

The Model as One Rater Among Several: Measuring Political Positions in Data-Sparse Regions with a Language-Model Panel

Tarek Gara*

2026

Abstract

Most tools for measuring political positions, manifesto coding, expert surveys, text-scaling models and the like, were built and validated on Western party systems. Outside that setting they work poorly, and often not at all. The regions where they break down worst tend to be the ones where many of the actors who matter never publish a Western-style manifesto, never contest a competitive election, or say one thing in a charter and do another in office. This paper is an attempt at a method for those settings. It treats a large language model not as a measurement device but as a single, fallible *rater* in a panel, roughly the way an expert survey treats one expert: the value comes from pooling many judges rather than trusting any one of them. I describe the panel and how its scores get combined, an applicability rule that keeps a score of zero distinct from a blank, and a lens system that separates what an actor says from what it does. I report three results that try to test whether the method works. First, a built-in experiment isolates the effect of the rubric. Holding a definition-free first round fixed, adding written axis definitions moves scores by a mean of 1.8 points on a 21-point scale, and it tightens agreement between raters on the same cells (mean absolute gap 2.81 \rightarrow 2.50; r 0.81 \rightarrow 0.89). So the definitions are doing more than nudging scores in a chosen direction. They make two independent raters agree more closely, which an arbitrary steer would not. Second, across nine models from eight laboratories in two countries, agreement is high, and it holds up whether the scale is treated as interval or ordinal: Krippendorff's α is 0.86 on both an interval and an ordinal metric, and it stayed put as the panel grew from five raters to nine. That is *reliability*, the reproducibility of a reading, and not *validity*, its correctness. Models trained on overlapping data can share a bias, and no statistic computed from the models alone can tell a reliable instrument apart from a reliably wrong one. Third, where the panel does disagree, the disagreement turns out to be informative. The sharpest single split, a full-scale, bimodal divergence on an actor's stance toward its state's foundational order, points to a *referent problem*. For a party caught in a moment of constitutional rupture, the models agree on the facts but differ on *which* order counts as "the regime," the one the actor toppled or the one it now defends, and so they split across the whole scale. A blind triple-coding of the widest such cells assigns about two-thirds of the disagreement to interpretation rather than to fact or unclear definitions. No sharper definition can take it away, because choosing the referent is itself a political judgement. I try to be plain about what the method can't do, including the human validation it still lacks, and I release the instrument and data in full. So what the paper offers is a method and a measurement scaffolding, with its reliability established and its validity an open question the design is meant to chip away at, rather than a finished reading of the region. An early grounded pilot makes the gap concrete: re-scoring the same cells from cited documents moves a single model's placements more than switching models does. The worked example is the Middle East and North Africa, but I'd expect the method to carry to any region these standard tools leave out.

*Independent researcher. Correspondence via tarekgara.com/tayyar. The author is responsible for all judgements, scores, and errors. The instrument, the rubric, and versioned data snapshots are public at the address above.

I Introduction

Ask where a Lebanese party sits relative to an Egyptian one, or whether Tunisia's Ennahda in 2024 stands where Egypt's Muslim Brotherhood stood in 2012, and you find that the tools built for this kind of comparison mostly stop at the edge of Europe. The spatial tradition that treats politics as movement along defined dimensions is old and well established [2, 1, 3], and the manifesto-coding and expert-survey programmes that put it to work are among the most-cited tools in the field [4, 13]. They were not built for, and do not transfer cleanly to, a region where many of the most important actors never publish a Western-style manifesto, never contest a competitive election, or commit one position to a founding charter and the opposite in office. The result is not a few missing rows; it is a tool that was never designed to take the reading.

This gap matters beyond area studies. Quantitative comparative politics increasingly uses position estimates as variables in models of coalition formation, party competition, conflict onset, and democratic backsliding, so the narrow geographic reach of these tools quietly limits what the field can say. A field that can place every party in the Bundestag to a tenth of a point but cannot place the actors of a twenty-state region at all is not neutral about which questions get asked.

The first thing most people would try now is to point a large language model at the problem and read the answer off. A fast-growing literature reports that frontier models match or beat crowd workers, and sometimes trained coders, on text-annotation tasks [21, 22], and that makes it tempting to treat the model as a measurement device that has simply read more than any human could. I think that's the wrong stance to take. A model's fluent, confident placement of a party isn't a measurement. It's one reading, produced by one system, carrying biases from a training corpus the analyst did not choose and cannot inspect, and if you treat it as an oracle those biases get laundered into apparent fact. Treat it instead as a *rater*, a judge whose output is pooled with others and whose disagreements are kept rather than hidden, and it becomes something an existing methodology already knows how to handle. For forty years expert surveys have built valid measures out of fallible human judges by pooling them and studying their agreement [13, 14]. What this paper does is put the language model in the same role, with the same tools and the same scepticism, and then ask the questions that methodology is designed to answer. Do the raters agree, and if they do, is it because they are right or because they share a mistake?

The instrument I build to make this concrete, Tayyar, is a positional dataset of the Middle East and North Africa: parties and public figures, current and historical, placed on a set of clearly defined, regionally anchored axes by a panel of raters that is, for now, mostly language models. I lean on it throughout as a worked example, though the contribution isn't really the dataset, and it isn't a new estimator either. It's a method and a stance. The method uses language models as raters in data-sparse measurement, taking their reliability seriously without mistaking it for validity. The design choices fall out of that stance: applicability rules, a *declared-versus-behavioural* lens, a synthesis that keeps disagreement, and an inclusion gate on aggregation.

A few things come out of this. First, I set out and defend the model-as-rater stance and the panel design behind it, including the measurement reasons to keep a blank distinct from a zero and a stated position distinct from a revealed one. Second, there's a *built-in experiment* that isolates the effect of the rubric by holding a definition-free round fixed: written definitions both move scores and improve agreement between raters, and that measurable "rubric effect" rules out the simplest version of the worry, that the rubric just imposes an arbitrary direction on a compliant model. Third, I try to treat reliability carefully, reporting Krippendorff's α under both interval and ordinal metrics so the result doesn't hang on an arguable equal-spacing assumption, and pressing the point that high agreement among models trained on overlapping data is not validity. The *pattern* of disagreement turns out to carry information too: the panel's widest split, on an actor's stance toward its state's foundational order, traces not to error but to a referent ambiguity that no rubric can resolve. And throughout I try to be explicit about what the instrument cannot yet do, so a reader can weigh it fairly.

If the stance holds up, what you get is a measurement where the field currently takes none: the actors of a data-sparse region placed on comparable, defined axes, with their disagreement reported rather than smoothed away, using raters that are available at scale today. The whole thing is built around its own central risk, which is that capable

models can be capably wrong in the same direction, and so it tries to measure that risk and keep it in view, leaving a place for the human validation that would actually close it.

I should say up front that the reliability figures below can't stand in for an external check. The usual external benchmarks (a manifesto corpus, an expert survey like the Chapel Hill placements, party-level democracy scores) barely reach these actors, which is the gap the method exists to address in the first place. So the validity test has to come from purpose-built human anchoring rather than from some existing dataset to benchmark against.

2 Related work

Tayyar sits where a few literatures meet, and what it really tries to do is borrow the discipline of each one without inheriting the blind spot that tends to come with it. I'll take the traditions one at a time. Most of the method's design choices make the most sense if you read them as responses to a specific limitation in each.

Spatial models of politics. The idea that political actors can be located as points in a low-dimensional issue space runs from Hotelling and Downs [1, 2] through the roll-call scaling of Poole and Rosenthal [3], and it's still the common language of quantitative comparison. Tayyar keeps the geometry but drops an assumption that makes good sense in two-party legislatures and much less sense here: that a single recovered dimension carries most of the signal. In the Middle East and North Africa a party's stance on the role of religion in the state, its position on the regional axis of resistance versus normalisation, and its posture toward the Palestinian question are empirically distinct, and you can't really fold them back into a left-right line. Collapse them too early and you throw away the very structure you were trying to recover. So the instrument hangs onto a full position vector and treats the familiar two-axis "compass" as a default view rather than the measurement itself.

Manifesto coding and text scaling. The Comparative Manifesto Project and its MARPOR successor [4, 6, 5] established the practice of estimating positions from a party's own published texts, hand-coded sentence by sentence against a fixed scheme. Its automated descendants ("words as data" [7], the Wordfish scaling model [8], and the scaling-from-coded-text refinements that came after [9]) dropped the human coder, but at the price of needing a large reference corpus and a language the method has actually seen at scale. Two things matter here. First, the coding is noisier than how widely it's used would suggest. Mikhaylov, Laver, and Benoit show non-trivial misclassification even among trained human coders of manifestos [10], which is a sobering baseline for any automated stand-in, and a reminder that "the human coder" isn't really a gold standard, just another rater. Second, and this is the one that decides it for our setting, bag-of-words scaling is tied to its corpus and its language. Arabic is comparatively under-served by the pretrained representations these methods lean on [36], and Hebrew more so. A Wordfish estimate comparing an Arabic charter to a Hebrew platform is, in effect, comparing two thinly modelled languages through a method that quietly assumes a dense one. Grimmer and Stewart are honest that automated text methods amplify careful reading rather than stand in for it [11, 12]. Tayyar keeps document grounding as the spine of its *declared* reading, but swaps word-frequency scaling for rubric-anchored judgement, which a multilingual model applies more evenly across Arabic, Hebrew, and English than a model of one language's word counts ever could.

Expert surveys and the aggregation of judgement. The Chapel Hill Expert Survey [13, 15] and Benoit and Laver's expert placements [14] ask informed humans to place parties directly and take the mean expert as the estimate. Their validity rests on inter-expert agreement and on the hope that idiosyncratic error washes out across a panel. The Varieties of Democracy project goes furthest with this, aggregating many country experts through an explicit item-response measurement model that estimates both latent positions and the raters' reliability and threshold differences [16, 17]. This is the tradition Tayyar most closely belongs to, and the analogy is pretty close to literal: the panel is an expert survey in which most of the "experts" happen to be language models. It inherits the tradition's main strength, a principled way to build a measure out of fallible judges. It also inherits the vulnerability V-Dem's own

methodologists are careful about. A panel that agrees for a shared wrong reason looks, from inside the agreement statistics, exactly like one that agrees because it's right. The catch is that human expert panels are deliberately assembled to be diverse, whereas a panel of language models can be a lot more correlated than its size lets on, which is something I come back to at length.

Cross-cultural incomparability. Once you put actors from very different systems on a single scale, you run into the problem that the scale may not mean the same thing in each. King and colleagues formalised this as differential item functioning and proposed anchoring vignettes to correct it [18]. The Aldrich–McKelvey family of methods, and its Bayesian extensions, recover latent positions while estimating and removing each rater's distortion of the common scale [19, 20]. For Tayyar the worry is live in two ways. Across actors, a single axis can simply fail to be commensurable: “West alignment” names an existential strategic baseline for an Israeli party and a contested ideological choice for an Arab one, so the same number ends up encoding different facts on the two sides of that divide. Across raters, different models may quietly anchor the -10 to $+10$ scale in different places. I don't solve incomparability. What I do is manage it, by anchoring the rubric in worked regional examples that pin down what the poles mean, by scoping axes so they're only asked where they apply, and by flagging the axes whose cross-divide comparability is weakest. Partial defences, not solutions.

Language models as annotators. The nearby literature reports that frontier models are strong text annotators. Gilardi, Alizadeh, and Kubli find ChatGPT beating crowd workers on several political-text tasks [21]; Törnberg reports much the same for annotating the ideology of political messages [22]; Ziems and colleagues survey the broader promise for computational social science while cataloguing where it can go wrong [23]. A parallel strand insists that LLM annotations be *validated* against human labels rather than just trusted, and shows that unvalidated automated coding can be confidently wrong [24]. Tayyar tries to take both messages at once. It uses models as annotators because, in a region where trained human coders and reference corpora are thin on the ground, they're the most capable raters you can get at scale. And it builds the validation in as planned human anchoring instead of asserting the models are right.

Language models as position estimators. Closest to this paper is a young literature that uses language models not to annotate text but to place actors directly in a policy or ideological space. Wu, Nagler, Tucker, and Messing scale US senators on a liberal–conservative dimension by asking a model for pairwise comparisons, and recover a scale that correlates strongly with NOMINATE [25]. Le Mens and Gallego push the approach the furthest: they ask an instruction-tuned model where a text stands on a focal dimension and *average* its responses, recovering manifesto and roll-call benchmarks at correlations above 0.9 across several languages, while warning, in so many words, that the method has to be re-validated in each new setting and not just trusted on the strength of those correlations [26]. Tayyar shares the premise that a model can read a position, but it parts ways in four places the rest of the paper develops. First, it treats the models as a *panel* of distinct raters and keeps their disagreement rather than averaging it into a point, since the disagreement is itself diagnostic (Section 7). Second, it runs in a region with no large validated reference corpus to calibrate against, so it can't lean on the benchmark correlations that make averaging look safe in well-served party systems; reliability has to be argued from the panel itself, and the threat of a shared prior met head-on. Third, it gates applicability, declining to score an axis that doesn't apply rather than emitting a number anyway. Fourth, it takes Le Mens and Gallego's warning, that a high correlation, or a high α , can hide a shared error, and runs with it as the organising idea, putting the reliability-versus-validity distinction at the spine of the method rather than tacking it on as a closing caveat.

The political leanings and non-independence of models. A panel of models can't be read like a panel of independent experts, for the simple reason that models aren't independent. A fair amount of work now documents that language models carry measurable, and often shared, political tendencies. Santurkar and colleagues show

whose opinions models reflect and whose they leave out [27]; Durmus and colleagues measure how some national viewpoints get over-represented relative to others on global-opinion tasks [28]; Atari and colleagues argue the representations skew toward Western, educated, industrialised populations [29]; Feng and colleagues trace political bias from pretraining data all the way through to downstream behaviour [30]; and several studies find a consistent left-of-centre tilt across widely used models [31, 32], with the deeper diagnosis that the training corpora, and the people who built them, shape the prior [35]. A related literature on “silicon sampling”, using models to stand in for human survey respondents, finds the simulations convincing but unreliable: faithful to some subpopulations and badly off for others [33, 34]. The lesson from all of this isn’t that models are useless raters. It’s that their agreement is suspect in a specific, measurable way. Drawn from overlapping corpora, they may share a prior, and about a region like the Middle East that prior is plausibly an Anglophone-media one. This is the threat the reliability analysis of Section 7 takes on directly, and that, in the end, only the planned human anchoring can settle.

Reliability theory. The discipline’s standard reliability coefficients are Cohen’s κ for chance-corrected categorical agreement [39] and, for ordered or interval scales where the raters don’t all score every item, Krippendorff’s α [37, 38], which copes with the missing data a partially-filled panel always ends up with. I lead with α . Following the advice that you report the coefficient under the measurement assumption you’re actually prepared to defend, I report it under both an interval and an ordinal difference function.

Regional scholarship. Finally, the frame and its exemplars lean on the area-studies literature: Lynch on the post-2011 uprisings [40], Hinnebusch on Ba’athist Syria [41], Wickham on the evolution of the Muslim Brotherhood [42]. The upshot is that the anchors fixing each axis come from a defensible reading of the region rather than a model’s folk theory of it. Anchoring in this literature is what lets the rubric, rather than the rater’s prior, carry the meaning of the scale.

3 The instrument: frame, applicability, and lenses

3.1 Axes

The dataset runs on sixteen active axes (Table 1). Each one has two named poles, a written definition that does the actual governing of how things get scored, and a scope that decides where it applies at all. Eight are fairly universal dimensions of political contestation. The rest are regional or context axes meant to catch cleavages a Western frame either misses or gets wrong, like pan-Arabism, the posture toward Iran, the stance on a confessional power-sharing system. The default two-axis compass picks economic and social, mostly to stay close to the political-compass tradition, but the clustering and comparison analyses use the full sixteen-dimensional position vector. This is one defensible way to carve up the region’s politics, not the only one, and Section 8 comes back to what the choice leaves out.

3.2 Applicability, and why a blank is not a zero

Most axes are universal, but a handful are scoped, and how you handle scope turns out to be a measurement decision that actually matters. Pan-Arabism just isn’t a meaningful question to ask of a Jewish-Israeli party. The stance on confessional power-sharing applies only where such a system exists, as in Lebanon’s Taif arrangement or Iraq’s *mubasasa*. Context axes apply only to their countries. So when an axis doesn’t apply to an actor, Tayyar leaves the cell *blank* instead of scoring it zero, and both scoring and rendering skip it.

This isn’t a cosmetic choice. In a spatial model a score of zero says the actor sits at the midpoint of a contested dimension, whereas an inapplicable axis says the actor isn’t on the dimension at all, and coding the second as the first quietly injects a phantom centrist position that drags every aggregate over that axis toward the middle. A region’s mean “sectarianism” score, if you zero the inapplicable cells, ends up measuring mostly how many actors the question

Axis	– pole	+ pole
Economic	Statist	Market
Social	Authority	Libertarian
State & religion	Religious state	Secular state
Liberal democracy	Weak / anti	Strong commitment
West alignment	Anti-Western	Pro-Western
Regional stance	Resistance / maximalist	Stability / normalisation
Palestinian question	Opposed	Pro-Palestinian rights
Civil liberties	Restrict	Expand
Regime stance	Anti-regime	Pro-regime
Pan-Arabism	Particularist	Pan-Arab
Federalism	Centralist	Federalist
Modernisation	Traditionalist	Modernising
Gender equality	Patriarchal traditionalism	Gender equality
Iran posture	Anti-Iran / adversarial	Pro-Iran / aligned
Press freedom	State-controlled press	Free press
Sectarian power-sharing	Consociational / quota	Post-sectarian / civic

Table 1. The axis frame. Each of the sixteen active axes carries a written definition and worked regional anchors that fix its poles; the definitions, not the labels, govern scoring.

doesn't even concern. "Outside the question" and "in the centre of the question" are just different facts, and a measurement that can't tell them apart will mislead any downstream model that averages, clusters, or correlates over the axis. I don't really see how you handle a region whose cleavages reach so unevenly without keeping the blank-versus-zero distinction.

3.3 The lens system: declared, behavioural, composite

For a lot of political actors a single number papers over the gap between what they profess and what they actually do. So Tayyar models a position as up to three readings. The *declared* lens is the stated position, read off platform, charter, and dated speech, and it's the reproducible spine of the measurement, since two raters reading the same document can be asked to converge. The *behavioural* lens is what the actor does with power, meaning its legislative record and its conduct in office, and it's there to test the declared position. The *composite* is the published synthesis.

Rather than apply one rule everywhere, each cell gets assigned to whichever construct fits it. The declared lens is the default. The behavioural lens is scored *selectively*, for actors who hold or have held power, on the conduct axes of liberal democracy, civil liberties, and press freedom, where a manifesto promise is cheap and the governing record is what you really go on. A record in office speaks to those three axes more directly than to the others. How a party treats courts, dissent, and the press shows up in its conduct in a way that its pan-Arab sympathies or its abstract stance toward the regime's legitimacy don't, and an actor that has never governed simply leaves no such record to read, so widening the behavioural lens to the other axes would mostly manufacture missing data. The approach borrows from the behavioural turn that distinguishes what regimes say from what they do [16]. An actor whose character changed across eras is modelled as separate era-stamped entities rather than asked to hold the contradiction inside one cell; the founding 1947 Ba'ath of Aflaq's pan-Arab charter is not the Assad regime that governed in its name. That also clears up a tension that would otherwise corrupt the reliability statistics, because a good share of the cells where the panel disagrees most aren't coding errors at all but declared and behavioural readings that an undifferentiated score had been quietly averaging together. Splitting the constructs apart turns that spurious disagreement into two well-posed measurements.

Layer	Count
Countries in the frame	20
Political parties (current and historical)	98
Public figures	274
Active axes	16
Source documents (platforms, charters, speeches)	286
Bills and legal instruments	49
Timeline events	98
Verified attributed quotes	101
Panel ratings recorded	12,867

Table 2. Coverage at the vo.2 snapshot. Parties are scored across fifteen of the twenty framed countries; the panel fills country by country.

4 Data and provenance

The unit of analysis is the party or the public figure. Both current and dissolved actors are in there and era-stamped, so historical comparison is supported directly. Scored positions are grounded, where grounding exists, in primary documents: party platforms and charters, dated speeches, official statements, founding constitutions. Secondary sources are used to corroborate biographical facts, not positions. Each source carries an authority tier and, for non-English originals, a record of the language it was written in. That last detail matters, because a model’s reading of a translated text and of the original don’t have to agree. Verification is explicit and runs through three states, *unverified*, *tarek-verified*, and *externally-verified*, and only *non-unverified* rows surface publicly.

I want to be plain about the provenance, since it bears on how the reliability evidence should be read. As it stands the snapshot is largely the work of one principal investigator with model assistance. The verification ladder, the document-grounding layer, and the human anchoring I have planned are all there to keep that provenance legible and improvable instead of hidden. My sense is that a reader should treat the current dataset as a well-instrumented pilot: its internal agreement is real, and its external validation, as I keep insisting myself, hasn’t been done yet.

5 The rater panel and its synthesis

5.1 The panel

A position in Tyyar comes out of a rater panel. The choice everything else hangs on is to treat each language model the way an expert survey treats a single expert [13, 14, 17]: a judge who’s worth something once you pool them and study them, but not on their own, and who can simply be wrong. The analogy does break down in one place that matters, and I come back to it in Section 7. You assemble a human expert panel for a spread of training and background. These raters share web-scale corpora instead, and two of them, Gemini and Gemma, come from the same builder. So whatever epistemic diversity the panel has is something to test, not something to assume.

Each cell can carry ratings from up to four kinds of rater. An *author-baseline* hand-coding seeds the panel as an explicit prior. It’s the one rating in the system that isn’t a language model, and the panel gets tracked against it. The *language models* number nine right now: Claude, Gemini, Grok, GPT, and Gemma, built in the United States, and Kimi, DeepSeek, Qwen, and MiniMax, built in China, drawn from eight laboratories across two countries (Appendix A lists the specific version each model ran). Each one scores every cell on its own, in a fresh session, from a blind template that shows the axis definition and poles and none of the other raters’ answers. Models get added as they show up. A model that joins partway through scores from that point on, and both the synthesis and the reliability statistics will tolerate a rater with partial coverage. A tenth model was trialled and dropped because it couldn’t get through the scoring instrument reliably, and a rater that can’t finish the task isn’t really a rater. *Human*

anchors, scored by people rather than models, are a planned round, and their job is the validity check the model panel can't run on itself (Section 7). They aren't in the data yet, and the paper doesn't pretend they are. A small *grounded pilot*, where a model scored straight from the cited documents rather than from its own knowledge, is kept separately as a reference for later document-grounded verification, and not counted as a panel vote.

5.2 Synthesis, and the preservation of disagreement

Each rater's ratings for a cell collapse to that rater's median, and the published score is the median *across distinct raters*, so a model that scored a cell in three rounds counts once and not three times. I use the median rather than the mean. With a panel of nine raters, one model that badly misreads a cell shouldn't get to drag the published number around, and the median is the robust choice an expert-survey methodology would land on anyway.

The synthesis keeps disagreement around rather than dissolving it. Dispersion is summarised by the median absolute deviation (MAD) of the raters' scores about the published median, on the -10 to $+10$ scale, and each cell gets a status from fixed cutoffs. A cell is *settled* when it has at least two raters and their MAD about the published median is within tolerance ($MAD \leq 1.5$). It is *contested* when they diverge sharply ($MAD \geq 3$), and in that case the cell is rendered and released as a range rather than a point. It is *provisional* otherwise, which means a single rater so far or a spread that sits between the cutoffs. The thresholds are pragmatic, and the underlying MAD travels with each cell so a reader can redraw them however they like. A companion agreement figure, $1 - MAD/10$, rides along with every synthesised cell as a convenience for the live tool. It just rescales dispersion onto $[0, 1]$ for display and isn't a statistical construct. The frozen definition-free round and the grounded pilot are both kept out of the published median: the first because it's the no-rubric baseline held back for the experiment of Section 6, the second because folding it in would let one model vote twice. At the current snapshot the panel has recorded 12,867 ratings. Of 1,108 active composite cells, 795 (72%) are settled, 58 (5%) are contested, and 255 (23%) remain provisional, and 998 cells (90%) already carry at least two independent model raters. Because the panel fills country by country, the balance shifts with every scoring session, so the published figures are interim and I'd treat them that way.

5.3 An inclusion gate on aggregation

A cell's status governs that cell, and nothing more. Field-level quantities are a second layer of inference: which axis the region is most divided on, how two party families compare, the correlations among axes. A sparse axis can quietly poison all of these. With only a handful of actors scored on an axis, a single outlier swings the mean, a high variance reads as "contested" when it's really just thin, and a spurious correlation can end up topping a table. So Tayyar lets an axis into field-level aggregates only once its coverage clears a floor, currently thirty scored entities. That figure is a convention I picked rather than something derived, but it does land in an empty band of the coverage distribution. No axis has between roughly twenty and sixty-five scored cells (its per-axis n in Table 4), so the core-versus-periphery split it produces comes out the same for any floor across that range. Coverage is the structural gate, because coverage is what makes an average mean anything at all. Cross-rater dispersion is reported alongside it as a separate signal, so a well-covered but genuinely contested axis gets flagged rather than quietly trusted, but coverage is what decides eligibility. Axes below the floor are reported in full, on their own, and never folded into a headline. In the data as it stands this separates a core of well-covered dimensions from a periphery of scoped axes (sectarian power-sharing, Iran posture, press freedom, gender) whose small cell counts make them informative about individual actors but not yet about the field. The gate isn't glamorous, but I don't think the dataset works without it. Most of the ways a positional dataset misleads at the macro level come down to treating thin coverage as if it were a finding.

6 The rubric effect: a built-in experiment

The standard objection to using a language model as a rater is that the rubric does the work: hand a compliant model a definition with a built-in direction and it'll dutifully echo that direction back, so any apparent agreement is really

just measuring how suggestive the prompt was. It’s a fair objection, and a real one. The round protocol is set up to answer it with a number rather than a reassurance.

Ratings are labelled by round. The first round (r1) ran on two models, Gemini and Grok, *without* the axis definitions: each model got the poles and was asked to place the actor from its own understanding, and then the round was locked. Every later round runs *with* the sharpened definitions and the construct rule of Section 3. Holding r1 fixed turns the worry into something you can measure. The movement from r1 to r2 on the cells scored in both is the effect of bringing in the rubric, and the shape of that movement is what tells us whether the rubric steered or corrected.

A couple of things come out of this. The first is about within-rater movement. Across the 376 rater-cell pairs scored in both rounds, the mean absolute shift is 1.79 points on the 21-point scale, with r1 and r2 scores correlated at $r \approx 0.90$. So the rubric leaves the central tendency of most placements where it was, which is what you’d expect if the models had reasonable priors to start with, but it does matter in the tail: roughly one cell in five (19%) moves by three points or more. When I look at those big moves they read as corrections rather than noise. In one case that’s fairly typical, a model placed Likud at +4 on the liberal-democracy axis without definitions, taking “holds elections” to mean “committed to liberal democracy,” and then at -3 once the rubric spelled out that the axis is about the *conduct* of power and not just the holding of elections. That move pulled an outlier back in line with the rest of the panel and with the behavioural record.

The second thing is that the rubric improves agreement *between* raters. On the 187 cells both models scored in both rounds, bringing in the definitions tightened the mean absolute gap between Gemini and Grok from 2.81 to 2.50 points and raised their correlation from 0.81 to 0.89. The steering objection has a hard time with this. A rubric that just imposed some arbitrary direction would shift the scores without necessarily making two independent raters agree any more closely about where an actor sits. The definitions did both, which rules out a purely arbitrary or orthogonal steer. By itself that still doesn’t separate convergence toward correctness from convergence toward an anchor the rubric is quietly supplying. Only lining the scores up against an external behavioural record, as in the Likud correction above, and eventually the planned human-anchor round, can do that. So what the experiment shows is narrower than you might want, but it holds: the rubric’s first-order effect is to turn vaguer individual readings into a more reproducible shared one.

Two raters are a thin base for a general claim, and I’m aware of that. The very thing that makes the test clean, that the *same* two models scored the same cells with and without definitions, is also what limits it. A bigger replication, even one that’s less perfectly paired, helps with generality. When the no-definition round is run across the panel on a fixed, randomly drawn 204-cell anchor (the persistent inter-rater sample) and compared with the rubric round on those same cells, mean pairwise between-rater disagreement falls from 2.43 to 2.07 points. Same direction, comparable magnitude, now across the nine surviving raters from eight laboratories rather than two. So the convergence isn’t an artefact of which two models happened to run the definition-free round first.

I think the design carries beyond this dataset. Any rubric-based use of a model rater can run the same held-fixed, definition-free baseline and report what the rubric is actually doing, instead of just asserting that it clarifies.

7 Reliability

The reliability question for a panel instrument is whether the same frame, applied by different raters, lands in the same place. The figures here are interim. They cover the 89 parties scored to date (positions for individual public figures come in a later round), across fifteen countries, and they will move as the panel fills. So read them as a snapshot of a measurement still in progress, not as a verdict.

7.1 Agreement across the model panel

On the 998 cells that carry at least two independent, blind model scores, the panel agrees closely. Krippendorff’s α across the model raters is **0.86** on an interval difference function. As in the published synthesis (Section 5), each

Model	Built	Laboratory	Cells	\bar{r}	mean $ \Delta $
Claude	US	Anthropic	997	0.90	1.89
GPT	US	OpenAI	961	0.89	1.95
Gemma	US	Google	1,001	0.87	2.21
Gemini	US	Google	997	0.87	2.30
DeepSeek	China	DeepSeek	934	0.86	2.25
Kimi	China	Moonshot	997	0.86	2.19
MiniMax	China	MiniMax	974	0.86	2.18
Qwen	China	Alibaba	997	0.85	2.33
Grok	US	xAI	997	0.84	2.34

Table 3. Per-model agreement. Each model’s agreement with the rest of the panel on the r2 cells, ordered by mean pairwise correlation. \bar{r} is its mean Pearson correlation with each other model over shared cells; mean $|\Delta|$ the matching mean absolute gap on the -10 to $+10$ scale; “Cells” its r2 coverage. Every \bar{r} falls in 0.84–0.90, and build country does not order the column—the highest and the lowest are both US-built, with the China-built models between.

rater contributes a single value per cell, its within-rater median, so the coefficient is computed over distinct raters and is not inflated by repeated rounds from the same model. The steps on the -10 to $+10$ scale may not be exactly equally spaced. Whether the distance from -2 to -1 is really the distance from 0 to $+1$ is an assumption, not a fact. So I also report α on an ordinal difference function, which assumes only that the values are ordered and weights a disagreement by how much of the observed distribution sits between the two ranks. The ordinal α is **0.86**, essentially identical to the interval figure. Because the two line up, the reliability claim doesn’t hang on treating the scale as evenly spaced, which is the assumption a lot of interval reliability reporting just leaves unsaid and the one a methods reviewer is most entitled to push on. The coefficient held at 0.86 as the panel grew from five raters to nine, which is at least some comfort against one narrow worry, that the early figure was an artefact of a handful of near-identical raters. What it can’t speak to is a prior shared across all nine; widening the panel would carry such a prior along rather than expose it.

Per-rater, each model’s average agreement with the others is high (Table 3: every model’s mean correlation with the rest of the panel sits between 0.84 and 0.90, with mean absolute gaps around two points). The panel as a whole tracks the independently authored, non-model baseline, the one comparison I have today that isn’t itself a language model. But that tracking is worth only so much. The baseline is my own hand-coding, written by the same hand that wrote the axes, the anchors, and the rubric the models are given, so the panel agreeing with it is what you’d see if the rubric had simply passed my prior along to the models faithfully. It’s a coherence check on the instrument, not an external validation of it. Only raters independent of the frame can give you that.¹

7.2 Where the panel converges, and where it frays

Agreement is far from uniform across the frame. The per-axis pattern (Table 4) is the part you can act on, since it points straight at the rubric anchors that most need tightening. The panel converges hardest where a position is loudly and repeatedly declared (gender, the Palestinian question, the social and state-and-religion dimensions), and two readers of the same texts land in the same place. (Iran posture and press freedom look comparably tight in Table 4 but rest on thin coverage, $n = 17$ and $n = 19$, so I wouldn’t lean on their agreement yet.) Where it frays, the reasons are worth naming, mostly because they aren’t the ones you’d guess.

Take the single widest disagreement in the data. On regime stance, Hayat Tahrir al-Sham draws scores spanning the *entire* scale (Figure 1): five models place it near the revolutionary floor (three at -10 , two at -9) and four place

¹A separate internal re-coding probe re-scored a 21-cell stratified sample across three blind rounds (Appendix B); the rubric’s test–retest consistency on it runs at $\alpha \approx 0.95$. That is a consistency floor, a measure of how stably the rubric maps a text to a number on re-application, which one framework reproduces almost by construction. It is not a reliability result, and it is recorded only for completeness; the useful content is the relative pattern across axes, not the figure.

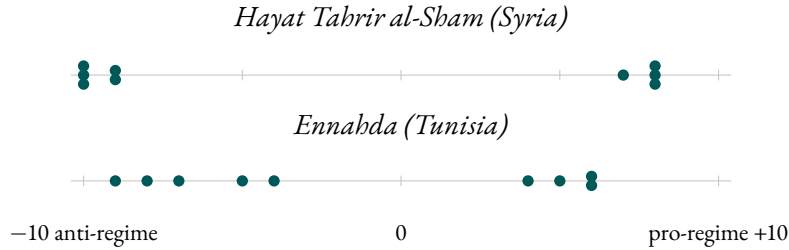


Figure 1. The referent problem on regime stance. Each dot is one of the nine models’ scores for the actor; coincident scores are jittered vertically. Both actors split into two clusters separated by most of the scale, not a spread around a centre: the models agree on the facts and disagree on *which* constitutional order “the regime” denotes. For Hayat Tahrir al-Sham the lower cluster scores its posture toward the Ba’athist order it overthrew (anti-regime) and the upper cluster toward the post-Assad order it now governs (pro-regime); for Ennahda, toward Kais Saied’s incumbent order, which it opposes, versus the 2011 republican order, which it defends. The split does not track build country; each cluster mixes US-built and China-built raters, so it is about interpretation rather than nationality. A blind triple-coding of the twenty-two widest regime-stance cells attributes 64% of such disagreement to interpretation of this kind rather than to fact or definition.

it near the incumbent ceiling (between +7 and +8). This isn’t a coding error, and it isn’t just a knowledge gap either. In the eighteen months before this writing HTS went from the most heavily sanctioned jihadist franchise in the Levant to the government in Damascus. You might guess the split just separates the models that learned of the transition from the ones that didn’t. It doesn’t. In their written rationales, raters on *both* poles state the same facts, that HTS toppled the Ba’athist order and now governs Syria, and they divide instead on *which* order “the foundational order” denotes. A rater scoring the actor’s posture toward the order it overthrew reads it as anti-regime (−10); a rater scoring its posture toward the order it is now building reads it as pro-regime (+8). This is a *referent problem*, not a factual one. The axis quietly assumes there’s a single stable order to take a stance toward, and an actor that has just swapped one order for another breaks that assumption. Ennahda splits the same way across fifteen points, opposing Kais Saied’s incumbent order while defending the 2011 constitutional order it helped write, and so do the Syrian Ba’ath, the Sudanese Communist Party, and Yemen’s Islah, each an actor whose country offers more than one plausible “regime” to be for or against. There’s a sharper version of the same thing, referent *absence*: in a state mid-collapse, with no settled order to be for or against, the axis isn’t ambiguous so much as ill-posed. The applicability rule (Section 3) keys on whether an axis applies to an actor in principle, not on whether a stable referent happens to exist at a given moment, so it doesn’t yet catch this case. Scoping the order-dependent axes to leave such cells blank, instead of rendering them as a contested range, is a refinement I think the rule should absorb.

To check this is really the mechanism and not just a flattering story about a few cells, three coders, working blind and independently, classified the source of disagreement on the twenty-two most divided regime-stance cells (they agreed on 76% of pairwise judgements). A clear majority traced to *interpretation* rather than to fact or to definitional confusion. Roughly two-thirds (64%) turned on which order is the referent, or on how to weigh an actor’s posture toward it, against 18% genuine factual conflict and 18% definitional. Written definitions are the rubric’s tool for removing that second kind of disagreement, and on this axis they don’t have much to work with. What divides the panel here isn’t what “regime stance” means but which regime an actor is scored against, and fixing that referent for a contested actor is a political judgement the instrument has no business making. So the synthesis marks these cells contested rather than averaging an insurgent with an incumbent, and the dataset flags the column to be read with that limit in mind.

This is the sort of thing I was hoping the panel would catch. A single-number method would average the two clusters into a meaningless midpoint. Keeping the split shows you there’s a real ambiguity in the question itself. For an actor that has just replaced one order with another, “stance toward the regime” has no single answer, and a confident zero would be the wrong one.

A second cluster sits on sectarian power-sharing, the loosest axis in the table, and what it exposes is a definitional

Axis	n cells	mean Δ
Palestinian question	80	1.35
Gender equality	20	1.52
Iran posture	17	1.55
Press freedom	19	1.82
Social	80	1.85
State & religion	80	1.88
Modernisation	76	1.90
West alignment	79	2.01
Economic	80	2.02
Civil liberties	79	2.07
Liberal democracy	80	2.26
Regional stance	79	2.46
Federalism	76	2.73
Regime stance	76	2.87
Pan-Arabism	66	3.22
Sectarian power-sharing	11	4.15

Table 4. Per-axis agreement. Mean pairwise absolute gap across the panel, tightest to loosest, on the -10 to $+10$ scale. Scoped axes have small cell counts and are noisier—read each row with its n ; the inclusion gate (Section 5) keeps the thinnest out of field-level aggregates.

trap rather than a factual one. Hezbollah, the Amal Movement, and the Syrian Ba’ath each draw spreads of thirteen to fourteen points, with raters scattered pole to pole. The axis is meant to measure a *stated position* on the confessional quota, whether an actor would abolish *mubasasa* or entrench it. But a model that hears “Hezbollah” and scores a sociologically sectarian actor is answering a different question from one that reads the party’s own anti-confessional rhetoric and scores the platform. Both readings are defensible. They’re just not answers to the same question. This is the construct-conflation the lens system (Section 3) is there to resolve.

The same split shows up across the frame. The panel agrees about what actors *say* and frays about what their position *is once you have to infer it*. That holds on the conduct axes (liberal democracy, civil liberties), where a charter and a record pull apart, and on the scoped regional axes (federalism, regime stance, pan-Arabism, and above all sectarian power-sharing), where raters differ about what the axis is even asking before they get to the answer. The conflict-defining regional axes are among the *least* settled, which is more or less the diagnosis a reliability programme is supposed to produce. It tells the next round of rubric work where to go, and it warns the reader which columns of the dataset to lean on and which to treat with caution.

There’s a real tension here. The panel is tightest on the universal, Western-legible axes, the social, economic, and state-and-religion core, and loosest on the regional axes (pan-Arabism, sectarian power-sharing, regime stance) that are its reason to exist in the first place. It’s most reliable where it adds least, and least reliable where it claims the most. I read that as diagnosing a problem rather than refuting the method. The regional, evaluatively loaded constructs are the ones carrying the interpretive disagreement the next subsection picks apart, and an instrument that reported them as settled would be hiding that problem rather than measuring it. The implication for use stands either way, and the inclusion gate of Section 5 enforces it: the regional columns inform claims about individual actors and are kept out of field-level aggregates until their coverage and agreement support the weight.

7.3 Reliability is not validity

Everything in this section is a statement about *reliability*, the reproducibility of a reading, and not yet about *validity*, its correctness. That distinction matters more than almost anything else in the method. Most of the panel’s raters are language models drawn from overlapping training data. Nine models agreeing is not nine independent

Axis	within-bloc	between-bloc	difference
Regime stance	2.69	3.01	+0.31
Federalism	2.59	2.85	+0.26
Iran posture	1.41	1.67	+0.26
Civil liberties	1.97	2.15	+0.18
Sectarian power-sharing	4.11	4.18	+0.07
Pan-Arabism	3.20	3.24	+0.04

Table 5. The provenance contrast is small. Mean absolute between-rater gap within build-country blocs versus across them, on the -10 to $+10$ scale, for the six axes with the largest between-minus-within difference (the rest are smaller or negative). Even the largest, on regime stance, is a third of a point, and the axes a geopolitical-bias account would target hardest—pan-Arabism and sectarian power-sharing—show essentially none. Build country is a coarse proxy for corpus independence (see text). Differences are computed on unrounded values and may differ from the rounded columns by 0.01 .

confirmations [23]. The nine can share a single prior, and where the subject is the politics of the Middle East, a prior coloured by Anglophone media coverage is just the kind of shared error I’d most want to rule out. A high α is perfectly consistent with a reliable instrument that’s also systematically wrong, and no statistic computed over models alone can tell the two apart. The defences against this live outside the panel: the independently authored baseline that the panel tracks (a single non-model anchor, helpful but not enough on its own), the document-grounding layer that ties scores to cited text, and, the one that matters most and isn’t collected yet, human anchors scored by people. No statistic computed over the models can stand in for that check.

One internal check makes the gap concrete instead of just rhetorical. The archived grounded pilot, in which a single model scored from the cited documents instead of from its own knowledge, overlaps the parametric panel on 117 cells. There its grounded scores correlate only $r = 0.65$ with its own parametric scores and differ from them by a mean of 3.7 points on the 21-point scale. That’s a wider gap than any model’s mean disagreement with the rest of the panel (Table 3, where those run 1.9 to 2.3 points), with a three-point-or-larger move on 44% of the cells. Re-reading the actual documents moves the same model more than swapping one model for another does. This is one model on a small, early sample, so it’s not a verdict on either reading, but it points the same way the rest of this section does. The panel’s close agreement is agreement about a shared parametric reading, and grounding that reading in cited text, rather than piling on more models, is the validity work that’s left.

There’s one check I can run from inside the panel itself. It mostly comes up empty, but it’s worth putting on the record. If the shared prior were specifically a US or Anglophone one, then models built in the United States and models built in China should disagree where that prior, rather than the text, is driving the score. I measure this per axis as the mean absolute gap between US-built and China-built raters on the cells both blocs scored (the *between-bloc* gap), set against the mean gap among raters of the same build country (the *within-bloc* gap, a five-rater quantity on the US side and a four-rater one on the China side). On no axis does the between-bloc gap exceed the within-bloc gap by more than a third of a point (Table 5). The largest gap is on regime stance ($2.69 \rightarrow 3.01$, $+0.31$); the next, on federalism and Iran posture, never exceeds $+0.26$; and on the two axes a naive geopolitical-bias story would go after hardest, pan-Arabism and sectarian power-sharing, the difference is basically zero. The split is small, and it doesn’t single out the China-built models.

I wouldn’t lean on this too hard, for a couple of reasons. Build country is about the crudest proxy you could pick for what actually matters, corpus independence: Google built two of the nine raters, “US-built” lumps together corpora that share little, and four raters is a thin base for a within-bloc dispersion. And a null here can’t rule out a prior shared across *all* nine models whatever their origin, which is the harder and probably likelier worry anyway. A fuller treatment of the provenance contrast I’m leaving to companion work. The test that would actually settle it is external to the models altogether, and the validity round that would supply it is the method’s main unfinished business.

8 Limitations

- **Correlated model error, and the missing validity check.** The panel is mostly language models drawn from overlapping corpora. Their agreement is evidence of *reliability*, not of independence (Section 7). A contrast of models by national provenance, developed in companion work, gets at whether a shared prior is doing the work, but spotting a bias and removing it are not the same thing. What would actually turn reliability into *validity* is human anchors, scored by people who are blind to the panel. That round is planned and not yet collected. Until it is, I'd ask the reader to keep that gap in mind for every reliability figure here. And the gap won't close evenly, for reasons that are built into the data rather than a matter of timing. Human anchoring is doable for present-day, reachable actors but hardest right where the data is thinnest: defunct parties, and actors in active conflict zones where local expertise is scarce or unsafe to ask for. For those cases the document-grounding layer, which ties a score to cited text rather than to some reachable living expert, is the validity path that has to carry the weight. So validity is going to show up unevenly across the dataset rather than all at once.
- **Stated versus revealed.** The *declared/behavioural* lens manages the oldest problem in positional measurement without making it go away. Which object a number refers to is a choice the analyst makes and has to declare. If you'd draw that line somewhere else, then some of the cells should read differently for you.
- **Ungrounded model rounds, and why grounding is the higher-value work.** The current rounds score from the models' parametric knowledge plus the rubric, not from the cited documents directly. They are best read as "what well-informed models believe, given clear definitions," with document-grounded scoring as a separate verification layer still being built [7, 24]. Reading parametric knowledge means the rounds may be pulling out the shared prior itself, so the reliability statistic and the scoring construct can line up for the same reason: both are reading the prior rather than the actor. Grounding the score in cited text (Section 7) would close that gap, and it's the unfinished work I'd put first.
- **Recency, and the moving target.** Politics in the region moves fast, and a model's knowledge of the most recent turns is bounded by its training cutoff. The prompt dates the question ("score the actor as it stands in 2026"). Even so, an actor mid-transition, say Hayat Tahrir al-Sham between insurgency and government, or a party turned out of office between one model's cutoff and the next, can still split the panel across the full scale, because the models are disagreeing about a fact only some of them have learned. The design reads such splits as signal rather than noise. The cells get flagged *contested*, and the synthesis won't average an insurgent with an incumbent. The blind triple-coding of the widest regime-stance cells (Section 7) is one go at telling the two apart. It put 18% of that disagreement down to factual conflict, the bucket that would hold any cutoff-driven split, against 64% interpretation, so cutoff-driven splits are real there but in the minority. Recency is a hard ceiling on any model panel, and the only real fix is fresh scoring once events settle.
- **The frame is one decomposition among several.** Every axis is a defensible choice, and every one is also an editorial call. The West-alignment axis in particular doesn't mean the same thing for Israeli and Arab actors, so cross-divide comparisons on it want some care [18]. How much the headline placements move when a given axis is dropped or redefined isn't something I report yet. A sensitivity analysis of that kind, which the regime-stance case (Section 7) more or less argues for, is a next step that's needed.
- **An uneven, drifting panel.** Not every cell is scored by the same set of models. Per-model coverage varies, so the reliability figures pool across cells whose panels aren't made up of the same models. A commercial model is a moving target too: a *rater* label just means whatever version answered in that session, and recording the model identifiers only partly pins that drift down.
- **Positionality.** An instrument that places contested political actors is not a view from nowhere. The choice of axes, the anchors that fix their meaning, and the applicability rules reflect a particular reading of the region,

written out in full so it can be argued with rather than mistaken for neutral measurement. And they are one investigator's choices. The active axes, the regional anchors, the applicability rules, and the seed baseline are all single-authored for now. I don't think that bottleneck should be allowed to stand: the planned human-anchor and external-expert rounds are there to widen the judgement *behind* the frame, not just the scoring within it. So this is a real limit on cultural validity. The frame bakes in contestable theory, since treating resistance-versus-normalisation as a primary regional cleavage is itself a choice, and no single author holds the range of expertise that axes running from Lebanese confessionalism to Gulf–Iran rivalry would call for. The external-expert round therefore has to mean participatory co-design of the frame, not only sign-off on the scores produced under it, and the frame is going to need revising as the region moves. If you'd draw a line somewhere else, you're using the instrument the way it's meant to be used.

9 Reproducibility and access

The instrument is built so its outputs can be traced back to where they came from. The live tools, the rubric, and a methodology document all read from the same tables as the figures reported here, so the public interface and the published numbers can't quietly drift apart. The dataset is kept as versioned snapshots, that is, CSV and JSON exports of every primary table with a manifest, so a citation can resolve to a fixed state even as the data keep growing. The figures here come from the first such freeze (v0.2). The formal release is still in preparation. On archival deposit the snapshot is to go out under CC BY-NC-SA 4.0 (source texts under fair use, with a documented takedown route) and the instrument code under the MIT licence, with a DOI minted at that point. Until then the live instrument and this paper are the public record, and the snapshot and code are available from the author on request.

Because the synthesis keeps disagreement around rather than hiding it, each cell is exported with its median, its dispersion (MAD), its *settled/contested/provisional* status, and, where contested, a range rather than a point. A downstream analysis ought to carry that uncertainty along rather than collapse it. The median does the job where a model needs a point estimate. The dispersion can go straight in as measurement-error variance in an errors-in-variables or Bayesian hierarchical model. A contested range can be multiply imputed over or entered as bounds. And a fixed rule, like dropping contested cells or down-weighting by the agreement figure, keeps the decision out in the open. Treating a contested cell as a confident point would just bring back the false precision the synthesis is there to avoid. Whichever rule a study goes with is a choice it should state rather than make quietly.

10 Conclusion

What I'm arguing in this paper is that the right way to use a language model in positional measurement is the way the discipline already knows how to use a fallible human expert: as one *rater* in a panel, pooled and studied rather than trusted, with its agreements scrutinised and its disagreements kept. Taking that stance turns a set of vague anxieties about "using AI for social science" into questions you can actually answer. The first is whether the rubric steers the model or sharpens it. The built-in experiment answers part of that. It moves scores in the tail and tightens agreement between raters, which rules out a merely arbitrary steer. Whether that convergence is toward correctness, rather than toward an anchor the rubric itself supplies, is for the external behavioural checks and the planned human-anchor round to settle. The second question is whether the raters agree, and whether the answer holds up once you drop the equal-spacing assumption. They do, on both an interval and an ordinal metric. The third is whether that agreement is evidence of correctness or of a shared prior. It isn't, on its own, and no figure the models can produce about themselves is going to settle it. And where the panel does come apart, most sharply over which constitutional order an actor should be scored against, the disagreement turns out to be interpretation rather than error, a limit that no amount of sharpening the rubric can legislate away.

What's left is the *validity* check the method is built to receive but hasn't yet been through: human coders, scoring blind, against the same rubric. That round and the document-grounding layer are the next steps, and they

Rater	Laboratory	Version	Built	Cells
Claude	Anthropic	Claude Opus 4.8	US	997
GPT	OpenAI	GPT-5.5	US	961
Gemma	Google	Gemma 4	US	1,001
Gemini	Google	Gemini 3.5 Flash	US	997
Grok	xAI	Grok 4.3	US	997
DeepSeek	DeepSeek	DeepSeek V3	China	934
Kimi	Moonshot	Kimi K2.6	China	997
MiniMax	MiniMax	MiniMax M3	China	974
Qwen	Alibaba	Qwen3.7-Plus	China	997

Table 6. The r2 model panel. The version each model ran, its builder and build country, and its r2 cell coverage at the vo.2 snapshot; the “Cells” column matches the per-model coverage in Table 3.

would let the instrument’s agreement be read as more than reproducibility. Where these actors do turn up in an existing instrument, and a few sit in cross-national democracy indices or in expert codings done for other projects, those coordinates make an obvious external check, and I’d welcome their use against the released snapshots. But the overlap is thin because the region is under-served, which is why purpose-built human anchoring, rather than a borrowed benchmark, is the validity work that counts. What I’m offering here is the method and the discipline around it: a regionally anchored frame, an applicability rule that respects the difference between a blank and a zero, a lens system that separates word from deed, a synthesis that keeps disagreement, and a sceptical, experiment-driven treatment of the model raters that won’t let their fluency stand in for their correctness. For a field that can place every party in a handful of well-documented democracies and almost none elsewhere, that discipline is what makes the rest of the map approachable at all. The running example here is the Middle East and North Africa, but the gap the method addresses is a general one: a comparative apparatus that goes quiet wherever the standard sources run out. The proposal is general too. Bring the language model in as a rater, and then hold it to the standards a discipline already applies to its human experts.

A The rater panel: models and versions

Table 6 lists the version of each model in the with-definitions (r2) panel, along with its laboratory and build country and its cell coverage at the vo.2 snapshot. A commercial model keeps shifting under you, so the panel just stores whichever version identifier actually answered in a given session (Section 8); what you see here are those identifiers. All of the r2 scoring happened in June 2026.

B Data provenance by round

Ratings are kept round by round. Only the with-definitions round (r2) feeds the published synthesis and the reliability numbers. The rest do smaller jobs: the seed prior, the definition-free baselines that let me measure the rubric effect at all, the document-grounded reference, and an internal consistency probe. Table 7 ties them back to the headline total of 12,867.

C The axis frame in full

The sixteen active axes, with the definition each rater actually saw. I reproduce the definitions as they went to the panel, since it was the definitions that did the work of scoring and not the pole labels (Section 3); the emphatic capitals are there in the originals. Eight are universal; the rest are scoped to wherever the cleavage actually applies.

Round	What it is	Raters	Cells	Ratings
Author baseline	Hand-coded prior; the one non-model rater	1	1,012	1,012
r1 (no definitions)	Gemini and Grok; the locked definition-free baseline	2	189	376
rib (no definitions)	Full panel, including the trialled tenth model	10	622	2,293
r2 (with definitions)	The published panel	9	1,121	8,855
Grounded pilot	One model scoring from the cited documents	1	205	205
Re-coding probe	Repeated blind re-reads of a 21-cell sample	2	21	126
Total				12,867

Table 7. Ratings by round. Every recorded rating, reconciling to the 12,867 of Table 2. Only r2 is synthesised into published scores and reliability; r1 and rib supply the rubric-effect baselines (Section 6), the grounded pilot seeds later document-grounded verification, and the re-coding probe is the consistency floor of Section 7.

- **Economic** (Statist ↔ Market; universal). Degree to which the actor favors market mechanisms (positive) vs. state allocation (negative).
- **Social** (Authority ↔ Libertarian; universal). Personal and moral social freedom—lifestyle, LGBTQ+ rights, individual expression, religion-in-private—permissive (+) vs traditional collective moral authority (–). This axis is about SOCIAL / MORAL liberty, NOT security policy, nationalism, or economics: a party can be nationalist-hawkish yet socially permissive (secular nationalists), or dovish yet socially conservative.
- **State & religion** (Religious state ↔ Secular state; universal). Separation of religion from state and law (positive) vs. religious authority in state and law (negative).
- **Liberal democracy** (Weak / anti ↔ Strong commitment; universal). Commitment to LIBERAL democracy— independent courts, free press, opposition and minority rights, limits on executive power. NOT mere electoral participation: a party that contests elections but works to weaken the judiciary, entrench ethnic or religious supremacy, or curb minority and dissident rights scores LOW (–). Strong commitment to liberal-democratic checks scores HIGH (+).
- **West alignment** (Anti-Western ↔ Pro-Western; universal). STRATEGIC and security alignment with Western (US/EU) blocs—military cooperation, diplomatic partnership, trade. This is geopolitical orientation, NOT cultural affinity for liberal Western values: a religious-nationalist party can be strongly pro-Western strategically (wants US arms and backing) while rejecting Western liberal social values. Score the strategic posture. Pro-Western (+), anti-Western or hostile-nonaligned (–).
- **Regional stance** (Resistance / maximalist ↔ Stability / normalization; universal). Posture toward the regional order: the “Axis of Resistance” (Iran–Hezbollah– Hamas–Houthis, anti-normalization, armed struggle) at the negative pole vs the normalization / stability camp (Abraham Accords, Gulf–Israel détente, negotiated settlement) at the positive pole. Pro-Palestinian sympathy ALONE does not make an actor “resistance”—score toward resistance (–) only for alignment with armed-resistance or anti-normalization politics; an actor that pursues its regional aims THROUGH negotiation and normalization leans positive (+).
- **Palestinian question** (Opposed ↔ Pro-Palestinian rights; universal). Position on Palestinian rights, statehood, and self-determination—strong support (positive) vs. opposition (negative).
- **Civil liberties** (Restrict ↔ Expand; universal). Speech, protest, dissent, association, press freedom—expand and protect (positive) vs. restrict (negative).
- **Regime stance** (Anti-regime ↔ Pro-regime; universal). Stance toward the state’s foundational constitutional order—for Israel the Zionist-democratic state; for republics and monarchies the existing political order. Pro-

regime (+) accepts and defends that order; anti-regime (−) rejects or seeks to overturn it. This measures stance toward the FOUNDATIONAL ORDER, not the incumbent government: an opposition party loyal to the constitutional order still scores positive.

- **Pan-Arabism** (Particularist ↔ Pan-Arab; Arab actors). Degree to which the actor frames its politics around pan-Arab solidarity (the Nasserist, Baathist, and contemporary anti-normalization tradition) versus country- or community-specific particularism (sectarian, ethnic, monarchic, or nation-state-first frames). One of the major regional cleavages.
- **Federalism** (Centralist ↔ Federalist; universal). Position on state structure: a strong unitary central state versus power-sharing among regions, sects, or ethnic groups. Salient in MENA, where Iraq, Lebanon, Sudan, and Syria have consociational or federal frames while Egypt, Tunisia, Algeria, and Morocco are strongly centralist; Kurdish parties and Yemeni southern movements sit at the federalist extreme.
- **Modernization** (Traditionalist ↔ Modernizing; universal). Position on social and institutional modernization versus tradition: women’s rights, religious authority in public life, reform of family law, openness to technological and economic transformation. Distinct from the secular–religious axis: an actor can be religious and modernizing (Vision 2030) or secular and traditionalist (state-corporatist nationalism).
- **Gender equality** (Patriarchal traditionalism ↔ Gender equality; universal). Position on women’s rights, gender roles in public life, family law (personal status, divorce, custody, inheritance), and political representation. Distinct from the broader modernization axis: a party can be tech-modernizing and gender-conservative.
- **Iran posture** (Anti-Iran / adversarial ↔ Pro-Iran / aligned; universal). Position toward the Islamic Republic of Iran as a regional actor: alliance, hostility, or neutrality. Independent from West alignment—a party can be anti-Western AND anti-Iran (most Sunni Salafi groups), pro-Western AND anti-Iran (Saudi, UAE, post-1979 Israel), or anti-Western AND pro-Iran (Hezbollah, Houthis, post-2003 Iraqi Shia parties).
- **Press freedom** (State-controlled press ↔ Free press; universal). Position on journalism, broadcast media, and digital information access. Distinct from broader civil liberties: a regime can selectively restrict the press while allowing other speech (or vice versa). Captures both legal frameworks (defamation law, licensing, internet shutdowns) and political enforcement.
- **Sectarian power-sharing** (Consociational / quota ↔ Post-sectarian / civic; confessional). Position on confessional or sectarian political organisation: Lebanese-style consociational quotas, Iraqi muhasasa, Syrian Alawite-network governance, versus post-sectarian / civic-state framings. Relevant primarily to Lebanon, Iraq, Syria, and Bahrain; less so to Egypt or Saudi Arabia, where the question does not structurally arise.

D The scoring prompt

Every rater—model or human—received the same blind template: one row per (actor, axis) cell, with the axis definition and pole meanings but no other rater’s scores. Models were given the instruction below verbatim, in a fresh session, and returned the filled table. The definition-free baseline (r1, r1b) used the same instruction with the `axis_definition` column blanked and the definition-following rules dropped, so the r1→r2 movement isolates the rubric (Section 6).

```
You are an independent political-science coder scoring MENA political parties and
public figures on a fixed set of ideological axes. I will give you a CSV. Each row
is one (actor, axis) cell.
```

```
For EVERY row, fill three columns and change nothing else:
- score: an integer from -10 to +10. Read the "axis_definition" column for the
```

- precise meaning of the axis -- it governs. -10 means the "scale_minus10_means" pole; +10 means the "scale_plus10_means" pole; 0 is centrist/mixed.
- confidence: 0.0 to 1.0, how sure you are.
- rationale: one short sentence justifying the score.

What to score:

- Default to the actor's DECLARED position -- what its platform, charter, and official statements commit to.
- Exception, the conduct axes (democracy, civil liberties, press freedom): if the actor holds or has held power, score its actual RECORD in office, not its stated commitment. A government that jails journalists scores low on press freedom whatever its charter says.
- Score the actor as it is today (2026). If its positions shifted across its history, score the current incarnation, not a founding-era doctrine.

Rules:

- The "axis_definition" column is authoritative -- if your intuition from the short pole labels conflicts with it, follow the definition.
- Score from your own knowledge of each actor. Do NOT look for or infer any "correct" answer -- give your independent read.
- Keep the "key" column EXACTLY as given. Do not add, remove, reorder, or rename columns or rows.
- Return EVERY row you were given -- do not truncate, summarize, or stop early. If the table is long, keep going to the last row.
- If you genuinely don't know an actor, still give your best estimate and set confidence low (e.g. 0.2).
- Output ONLY the full CSV (all rows, same columns), nothing before or after it.

Here is the CSV:

References

- [1] Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, 39(153), 41–57.
- [2] Downs, A. (1957). *An Economic Theory of Democracy*. Harper & Brothers.
- [3] Poole, K. T., & Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.
- [4] Budge, I., Klingemann, H.-D., Volkens, A., Bara, J., & Tanenbaum, E. (2001). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford University Press.
- [5] Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., & McDonald, M. (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford University Press.
- [6] Volkens, A., Burst, T., Krause, W., Lehmann, P., et al. (2021). *The Manifesto Project Dataset (MARPOR), Version 2021a*. WZB Berlin Social Science Center.
- [7] Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2), 311–331.
- [8] Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3), 705–722.

- [9] Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling Policy Preferences from Coded Political Texts. *Legislative Studies Quarterly*, 36(1), 123–155.
- [10] Mikhaylov, S., Laver, M., & Benoit, K. (2012). Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis*, 20(1), 78–91.
- [11] Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297.
- [12] Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- [13] Bakker, R., de Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., et al. (2015). Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999–2010. *Party Politics*, 21(1), 143–152.
- [14] Benoit, K., & Laver, M. (2006). *Party Policy in Modern Democracies*. Routledge.
- [15] Hooghe, L., Marks, G., & Wilson, C. J. (2002). Does Left/Right Structure Party Positions on European Integration? *Comparative Political Studies*, 35(8), 965–989.
- [16] Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., et al. (2024). *V-Dem Codebook v14*. Varieties of Democracy Project.
- [17] Pemstein, D., Marquardt, K. L., Tzelgov, E., Wang, Y.-T., Krusell, J., & Miri, F. (2018). The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data. *V-Dem Working Paper 21* (3rd ed.).
- [18] King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98(1), 191–207.
- [19] Aldrich, J. H., & McKelvey, R. D. (1977). A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections. *American Political Science Review*, 71(1), 111–130.
- [20] Hare, C., Armstrong, D. A., Bakker, R., Carroll, R., & Poole, K. T. (2015). Using Bayesian Aldrich-McKelvey Scaling to Study Citizens’ Ideological Preferences and Perceptions. *American Journal of Political Science*, 59(3), 759–774.
- [21] 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.
- [22] Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv:2304.06588*.
- [23] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1), 237–291.
- [24] Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated Annotation with Generative AI Requires Validation. *arXiv:2306.00176*.
- [25] Wu, P. Y., Nagler, J., Tucker, J. A., & Messing, S. (2023). Large Language Models Can Be Used to Estimate the Latent Positions of Politicians. *arXiv:2303.12057*.
- [26] Le Mens, G., & Gallego, A. (2025). Positioning Political Texts with Large Language Models by Asking and Averaging. *Political Analysis* (arXiv:2311.16639).

- [27] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- [28] Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Aspell, A., et al. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv:2306.16388*.
- [29] Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? *PsyArXiv preprint*.
- [30] Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 11737–11762.
- [31] Rozado, D. (2024). The Political Preferences of LLMs. *PLOS ONE*, 19(7), e0306621.
- [32] Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More Human than Human: Measuring ChatGPT Political Bias. *Public Choice*, 198, 3–23.
- [33] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351.
- [34] Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4), 401–416.
- [35] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.
- [36] Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*.
- [37] Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Sage.
- [38] Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77–89.
- [39] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- [40] Lynch, M. (2016). *The New Arab Wars: Uprisings and Anarchy in the Middle East*. PublicAffairs.
- [41] Hinnebusch, R. (2001). *Syria: Revolution from Above*. Routledge.
- [42] Wickham, C. R. (2013). *The Muslim Brotherhood: Evolution of an Islamist Movement*. Princeton University Press.