM measurement based on the evaluated designs

Research situation	Supported procedure	Cost and interpretation
New data: one call per item	Draw the option order independently and uniformly for each item; map the answer back to its substantive option and retain the assignment	Same calls as a common order. Estimates the declared order-average target; does not validate that target
New data: four calls per item	Draw a position-balanced set independently for each item. For larger budgets, use further independent balanced sets	Same calls as reusing a four-order set. Cancels additive position effects; assess the remaining nonadditive and generation error
Legacy fixed-order data	Fit a correction using separate real pilot items under every order; account for pilot uncertainty and check held-out performance	The evaluated 30-item pilot costs 720 calls. This is an evaluated size, not a universal requirement or a free repair; content-free correction failed here
Changed instrument	Reassess order profiles and design-specific risk when policy domain, task, model, language, instructions or reasoning changes. Two complete blocks identify the components under stationary, independent generation	A 30-item, two-block diagnostic costs 1,440 calls. Validate any transported correction; equal token caps do not imply equal cost
External construct	Validate against suitable external evidence, separately from repeatability and assignment precision	Narrower intervals or order stability do not establish agreement with people, reference labels or political reality

Note: Report the target, admissible orders, calls per item and assignment mechanism. Conventional randomized intervals require independent calls and an adequate central limit approximation; across-item variance is conservative for a fixed item set. Compare learned designs at the full budget: modest average deployment gains in the prespecified EU panels did not repay the pilot.

The instrument changes show which decisions are portable and which quantities need local evidence. The parallel language, task and prompt comparisons support independent assignment across the recorded conditions. Fitted-repair portability depends on the change: out-of-domain correction often helps, while language transfer and cross-task position templates usually lose against local fits. Thinking changes both generation and order variation, so selecting a mode solely for its lower order sensitivity can increase the risk remaining after balance. These comparisons concern recorded configurations with their own targets and operational response processes; they do not isolate a neural mechanism or show that one mode produces better substantive judgments.

The contribution complements rather than replaces existing reliability and prompt-uncertainty work. Repeated-generation consistency (Hsu and Hsu 2026), prompt optimization against external labels (Abraham et al. 2025), audits of design sensitivity (Baumann et al. 2025; Reiter et al. 2026) and uncertainty across governed pipeline choices (Messing 2026; Leng 2026) address related but different decisions. Our evidence adds a detailed assessment of shared order error, equal-call assignment and the portability of fitted repairs. The practical implication is to match a reliability diagnostic to the way measurements will be aggregated, rather than relying on a single consistency measure.

A categorical extension clarifies the boundary with validity. Four models classify 200 Congressional Bills descriptions into policy topics (Supplement A3.2). Their shared order errors differ, and independent assignment gives high coverage for the model-specific order-average targets. Yet order-averaged shares still differ from the reference codes. The stance extension likewise separates design-based precision from agreement with existing human annotations. These checks broaden the kinds of outputs studied, while retaining the distinction between a well-measured model target and a valid measure of an external construct (Bisbee et al. 2024; Halterman and Keith 2026). For applications aimed at human-coded constructs, a probability sample of high-quality human annotations can supply a separate inferential bridge: Egami et al. (2023) combine such labels with imperfect surrogate annotations in design-based downstream estimation. The assignment rule and the human-validation design then address distinct components of measurement error.

Using open-weight systems of about 2–35 billion parameters enabled broad comparisons across model families, reasoning settings and decoding configurations. Local control also permits retaining weights and runtimes that proprietary APIs do not uniformly expose, supporting reproducible model specifications after hosted versions change or retire. Stochastic decoding and hardware differences can still change individual outputs.

Several limitations qualify the scope. The shared-order diagnostics and fitted repairs were developed after inspecting the corpus; their uncertainty does not account for that analytical development. The calibration protocol was frozen before its fresh responses were collected. Models, languages and issues are not probability samples, and the treatment matrix is only partially crossed. Fixed-order coverage is higher for the four largest (Table 5) but still far below nominal, and magnitudes for larger proprietary models are untested. The single-block arms use their own recorded balanced means as references and cannot identify separate order and generation components or establish coverage for new generations. Two blocks in the replicated arms cannot rule out drift or dependence between calls. Translations can change meaning as well as response processes, and the forced score spread restricts the EU instrument even though removing it retains the qualitative assignment result.

For applied social research, the evidence supports a simple workflow. Define the target and the admissible presentation changes, assign those changes independently across items, retain the assignment with each output and assess the uncertainty relevant to the planned aggregate. Reassess fitted corrections after changing an instrument, and charge them for the information used to estimate them. More repeated calls and more elaborate prompting can be useful, but their value depends on the error left by the design and the external validity required of the measurement.

Data availability

Supplementary methods, code, saved model responses, derived results and cached vectors are archived together as Online Resources 1–4 at Zenodo (https://doi.org/10.5281/

zenodo.24155464). The Congressional Bills sample and included source-audit inputs retain the Comparative Agendas Project terms (Wilkerson et al. 2025; Jones et al. 2025). SemEval source texts and individual source annotations are excluded under the source-owner redistribution terms; the source is documented by Mohammad et al. (2016) and Barbieri et al. (2020). They can be obtained for research from https://saifmohammad.com/WebPages/StanceDataset.htm under the stated terms. Online Resource 2 identifies the pinned TweetEval source files, original hashes and exact local restoration procedure in WHERE_TO_GET_EXTERNAL_DATA.md; complete stance reanalysis requires those separately obtained files.

Supplementary information (Online Resources 1–4)

Online Resource 1 (ESM_1.pdf) documents instruments, derivations, detailed results, uncertainty procedures, evaluations and reproducibility qualifications. Online Resource 2 (ESM_2.zip) contains the anonymous version 1.7 analysis code, saved responses, derived results and source-restoration instructions. Online Resource 3 (ESM_3.zip) contains the cached embedding vectors and indexes required for the semantic reanalysis; no embedding model is needed.

Online Resource 4 (ESM_4.zip) supplies the analysis specification, reproducible scripts, complete domain/task results, the pseudo-domain benchmark and independent checks for Supplement A7.8; it uses the EU inputs in Online Resource 2 and needs no model calls.

AI-use statement

The authors used Anthropic Claude Opus 5.5 and OpenAI ChatGPT-6.1 Sol for literature search, text editing, and code completion and correction. The large language models described in the manuscript were used to generate the analyzed data. The authors checked and verified all claims and take full responsibility for the content of this article.

References

Abraham L, Arnal C, Marie A (2025) Prompt selection matters: enhancing text annotations for social sciences with large language models. Journal of Computational Social Science 8:73. https://doi.org/10.1007/s42001-025-00388-6

Argyle LP, Busby EC, Fulda N, et al (2023) Out of one, many: Using language models to simulate human samples. Political Analysis 31(3):337–351. https://doi.org/10.1017/pan.2023.2

Barbieri F, Camacho-Collados J, Espinosa Anke L, et al (2020) TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, pp 1644–1650, https://doi.org/10.18653/v1/2020.findings-emnlp.148

Barrie C, Palaiologou E, Törnberg P (2024) Prompt stability scoring for text annotation with large language models. arXiv preprint arXiv:2407.02039v3, https://doi.org/10.48550/arXiv.2407.02039, first posted 2024; version 3, May 15, 2026

Baumann J, Röttger P, Urman A, et al (2025) Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. arXiv preprint arXiv:2509.08825, https://doi.org/10.48550/arXiv.2509.08825

Bisbee J, Clinton JD, Dorff C, et al (2024) Synthetic replacements for human survey data? The perils of large language models. Political Analysis 32(4):401–416. https://doi.org/10.1017/pan.2024.5

Brennan RL (2001) Generalizability Theory. Springer, New York, https://doi.org/10.1007/978-1-4757-3456-0

Brucks M, Toubia O (2025) Prompt architecture induces methodological artifacts in large language models. PLOS ONE 20(4):e0319159. https://doi.org/10.1371/journal.pone.0319159

Debevc L, Chatterjee N, Doucet A, et al (2026) Navigating the digital spectrum: Assessing political bias, stability, and downstream fairness in large language models. arXiv preprint

arXiv:2609.08637, https://doi.org/10.48550/arXiv.2609.08637

Dominguez-Olmedo R, Hardt M, Mendler-Dünner C (2024) Questioning the survey responses of large language models. In: Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 45850–45878, https://doi.org/10.52202/079017-1458

Egami N, Hinck M, Stewart BM, et al (2023) Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In: Advances in Neural Information Processing Systems, vol 36. Curran Associates, Inc., pp 68589–68601, https://doi.org/10.52202/075280-3000

Feng P, Meng C, Qu D, et al (2026) Second-order response laws for LLM judges: Debiased estimation of prompt instability. arXiv preprint arXiv:2608.16253, https://doi.org/10.48550/arXiv.2608.16253

Galesic M, Tourangeau R, Couper MP, et al (2008) Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. Public Opinion Quarterly 72(5):892–913. https://doi.org/10.1093/poq/nfn059

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences 120(30):e2305016120. https://doi.org/10.1073/pnas.2305016120

Groves RM, Lyberg L (2010) Total survey error: Past, present, and future. Public Opinion Quarterly 74(5):849–879. https://doi.org/10.1093/poq/nfq065

Halterman A, Keith KA (2026) Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. Political Analysis 34(2):188–204. https://doi.org/10.1017/pan.2025.10017

Helm H, Priebe CE (2026) When prompt perturbations break your A/B test: A valid statistical test for generative surveying. arXiv preprint arXiv:2605.27463v1, https://doi.org/10.48550/arXiv.2605.27463, version 1, 26 May 2026

Hsu HY, Hsu ML (2026) Semantic stability protocol: intercoder reliability for zero-shot classification via AI coders. Quality & Quantity 60(5):16545–16567. https://doi.org/10.1007/s11135-026-02832-9

Jones BD, Baumgartner FR, Theriault SM, et al (2025) Policy agendas project: Codebook. Comparative Agendas Project, URL https://www.comparativeagendas.net/pages/master-codebook, official master-codebook page, accessed October 25, 2026

Kish L (1962) Studies of interviewer variance for attitudinal variables. Journal of the American Statistical Association 57(297):92–115. https://doi.org/10.1080/01621459.1962.10482153

Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5(3):213–236. https://doi.org/10.1002/acp.2350050305

Krosnick JA, Alwin DF (1987) An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly 51(2):201–219. https://doi.org/10.1086/269029

Le Mens G, Gallego A (2025) Positioning political texts with large language models by asking and averaging. Political Analysis 33(3):274–282. https://doi.org/10.1017/pan.2024.29

Leng Y (2026) When prompts matter: Quantifying prompt uncertainty and improving empirical inference with LLM-based measurement. Tech. Rep. 6381558, SSRN, https://doi.org/10.2139/ssrn.6381558, working paper, 9 March 2026

Meng XL (2018) Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics 12(2):685–726. https://doi.org/10.1214/18-AOAS1161SF

Messing S (2026) Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking. Accepted at NeurIPS 2026, https://doi.org/10.48550/arXiv.2604.11581, consulted preprint: arXiv:2604.11581v6, 13 May 2026; conference presentation forthcoming

Mohammad S, Kiritchenko S, Sobhani P, et al (2016) SemEval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, pp 31–41, https://doi.org/10.18653/v1/S16-1003

Owen AB (2013) Monte Carlo theory, methods and examples. Author's online monograph, URL https://artowen.su.domains/mc/, variance reduction and importance sampling chapters; updated versions accessed October 25, 2026

Parschan P, Jakob C (2026) Computational measurement of political positions: a review of text-based ideal point estimation algorithms. Quality & Quantity 60(2):5663–5690. https://doi.org/10.1007/s11135-025-02500-4

Pezeshkpour P, Hruschka E (2024) Large language models sensitivity to the order of options in multiple-choice questions. In: Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, pp 2006–2017, https://doi.org/10.18653/v1/2024.findings-naacl.130

Podsakoff PM, MacKenzie SB, Lee JY, et al (2003) Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology 88(5):879–903. https://doi.org/10.1037/0021-9010.88.5.879

Reiter T, Kern C, Miasnikov F, et al (2026) Reliable but design-sensitive: Instrument uncertainty in LLM annotation. In: The Proceedings of the Third Workshop on Uncertainty-Aware NLP. Association for Computational Linguistics, pp 312–327, https://doi.org/10.18653/2026.uncertainlp-main.20

Röttger P, Hofmann V, Pyatkin V, et al (2024) Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 15295–15311, https://doi.org/10.18653/v1/2024.acl-long.816

Rupprecht J, Ahnert G, Strohmaier M (2026) Prompt perturbations reveal human-like biases in large language model survey responses. In: Proceedings of the Seventh Workshop

on Natural Language Processing and Computational Social Science. Association for Computational Linguistics, pp 1–21, https://doi.org/10.18653/v1/2026.nlpcss-1.1

Sclar M, Choi Y, Tsvetkov Y, et al (2024) Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In: The Twelfth International Conference on Learning Representations, URL https://openreview.net/forum?id=RIu5lyNXjT

Searle SR, Casella G, McCulloch CE (1992) Variance Components. Wiley, New York, https://doi.org/10.1002/9780470316856

Tjuatja L, Chen V, Wu T, et al (2024) Do LLMs exhibit human-like response biases? A case study in survey design. Transactions of the Association for Computational Linguistics 12:1011–1026. https://doi.org/10.1162/tacl_a_00685

Törnberg P (2024) Best practices for text annotation with large language models. Sociologica 18(2):67–85. https://doi.org/10.6092/issn.1971-8853/19461

Wang P, Li L, Chen L, et al (2024) Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 9440–9450, https://doi.org/10.18653/v1/2024.acl-long.511

Wilkerson J, Adler ES, Jones BD, et al (2025) Policy agendas project: Congressional bills. Comparative Agendas Project, URL https://www.comparativeagendas.net/project/us/datasets, downloaded October 25, 2026; snapshot and source hash retained in replication materials

Zhao Z, Wallace E, Feng S, et al (2021) Calibrate before use: Improving few-shot performance of language models. In: Proceedings of the 38th International Conference on Machine Learning, pp 12697–12706, URL https://proceedings.mlr.press/v139/zhao21c.html

Zheng C, Zhou H, Meng F, et al (2024) Large language models are not robust multiple choice selectors. In: The Twelfth International Conference on Learning Representations, URL https://openreview.net/forum?id=shr9PXz7T0