Normative Verbal Reasoning Profiles From 99 Podcast Guests: A Three-Model Scoring Study Using the Expressive Cognition Rubric

A research paper from expressivecognition.org

CONFLICT OF INTEREST STATEMENT

This research was conducted by the developer of the Expressive Cognition assessment tool (expressivecognition.org), a freely accessible assessment with an optional paid report tier. This relationship is disclosed in the interest of full transparency. No external funding was received for this research.

AI USAGE STATEMENT

This study employs Claude Sonnet 4 (Anthropic), GPT-5 mini (OpenAI), and Mistral Large (Mistral AI) as automated scoring agents within a predefined behavioral rubric and a blinded multi-pass evaluation protocol. All three models function as measurement instruments: they apply the rubric to speech transcripts and generate dimension-level scores under controlled prompting conditions. Claude was also used for editorial assistance in the preparation of this manuscript. The theoretical framework, research design, analyses, and all interpretive conclusions are the work of the Expressive Cognition research program. Full responsibility for the accuracy, integrity, and originality of the manuscript rests with the project.

ABSTRACT

This study establishes normative Verbal Reasoning Index (VRI) scores for 99 guests from *Conversations with Tyler*, a long-form intellectual interview podcast, scored across six core reasoning dimensions and two moderators by three independent large language models from three different vendors (Claude Sonnet 4, GPT-5 mini, and Mistral Large) in three blinded passes each. The resulting corpus — balanced across nine disciplinary cells with 10–12 speakers each — constitutes the largest labeled spontaneous verbal reasoning dataset in healthy adults currently available. Pairwise cross-model VRI agreement averages r = .668 across the three vendors (Sonnet↔GPT-5m r = .758, Sonnet↔Mistral r = .657, GPT-5m↔Mistral r = .590), with systematic calibration offsets that form a strict generosity gradient: Mistral scores highest on average, Sonnet intermediate, GPT-5 mini lowest. Discipline rank ordering is preserved across all three models: philosophy and hard science at the top, literary arts at the bottom. Inter-pass reliability for Sonnet is excellent (mean VRI spread across three passes = 0.19). Confirmatory factor analyses estimated independently on each scorer's data reject a

unidimensional model and recover an identical two-factor structure under all three scorers — Abstraction, Compression, and Originality clustering as Generative Range, and Epistemic Calibration and Generative Self-Monitoring clustering as Calibrative Control — with factor composition stable across models and factor separation (φ = .38 Sonnet → .77 GPT-5 mini → .89 Mistral) varying systematically with each scorer's within-factor correlation pattern. Each of the three models selects a different best-fitting home for Conceptual Continuity, identifying it empirically as a boundary dimension whose factorial placement is scorer-convention-determined. The finding that three independently-prompted frontier LLMs reproduce the same disciplinary hierarchy and the same factor composition — despite having no access to speaker identity, discipline, or each other's scores — provides convergent validity evidence for the EC rubric's capacity to detect real differences in spontaneous verbal reasoning across intellectual domains.

Keywords: verbal reasoning, normative data, LLM scoring, cross-model agreement, spontaneous speech, podcast discourse, Conversations with Tyler

INTRODUCTION

The Expressive Cognition (EC) rubric scores spontaneous speech across six core dimensions of verbal reasoning — Abstraction, Compression, Originality, Conceptual Continuity, Epistemic Calibration, and Generative Self-Monitoring — plus two moderator dimensions (Vocabulary and Syntactic Control) that are reported but excluded from the composite Verbal Reasoning Index (VRI). Prior work has established construct validity for the rubric in known-groups designs using Supreme Court oral arguments and academic seminar speech (companion paper), and ecological validity in a 30-guest subset of *Conversations with Tyler* (CWT) guests correlated against external intellectual reputation (companion paper).

The present study extends this work in two directions. First, it expands the CWT sample from 30 to 99 guests balanced across nine disciplinary cells, producing normative data that allows VRI scores to be interpreted relative to a reference population of high-ability conversational speakers. Second, it introduces cross-model scoring — the same 99 transcripts scored independently by three frontier LLMs from three different vendors (Anthropic, OpenAI, and Mistral AI) — providing the first published inter-model reliability data for an LLM-applied psychometric rubric at this scale, and the first three-way factor-invariance test of the underlying construct.

METHOD

Sample

Ninety-nine guests from *Conversations with Tyler* were selected using a purposive sampling design stratified across nine disciplinary cells. A sampling script enumerated all CWT guests from the public

episode index, assigned discipline tags, and selected approximately 11 guests per cell, force-including 30 guests from the prior ecological validity study. One joint-guest episode (Noel Johnson and Mark Koyama) was excluded because the rubric assumes a single speaker. The final sample comprised 99 unique speakers.

Table 1. Sample composition by discipline cell.

CELL	N	EXAMPLE GUESTS
Philosophy	12	Agnes Callard, Slavoj Žižek, Noam Chomsky, Peter Singer
Economics	10	Daron Acemoglu, Esther Duflo (not in sample), Larry Summers
Hard Science	10	Alison Gopnik, David Deutsch, Steven Pinker, Ed Boyden
Social Science	11	Daniel Kahneman, Jonathan Haidt, Philip Tetlock
History	11	Niall Ferguson, Jill Lepore, Ada Palmer
Law/Policy	11	Cass Sunstein, Samantha Power, Jamal Greene
Lit/Arts	11	Margaret Atwood, Camille Paglia, Dana Gioia
Tech/Entrepreneurship	11	Vitalik Buterin, Sam Altman, Marc Andreessen
Journalism/Public	11	Malcolm Gladwell, Ezra Klein, Nate Silver

Transcript Extraction and Screening

For each guest, the full CWT transcript was fetched from the public CWT website and processed through a three-stage screening protocol designed to isolate spontaneous reasoning from rehearsed or recited material.

Stage 1 — Pre-filtering. Host speech was stripped; only guest turns were retained. Turns below a minimum word threshold were excluded.

Stage 2 — Spontaneity screening. Each remaining turn was evaluated by Claude Sonnet 4 for spontaneity. Turns classified as rehearsed set-pieces, recitations, memorized factual lists, or pre-drafted statements were excluded. The screener operated blind to the EC scoring rubric.

Stage 3 — Inclusion threshold. Guests were included only if their screened transcript contained ≥1,500 words across ≥8 retained turns. All 99 candidates passed this threshold. Median screened transcript length was 7,500 words (range: 1,515–11,267).

Scoring

Three scoring runs were conducted independently on the same 99 transcripts. All three used identical prompts, rubric text, dimension definitions, band descriptors, JSON output schema,

temperature (0.3), batch structure, and three-pass shuffled-blinding protocol. The only thing that varied was the scoring model itself.

Sonnet 4 scoring. Claude Sonnet 4 (claude-sonnet-4-20250514, Anthropic) scored each transcript blinded — identified only as "Speaker A," "Speaker B," etc. — using the full v3 EC behavioral descriptor rubric. Speaker labels were randomized independently on each of three scoring passes. Total scoring passes: 297 (99 guests × 3 passes). Guests were scored in shuffled batches of 6 to enable cross-guest blinding within each batch.

GPT-5 mini scoring. GPT-5 mini (OpenAI) scored the same transcripts using the same rubric and protocol. Total scoring passes: 294 (98 guests × 3 passes; one guest was excluded due to a scoring pipeline error).

Mistral Large scoring. Mistral Large (mistral-large-latest, Mistral AI) scored the same transcripts via the La Plateforme REST API using the same rubric, protocol, and schema. Total scoring passes: 297 (99 guests × 3 passes). All 99 guests produced valid scores.

The three scoring runs were completely independent: no model had access to any other's scores, and no post-hoc calibration was applied. Cross-model analyses that require a common sample use the 98 guests valid under all three scorers.

Measures

Verbal Reasoning Index (VRI). A weighted composite of six core dimensions: Abstraction (.18), Compression (.16), Originality (.16), Conceptual Continuity (.16), Epistemic Calibration (.18), and Generative Self-Monitoring (.16). Weights reflect the theoretical priority of Abstraction and Epistemic Calibration as the dimensions most closely linked to the Gf-dominant construct EC targets.

Moderator dimensions. Vocabulary and Syntactic Control are scored but excluded from the VRI composite. They capture Gc-linked linguistic competence that correlates with education and language background rather than with the reasoning construct.

Factor-structure models. To test whether the six core dimensions reflect a single general factor or a more differentiated structure, we ran confirmatory factor analyses independently on each scoring model's data. Four nested models were estimated by maximum likelihood on the 6×6 dimension-level correlation matrix (three-pass mean per dimension): a one-factor baseline (M1) and three two-factor specifications that differ in where Conceptual Continuity is assigned (M2a: on Generative Range only; M2b: on Calibrative Control only; M2c: cross-loading both). Fit was evaluated by χ^2, RMSEA, CFI, and SRMR against conventional cutoffs, with AIC used for model selection. The same nested-model comparison was run on Sonnet, GPT-5 mini, and Mistral Large data separately, yielding three independent CFA results that can be compared for structural replication. Fit computations were implemented directly (see `scripts/cwt-norms/cfa.mjs`) rather than via an external SEM package to keep the full analysis pipeline reproducible from a single repository.

RESULTS

Overall Descriptive Statistics

Table 2. Overall VRI descriptive statistics by model.

STATISTIC	SONNET 4	GPT-5 MINI	MISTRAL LARGE
n	98	98	98
Mean VRI	7.01	6.67	7.74
SD	0.45	0.34	0.44
Min	5.25	5.25	6.63
Max	7.68	7.35	8.84
Median	7.02	6.70	7.73

The three scoring models produce a strict generosity gradient: Mistral scores highest on average (mean VRI 7.74), Sonnet is intermediate (7.01), and GPT-5 mini scores lowest (6.67). The difference between the most-generous and least-generous model is more than one full scale point on VRI (1.07). Mistral and Sonnet have similar distributional spread (SD = 0.44 and 0.45 respectively), while GPT-5 mini compresses the range somewhat (SD = 0.34). Despite the calibration differences in absolute level, all three models place the lowest scorer and the highest scorers in the same guests: Ana Vidovic at the bottom on both Sonnet and GPT-5 mini, and philosophy/hard-science guests at the top on all three models.

Discipline Cell Means

Table 3. Mean VRI by discipline cell, all three models.

CELL	N	SONNET 4	GPT-5 MINI	MISTRAL LARGE
Philosophy	12	7.44	6.88	8.17
Hard Science	10	7.42	6.89	8.16
Social Science	11	7.08	6.80	7.77
Tech/Entrepreneurship	11	6.95	6.72	7.71
History	11	7.11	6.76	7.62
Journalism/Public	11	6.70	6.60	7.61
Economics	10	6.97	6.65	7.56
Law/Policy	11	6.78	6.49	7.52
Lit/Arts	11	6.61	6.27	7.50

All three models preserve the top-two and bottom-one discipline ranking: philosophy and hard science are the highest-scoring cells on every scorer, and literary arts is the lowest on every scorer. The ordering of intermediate cells shifts somewhat between scorers — Mistral places social science and tech higher than history, while Sonnet places history above both — but the differences within the middle band are small and generally within the standard error of the cell means. The discipline rank ordering of cells is substantially stable across all three scoring models even though their absolute level calibrations differ by more than a full scale point.

Dimension-Level Cell Profiles

Table 4. Sonnet 4 mean dimension scores by discipline cell.

CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN
Philosophy	8.00	7.00	7.39	7.75	7.75	6.67	8.03	6.97
Hard Science	7.70	6.70	7.53	7.87	7.90	6.73	7.97	6.97
Social Science	7.52	6.52	7.03	7.64	7.52	6.18	7.79	6.79
History	7.42	6.42	7.24	7.94	7.39	6.18	7.85	6.91
Tech/Entrepreneurship	7.24	6.27	7.36	7.48	7.12	6.15	7.39	6.39
Economics	7.27	6.33	6.80	7.43	7.57	6.30	7.53	6.50
Law/Policy	7.21	6.09	6.52	7.39	7.27	6.06	7.42	6.48
Journalism/Public	6.82	5.76	6.73	7.52	7.15	6.15	7.21	6.27
Lit/Arts	6.94	5.85	7.00	7.12	6.82	5.88	7.45	6.48

Notable discipline-specific patterns:

- **Philosophy** achieves the maximum mean Abstraction (8.00) — every philosopher in the sample operates at the "Principled" band. This is a ceiling effect consistent with the prior ecological validity study.
- **Hard Science** leads on Epistemic Calibration (7.90) and Originality (7.53), reflecting the epistemic marking and novel-framing demands of scientific discourse.
- **History** leads on Conceptual Continuity (7.94), consistent with the narrative coherence demands of historical analysis.
- **Lit/Arts** scores lowest on five of six core dimensions and lowest on VRI. This is interpreted as a construct-appropriate finding: the EC rubric measures analytical verbal reasoning, not narrative or associative reasoning. Literary discourse deploys different cognitive operations than the analytical register the rubric targets.

Cross-Model Agreement

Table 5. Pairwise cross-model agreement statistics (Pearson *r*), n = 98 common sample.

DIMENSION	SONNET↔GPT-5M	SONNET↔MISTRAL	GPT-5M↔MISTRAL
Abstraction	.643	.642	.606
Compression	.397	.611	.272
Originality	.823	.718	.726
Conceptual Continuity	.556	.460	.501
Epistemic Calibration	.666	.497	.368
Generative Self-Monitoring	.518	.355	.391
Vocabulary	.620	.640	.656
Syntax	.513	.514	.496
VRI	**.758**	**.657**	**.589**
VRI Spearman *ρ*	**.721**	**.595**	**.536**

Table 5b. Mean calibration offsets on VRI (higher-scoring model − lower-scoring model).

PAIR	OFFSET (POINTS)	DIRECTION
Sonnet − GPT-5 mini	+0.33	Sonnet more generous
Mistral − Sonnet	+0.73	Mistral more generous
Mistral − GPT-5 mini	+1.06	Mistral more generous

The three models form a strict generosity gradient on VRI: **Mistral > Sonnet > GPT-5 mini**, with a 1.06-point total spread from the most-generous to the least-generous scorer. Pairwise agreement is highest between Sonnet and GPT-5 mini (r = .758) and lowest between GPT-5 mini and Mistral (r = .589), with Sonnet↔Mistral intermediate (r = .657). The mean pairwise r across the three vendors is .668. All three pairwise correlations are substantial and well above chance, indicating that despite the absolute-level calibration differences, the three models substantially agree on the rank ordering of speakers.

At the dimension level, Originality shows the highest pairwise agreement for the Sonnet↔GPT-5m pair (r = .823), and also high agreement for the Sonnet↔Mistral and GPT-5m↔Mistral pairs (r = .718 and .726 respectively). Compression is the dimension with the most scorer-specific disagreement: Sonnet and Mistral agree moderately on Compression (r = .611), but GPT-5 mini diverges from both (r = .397 with Sonnet and r = .272 with Mistral). This suggests that GPT-5 mini operationalizes propositional density somewhat differently from the other two scorers, which may reflect prompt-interpretation sensitivity on that dimension. Epistemic Calibration and Generative Self-Monitoring also show lower pairwise agreement with Mistral than with the Sonnet↔GPT-5m pair, consistent with Mistral's broader interpretation of what counts as evidence of real-time epistemic marking and self-revision.

The Mistral–Sonnet–GPT-5 mini generosity gradient is observable on every dimension except Syntactic Control. Mistral scores substantially higher than Sonnet on Abstraction (+0.65), Compression (+0.69), Originality (+0.81), Conceptual Continuity (+0.72), and Vocabulary (+0.56), and higher than GPT-5 mini on the same dimensions by even larger margins. The largest single inter-model dimension offset is the Mistral–GPT-5 mini gap on Originality (+1.80 points), reflecting the fact that Mistral credits far more reframings as genuinely novel than GPT-5 mini does.

Inter-Pass Reliability

Table 6. Inter-pass reliability (Sonnet 4): mean spread across three passes per dimension.

DIMENSION	MEAN SPREAD	SD
Abstraction	0.09	0.29
Compression	0.15	0.44
Originality	0.18	0.39
Conceptual Continuity	0.29	0.56
Epistemic Calibration	0.29	0.50
Gen. Self-Monitoring	0.38	0.49
Vocabulary	0.15	0.36
Syntax	0.16	0.37
VRI	**0.19**	**0.23**

Inter-pass reliability is excellent. Mean VRI spread across three passes is 0.19 points — the same speaker scored three times by the same model under different blinding labels produces VRI scores that differ by less than two-tenths of a scale point on average. Abstraction is the most stable dimension (mean spread 0.09), and Generative Self-Monitoring is the least stable (0.38), consistent with the finding that GSM is more context-sensitive than other dimensions.

Factor Structure: Confirmatory Factor Analysis

To test whether the six core EC dimensions reflect a single general verbal-reasoning factor or a more differentiated structure, we ran a confirmatory factor analysis on each model's scoring data independently. Four nested models were estimated by maximum likelihood on the dimension-level correlation matrix (n = 99 guests, each with the three-pass mean for each dimension):

- **M1 (1-factor):** All six core dimensions load on a single general Verbal Reasoning Capacity factor.
- **M2a (2-factor, Cont → GR):** Generative Range (GR) = Abstraction, Compression, Originality, Conceptual Continuity; Calibrative Control (CC) = Epistemic Calibration, Generative Self-Monitoring.
- **M2b (2-factor, Cont → CC):** GR = Abstraction, Compression, Originality; CC = Conceptual Continuity, Epistemic Calibration, Generative Self-Monitoring.
- **M2c (2-factor, Cont cross-loads):** As M2a/M2b, but Conceptual Continuity freely loads on both factors.

Fit indices reported follow standard cutoffs: RMSEA < .08 acceptable, < .06 excellent; CFI > .90 acceptable, > .95 excellent; SRMR < .08 acceptable.

Table 7. Confirmatory factor analysis fit indices by scoring model.

MODEL	SONNET 4		GPT-5 MINI			MISTRAL LARGE		
χ^2 (df)	RMSEA	CFI	χ^2 (df)	RMSEA	CFI	χ^2 (df)	RMSEA	CFI
M1 (1-factor)	135.6 (15)	.286	.712	30.7 (15)	.103	.918	41.3 (15)	.134	.959
M2a (Cont → GR)	35.8 (14)	.126	.948	**18.8 (14)**	**.059**	**.975**	32.2 (14)	.115	.971
M2b (Cont → CC)	34.7 (14)	.123	.951	30.1 (14)	.108	.916	**27.9 (14)**	**.101**	**.978**
M2c (Cont cross-loads)	**21.1 (13)**	**.080**	**.981**	18.8 (13)	.067	.970	26.1 (13)	.101	.979

Best-fitting model per scorer in **bold**. Winning models by AIC: Sonnet → M2c (AIC = -4.9); GPT-5 mini → M2a (AIC = -9.2); Mistral → M2b (AIC = -0.1). Each of the three scorers selects a different best-fitting specification — and the difference between them is entirely about where Conceptual Continuity belongs in the two-factor structure. Every scorer rejects the one-factor model in the direction of a two-factor solution, but the three scorers disagree on Continuity's factorial home.

Table 8. Standardized factor loadings under each model's best-fitting CFA solution.

DIMENSION	SONNET (M2C) GR	SONNET (M2C) CC	GPT-5M (M2A) GR	GPT-5M (M2A) CC	MISTRAL (M2B) GR	MISTRAL (M2B) CC
Abstraction	.921	—	.707	—	.988	—
Compression	1.000	—	.679	—	.993	—
Originality	.680	—	.562	—	.833	—
Conceptual Continuity	.342	.357	.765	—	—	.798
Epistemic Calibration	—	.824	—	.858	—	.730
Gen. Self-Monitoring	—	1.000	—	.700	—	.742
Factor correlation (φ)	**.384**	**.766**	**.886**			

Three scorers produce three *different* best-fitting placements for Conceptual Continuity — Sonnet cross-loads it on both factors, GPT-5 mini assigns it to Generative Range only, and Mistral assigns it to Calibrative Control only. This is a striking and consistent finding: the *one* dimension whose factorial home varies across scorers is the same one under all three, and each scorer picks a different

home for it. Abstraction, Compression, and Originality cluster as Generative Range on every scorer. Epistemic Calibration and Generative Self-Monitoring cluster as Calibrative Control on every scorer. No scorer ever assigns any of these five dimensions differently.

Four findings are substantively important across the three scoring models:

1. A one-factor model is rejected by every scorer. No model is consistent with a unidimensional general-verbal-reasoning factor. Sonnet rejects M1 decisively (RMSEA = .286, CFI = .712). GPT-5 mini rejects it more mildly but still clearly (RMSEA = .103, CFI = .918; $\Delta\chi^2(1)$ M1 vs. M2a = 11.91, p = .0006). Mistral also rejects it in the direction of a two-factor solution (RMSEA = .134, CFI = .959; $\Delta\chi^2(1)$ M1 vs. M2b = 13.32, p = .0003). Under all three scorers, the two-factor structure fits significantly better than the unidimensional alternative.

2. Factor composition is invariant across scorers. All three models independently cluster Abstraction, Compression, and Originality as one factor (Generative Range) and Epistemic Calibration and Generative Self-Monitoring as the other (Calibrative Control). No scorer assigns any of these five dimensions differently. The empirical clustering of the five core dimensions is model-invariant across three independent LLMs from three different vendors.

3. Factor separation varies monotonically with scorer calibration. The factor correlation φ — how distinct GR and CC are from one another — rises monotonically from Sonnet (φ = .384, well-separated factors) through GPT-5 mini (φ = .766) to Mistral (φ = .886, nearly-merging factors). The raw correlation matrices clarify the source: the three scorers differ dramatically in their within-factor correlations. Sonnet's Abs↔Comp correlation is .921; Mistral's is **.981** (nearly rank-degenerate); GPT-5 mini's is .471. The scorers with tighter within-factor correlations produce factors that are also more correlated with one another — because when every dimension inside a factor is near-identical, any cross-factor signal dominates the residual. This is a calibration phenomenon, not a structural disagreement about what the factors are.

4. Conceptual Continuity is empirically a boundary dimension with three different best-fitting homes. Sonnet's best-fitting solution has Continuity cross-loading both factors (.342 on GR, .357 on CC). GPT-5 mini's best-fitting solution (M2a) places Continuity entirely on GR (.765), with any CC cross-loading collapsing to essentially zero (−0.014) when estimated. Mistral's best-fitting solution (M2b) places Continuity entirely on CC (.798), with no GR loading. Three frontier LLMs, three distinct homes for the same dimension. This is the strongest empirical evidence to date that Conceptual Continuity is a boundary dimension whose factorial placement is scorer-convention-determined rather than construct-determined. It is therefore excluded from both the reported Generative Range and Calibrative Control subscores in production, retaining only the five unambiguously-loading dimensions (Abstraction, Compression, Originality for GR; Epistemic Calibration, Generative Self-Monitoring for CC). The production decision is not a theoretical choice; it is what the three-way CFA comparison forces.

Individual Speaker Results

Table 9. Top 10 highest VRI guests under Sonnet 4, with matched scores from the other two models.

RANK	GUEST	CELL	SONNET	GPT-5M	MISTRAL
1	Rebecca Kukla	Philosophy	7.68	6.97	8.22
1	Ed Boyden	Hard Science	7.68	7.35	8.05
1	Michelle Dawson	Hard Science	7.68	7.13	8.73
4	Agnes Callard	Philosophy	7.63	7.03	7.62
4	Henry Farrell	Social Science	7.63	7.09	7.73
4	Cass Sunstein	Law/Policy	7.63	6.74	7.83
7	Alison Gopnik	Hard Science	7.57	7.09	8.56
7	David Deutsch	Hard Science	7.57	7.03	8.68
7	Vitalik Buterin	Tech	7.57	7.20	7.90
10	Russ Roberts	Economics	7.52	6.63	7.79

Table 10. Top 10 largest Sonnet↔GPT-5m VRI disagreements (with Mistral comparison).

GUEST	CELL	SONNET	GPT-5M	\	Δ S−G\	MISTRAL
Camille Paglia	Lit/Arts	6.67	5.57	1.11	7.94	
Dana Gioia	Lit/Arts	7.34	6.42	0.92	7.67	
Cass Sunstein	Law/Policy	7.63	6.74	0.89	7.83	
Russ Roberts	Economics	7.52	6.63	0.89	7.79	
Peter Singer	Philosophy	7.36	6.49	0.87	7.68	
Diarmaid MacCulloch	History	7.47	6.69	0.78	7.84	
Abhijit Banerjee	Economics	7.29	6.58	0.71	7.69	
Rebecca Kukla	Philosophy	7.68	6.97	0.71	8.22	
Jess Wade	Hard Science	7.07	6.37	0.70	7.84	
Marc Andreessen	Tech	7.45	6.75	0.69	8.00	

The largest Sonnet↔GPT-5 mini disagreements are asymmetric: in all 10 cases Sonnet scores higher than GPT-5 mini, and in all 10 cases Mistral scores higher than both. This pattern is consistent across the full sample — 87 of 98 guests receive higher VRI from Sonnet than from GPT-5 mini, and 96 of 98 receive higher VRI from Mistral than from GPT-5 mini. The disagreements are concentrated among speakers whose discourse style is rhetorically confident and compressed (Paglia, Sunstein, Andreessen). The consistent direction of disagreement across all three model pairs — Mistral highest, Sonnet intermediate, GPT-5 mini lowest — suggests GPT-5 mini evaluates such speakers more

strictly against the rubric's behavioral anchors, while the more generous scorers credit rhetorical confidence as evidence of underlying reasoning capacity even when the behavioral markers are ambiguous.

DISCUSSION

The Normative Contribution

This study provides the first large-scale normative dataset for the EC Verbal Reasoning Index. Ninety-eight speakers scored by three independent frontier language models across nine balanced disciplinary cells produce VRI distributions that can serve as a reference population for interpreting individual scores. Median VRI in this population is approximately 7.0 on Sonnet, 6.7 on GPT-5 mini, and 7.7 on Mistral — a population pre-selected by the same interviewer for intellectual distinction. The scorer-specific median matters because, as documented above, the three scoring models differ by more than a full scale point in absolute-level calibration.

The compressed VRI ranges (5.25–7.68 on Sonnet, 5.25–7.35 on GPT-5 mini, 6.63–8.84 on Mistral) reflect this pre-selection. The normative data should be interpreted as norms for the upper end of the ability distribution, not as population norms. A broader normative study — currently planned using Prolific recruitment with concurrent ICAR fluid reasoning assessment — will provide norms across the full ability range.

Cross-Model Agreement as Validity Evidence

The pairwise inter-model agreement reported here (mean r = .668 on VRI across three pairs, range .589 to .758) is, to our knowledge, the first published three-way cross-vendor LLM-as-judge reliability statistic for a psychometric rubric. Three models from three different vendors — Anthropic (Sonnet 4), OpenAI (GPT-5 mini), and Mistral AI (Mistral Large) — trained on different data, with different architectures, applied the same rubric to the same transcripts without any coordination, and produced scores whose rank-orderings substantially agree.

This agreement is stronger than most published inter-rater reliability estimates for human-scored performance assessments in educational measurement, even when averaged across all three independent pairs. It does not mean the models are "correct" — all three could share systematic biases inherited from shared web-scale pretraining data. But it establishes that the construct measured by the EC rubric is scorer-recoverable across vendor boundaries: the same rubric, applied by three frontier LLMs that share neither architecture nor training corpus nor vendor lineage, produces broadly the same rank ordering of speakers. This is the fundamental requirement for measurement reliability, and it holds across the three-way comparison even though the absolute-level calibration of the three scorers differs by more than a full scale point.

A Two-Factor Reasoning Structure, Stable in Composition and Variable in Separation

The confirmatory factor analyses converge on a substantively important claim: the six core EC dimensions are not a single lump. A one-factor model is rejected under all three scoring models, and all three independently recover the *same* two-factor composition — Abstraction, Compression, and Originality clustering as a Generative Range factor, and Epistemic Calibration and Generative Self-Monitoring clustering as a Calibrative Control factor. That three independently-prompted frontier LLMs, with no access to each other's scoring and no shared intermediate representations, partition the same six dimensions into the same two empirical clusters is the strongest single piece of convergent-validity evidence in this study — and the three-way replication is meaningfully stronger than a two-way one, because it essentially closes the door on "both models inherited the same partition from shared pretraining."

The theoretical interpretation is straightforward. Generative Range captures the *productive* side of verbal reasoning — the ability to operate at high levels of abstraction, to pack propositions densely, and to generate non-obvious reframings. Calibrative Control captures the *monitoring* side — the ability to mark epistemic status explicitly and to revise one's own formulations upward in real time. Both factors contribute to what the VRI composite measures, and they are moderately-to-highly correlated across all three scorers (φ = .384 Sonnet, .766 GPT-5 mini, .886 Mistral) — distinguishable but not independent, as expected for two aspects of a broader verbal-reasoning capacity.

Two findings vary systematically across scorers rather than replicating cleanly. First, the magnitude of factor separation differs substantially: Sonnet produces well-separated factors (φ = .384), GPT-5 mini produces moderately separated factors (φ = .766), and Mistral produces nearly-merging factors (φ = .886). The source is visible in the raw correlation matrices: each scorer shows a characteristic within-factor correlation pattern. Sonnet has Abs↔Comp r = .921; Mistral has .981 (nearly rank-degenerate); GPT-5 mini has .471 (the most dimension-independent of the three). The stricter the within-factor correlations, the more correlated the resulting factors also become, because near-redundant within-factor dimensions leave only cross-factor residual signal. This is a scoring-calibration phenomenon, not a structural disagreement about what the factors are — all three scorers still prefer the two-factor structure to the one-factor alternative by conventional fit criteria.

Second, Conceptual Continuity — and Continuity alone — receives three different best-fitting factorial homes across the three scorers. Sonnet's preferred model cross-loads Continuity on both factors. GPT-5 mini's preferred model places Continuity on Generative Range only. Mistral's preferred model places Continuity on Calibrative Control only. No other dimension varies in its factor assignment across scorers. We interpret this as evidence that Continuity is empirically a boundary dimension — its correlations with the other five are ambiguous enough that the "right" factor for it depends on idiosyncratic scoring calibration rather than underlying construct structure. This finding is not a defect; it is the clearest possible empirical justification for the production decision to exclude Continuity from both the reported Generative Range and Calibrative Control subscores. The reported

subscores use only the five dimensions whose factor assignment is invariant across all three independent scorers. Users of the scale should specify the scoring model just as they would specify the norming sample, because absolute-level calibration and factor-separation magnitude both vary across scorers; the rank-ordering of speakers does not.

Disciplinary Reasoning Signatures

The finding that all three models independently reproduce the same top-of-hierarchy and bottom-of-hierarchy disciplinary ranking — philosophy and hard science at the top, literary arts at the bottom — provides convergent validity evidence that the rubric detects real differences in discourse register rather than random variation. Intermediate-cell orderings shift slightly across scorers, but always within the standard error of the cell means. The dimension-level profiles are consistent with what is known about how disciplinary training shapes discourse: philosophers reason at the level of principles (Abstraction at or near the scale ceiling), scientists mark epistemic boundaries explicitly (the highest Epistemic Calibration means), historians build cumulative narrative arguments (the highest Conceptual Continuity means), and literary speakers use associative and narrative reasoning that the analytical rubric does not fully capture.

Limitations

- **The sample is pre-selected.** All speakers are guests on a single interview podcast, selected by the same host for intellectual distinction. This compresses the VRI range and limits generalizability.
- **No external criterion.** Unlike the prior ecological validity study, this study does not correlate VRI with an external measure. The disciplinary patterns are interpretable but not validated against independent criteria.
- **The cross-model calibration offsets are systematic and large.** The three scorers span a strict 1.06-point generosity gradient on VRI: Mistral > Sonnet > GPT-5 mini. Norms based on one model's scores are not interchangeable with norms based on another's. Any production use of the EC rubric must specify the scoring model, and any longitudinal comparison of a single speaker over time must use the same scorer throughout.
- **Compression shows scorer-specific disagreement.** GPT-5 mini operationalizes propositional density notably differently from Sonnet (r = .397) and Mistral (r = .272), while Sonnet and Mistral agree more substantially with each other (r = .611). This suggests the GPT-5 mini prompt interpretation on Compression may be an outlier, and the dimension may benefit from additional behavioral anchoring to align scorers.
- **Literary arts scores reflect a construct boundary.** The rubric measures analytical verbal reasoning. Speakers whose primary mode is narrative, associative, or performative will score lower

not because they are less intelligent but because the rubric is not designed to measure their kind of reasoning. This is a scope limitation, not a measurement failure.

- **Single-pass written-vs-spoken comparison is not included.** The prior ecological validity study included a written-vs-spoken pilot for three guests. The present study does not extend this comparison.
- **Conceptual Continuity's factor placement is scorer-dependent.** The three-scorer CFA comparison reveals that Conceptual Continuity has no single empirical home in the two-factor structure — each scorer places it differently. This is a limitation of the current rubric insofar as it suggests the Continuity behavioral descriptors are ambiguous with respect to the Generative Range / Calibrative Control partition. It is also the clearest empirical justification for excluding Continuity from the reported GR and CC subscores. Future rubric revisions could either sharpen the Continuity anchors to force a consistent factorial home or formally recognize it as a cross-cutting dimension that contributes to VRI without loading on either subscore.

CONCLUSION

Ninety-eight speakers from *Conversations with Tyler*, scored by three independent frontier large language models from three different vendors across six core dimensions of verbal reasoning, produce normative VRI data that is internally consistent (inter-pass reliability = 0.19), cross-model reliable (mean pairwise r = .668), discipline-differentiated in theoretically predicted directions, and factorially coherent under every scorer.

Two complementary sources of convergent validity support the EC rubric's construct claim. First, all three scoring models independently produce the same disciplinary hierarchy — philosophy and hard science at the top, literary arts at the bottom, with theoretically interpretable dimension-level profiles for each discipline — despite having no access to speaker identity, discipline, or each other's scores. Second, confirmatory factor analyses estimated independently on each of the three scorers' data reject a unidimensional model and recover an identical two-factor composition: Abstraction, Compression, and Originality cluster as Generative Range; Epistemic Calibration and Generative Self-Monitoring cluster as Calibrative Control. The factor composition is invariant across scorers. Only factor separation magnitude and the placement of one boundary dimension (Conceptual Continuity) vary, and both variations are attributable to scoring-calibration differences rather than to disagreement about what the factors are.

Together, these findings suggest that what the rubric measures is real, even if what it measures is narrower than intelligence writ large. The EC rubric measures analytical verbal reasoning as performed in spontaneous speech, and the construct is internally structured as two moderately-to-highly correlated but empirically distinguishable components — one productive, one calibrative. Within that scope, the rubric measures reliably, it differentiates meaningfully, and its underlying

factor structure is recoverable by three independent frontier LLMs from three different vendors. The three-way replication of the Generative Range / Calibrative Control partition, with a three-way disagreement confined to a single boundary dimension, constitutes the strongest evidence currently available that the two-factor structure of verbal reasoning recovered here is a property of the construct rather than an artifact of any individual scorer.

REFERENCES

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* Academic Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*(10), 906–911.

Guilford, J. P. (1967). *The nature of human intelligence.* McGraw-Hill.

Kintsch, W. (1974). *The representation of meaning in memory.* Erlbaum.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*(1), 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749.

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity.* Oxford University Press.

Simpson, R., Briggs, S., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English.* The Regents of the University of Michigan.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* Academic Press.

Vygotsky, L. S. (1962). *Thought and language.* MIT Press.

APPENDIX A: COMPLETE PER-GUEST DIMENSION SCORES

The full per-guest dimension-level scores are presented in three tables, one per scoring model, to keep each table within the page width. Scores are 3-pass blinded averages on the 1–9 scale. Abs = Abstraction; Cmp = Compression; Ori = Originality; CC = Conceptual Continuity; EC = Epistemic Calibration; GSM = Generative Self-Monitoring; Voc = Vocabulary (moderator); Syn = Syntactic Control (moderator). VRI = Verbal Reasoning Index composite weighted from the six core dimensions. Rows sorted by discipline cell, then by that scorer's VRI descending within cell.

Appendix A.1: Claude Sonnet 4 scores (n = 98)

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Russ Roberts	economics	8	7	7	8	8	7	7.7	6.7	7.52
Daron Acemoglu	economics	8	7	8	8	7	6	8	7	7.34
Abhijit Banerjee	economics	7.3	6.3	7	8	8	7	8	7	7.29
Paul Krugman	economics	7.3	6.3	6.3	7.3	8	7	7.7	6.7	7.08
Raj Chetty	economics	7	6	6.7	7.7	8	6.7	7	6.3	7.02
Alain Bertaud	economics	7	6	7	8	7	6	7	6	6.84
Larry Summers	economics	7	6	6	7	8	6.3	7.7	7	6.75
Simon Johnson	economics	7	6	6	7	7.7	6	7.3	6.3	6.64
Nassim Nicholas Taleb	economics	7	6.7	8	6.3	6.3	5.3	8	6	6.61
Alan Taylor	economics	7	6	6	7	7.7	5.7	7	6	6.59
Ed Boyden	hard_science	8	7	8	8	8	7	8	7	7.68
Michelle Dawson	hard_science	8	7	8	8	8	7	8	7	7.68
Alison Gopnik	hard_science	8	7	8	8	7.7	6.7	8	7	7.57
David Deutsch	hard_science	8	7	8	8	7.7	6.7	8	7	7.57
Steven Pinker	hard_science	8	7	7	8	8	7	8	7	7.52
Michael Nielsen	hard_science	8	7	8	7	8	6	8	7	7.36
Paul Bloom	hard_science	7.7	6.7	7.3	7.7	8	6.7	7.7	6.7	7.35
Philip Ball	hard_science	7.3	6.3	7	8	8	6.7	8	7	7.24
Atul Gawande	hard_science	7	6	7	8	8	7	8	7	7.18
Jess Wade	hard_science	7	6	7	8	7.7	6.7	8	7	7.07
Diarmaid Macculloch	history	8	7	7.3	8	8	6.3	8	7	7.47
Ada Palmer	history	8	7	8	8	7	6	8	7	7.34
Adam Tooze	history	8	7	8	8	7	6	8	7	7.34
Jill Lepore	history	8	7	8	8	7	6	8	7	7.34
Roy Foster	history	7.7	6.7	7	8	8	6.3	8	7	7.30
Helen Castor	history	7	6	7	8	8	7	8	7	7.18
Jennifer Burns	history	7	6	7	8	8	6.3	7	6.3	7.07
Paul Gillingham	history	7	6	7	8	7	6	8	7	6.84
Niall Ferguson	history	7	6	7	8	7	6	8	7	6.84
Patricia Fara	history	7	6	6.7	7.7	7.3	6	7.3	6.7	6.79
Reza Aslan	history	7	6	6.7	7.7	7	6	8	7	6.73
Ezra Klein	journalism_public	7	6	7	8	8	7	7	6.7	7.18

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Nate Silver	journalism_public	7	6	6.3	7.3	8	7	7	6	6.97
David Brooks	journalism_public	7	6	7	8	7	6	7.3	6.3	6.84
Malcolm Gladwell	journalism_public	7	6	7	8	7	6	7	6	6.84
Larissa Macfarquhar	journalism_public	7	6	7	8	7	6	8	7	6.84
Andrew Sullivan	journalism_public	7	6	7	8	7	6	8	7	6.84
Ben Thompson	journalism_public	7	6	6.3	7.3	7	6	7	6	6.63
Barkha Dutt	journalism_public	7	6	6	7	7	6	7	6	6.52
Ben Westhoff	journalism_public	6	6	7	7	7	6	7	6	6.50
Annie Jacobsen	journalism_public	7	4.7	6.3	7	7.7	6	7	6	6.48
Andrew Ross Sorkin	journalism_public	6	4.7	7	7	6	5.7	7	6	6.05
Cass Sunstein	law_policy	8	7	7.7	8	8	7	8	7	7.63
Jamal Greene	law_policy	8	7	7	8	8	6.7	8	7	7.47
Rachel Harmon	law_policy	7	6	6.3	7.3	8	6.3	7.3	6.3	6.86
Ben Sasse	law_policy	7	6	7	8	7	6	7.7	7	6.84
Bruno Macaes	law_policy	7.3	6.3	7.3	7.7	6.7	5.7	7.7	6.3	6.84
Jennifer Pahlka	law_policy	7	6	7	8	7	6	7	6	6.84
Samantha Power	law_policy	7	6	6	7	8	6.3	8	7	6.75
Stanley Mcchrystal	law_policy	7	6	6	7	7	6	7	6	6.52
Tom Tugendhat	law_policy	7	6	6	7	6.7	5.7	7	6.7	6.41
John O Brennan	law_policy	7	4.7	5.3	6.3	7.7	6	7	6	6.21
Leopoldo Lopez	law_policy	7	6	6	7	6	5	7	6	6.18
Dana Gioia	lit_arts	8	7	8	8	7	6	8	7	7.34
Margaret Atwood	lit_arts	7	6	8	7	8	7	8	7	7.18
Brian Koppelman	lit_arts	7	6	7	8	7.7	6.7	7.3	6.3	7.07
Fuchsia Dunlop	lit_arts	7	6	7	8	7	6	8	7	6.84
Alex Ross	lit_arts	7	6	7	8	7	6	8	7	6.84
Camille Paglia	lit_arts	8	7	8	6.7	5.7	4.7	8	7	6.67
Benjamin Moser	lit_arts	7	6	7	7.3	6.7	5.7	8	7	6.62
Andy Weir	lit_arts	6.3	5.3	7	7.3	6.3	6	6.7	6	6.39
Cynthia Haven	lit_arts	7	6	6	6	7	6	7	6	6.36
Emily St John Mandel	lit_arts	6	5	7	6	7	6	7	6	6.18
Ana Vidovic	lit_arts	6	4	5	6	5.7	4.7	6	5	5.25

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Rebecca Kukla	philosophy	8	7	8	8	8	7	8	7	7.68
Agnes Callard	philosophy	8	7	7.7	8	8	7	8	7	7.63
David Bentley Hart	philosophy	8	7	7	8	8	7	8.3	7.7	7.52
Elijah Millgram	philosophy	8	7	8	7	8	7	8	7	7.52
William Macaskill	philosophy	8	7	7	8	8	7	8	7	7.52
Amia Srinivasan	philosophy	8	7	7	8	8	6.7	8	7	7.47
Rabbi David Wolpe	philosophy	8	7	7.3	7.7	8	6.7	8	7	7.47
John Gray	philosophy	8	7	8	8	7.3	6.3	8	7	7.45
Kwame Anthony Appiah	philosophy	8	7	7	8	8	6.3	8	7	7.41
Peter Singer	philosophy	8	7	6	8	8	7	8	7	7.36
Noam Chomsky	philosophy	8	7	7.7	8	7	6	8	7	7.29
Slavoj Zizek	philosophy	8	7	8	6.3	6.7	6	8	6	7.01
Henry Farrell	social_science	8	7	7.7	8	8	7	8	7	7.63
Daniel Kahneman	social_science	8	7	8	7.3	8	6.3	8	7	7.47
Jonathan Haidt	social_science	8	7	7	8	8	6	8	7	7.36
Joseph Henrich	social_science	8	7	8	8	7	6	8	7	7.34
Arthur Brooks	social_science	7.7	6.7	7	8	7.7	6.7	8	7	7.29
Philip E Tetlock	social_science	7	6	7	7.7	8	6.3	8	7	7.02
Chris Blattman	social_science	7	6	6.7	7.7	8	6.7	7.7	6.7	7.02
Harvey Mansfield	social_science	8	7	7	8	6	5.3	8	7	6.89
Ashley Mears	social_science	7	6	7	7.3	7	6	7	6	6.73
Daniel Carpenter	social_science	7	6	6	7	8	6	8	7	6.70
Eric Kaufmann	social_science	7	6	6	7	7	5.7	7	6	6.47
Vitalik Buterin	tech_entrepreneurship	8	7	8	8	7.7	6.7	8	7	7.57
Marc Andreessen	tech_entrepreneurship	8	7	8	8	7	6.7	8	7	7.45
Audrey Tang	tech_entrepreneurship	8	7	8	7	7.7	6	8	7	7.30
Balaji Srinivasan	tech_entrepreneurship	8	7	8	7	7	6.3	8	7	7.23
Daniel Gross	tech_entrepreneurship	7	6	7	8	7.7	6.7	7	6	7.07
Sam Altman	tech_entrepreneurship	7	6	7	7.7	7.7	6	7	6	6.91
Brian Armstrong	tech_entrepreneurship	7	6	7	7.7	7.7	6	7	6	6.91
Chris Dixon	tech_entrepreneurship	7	6	7	8	7	6	7.3	6.3	6.84
Blake Scholl	tech_entrepreneurship	7	6	8	7	6	6	7	6	6.66

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Patrick Collison	tech_entrepreneurship	6.7	6	7	7	6	5.3	7	6	6.33
David Rubenstein	tech_entrepreneurship	6	5	6	7	7	6	7	6	6.18

Appendix A.2: GPT-5 mini scores (n = 99)

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Daron Acemoglu	economics	7.7	6.3	7	7.7	7.7	6.7	6.7	6.7	7.19
Larry Summers	economics	7.7	6	5.7	6.3	8	7.3	7	7	6.87
Raj Chetty	economics	7	5.7	6	7	7.7	7	6	6	6.75
Nassim Nicholas Taleb	economics	7.7	6	7	6	6.7	7	6.7	6	6.74
Russ Roberts	economics	7.3	5.7	6	6.7	7	7	6	6.7	6.63
Alan Taylor	economics	7.3	6	5	6.7	7.7	6.7	6.3	6.7	6.59
Abhijit Banerjee	economics	7	5.7	6	6.3	7.3	7	6	6	6.58
Paul Krugman	economics	7.3	6	6	6.3	7	6.7	6.3	6.7	6.58
Alain Bertaud	economics	7.7	5.3	6	7	6.7	6	6.7	6.7	6.47
Simon Johnson	economics	7.3	5	5.3	6.3	6.3	6	6.3	6.3	6.09
Ed Boyden	hard_science	7.7	6	7.7	7.7	8	7	7	6.7	7.35
Michelle Dawson	hard_science	7.7	6	7	7.3	7.7	7	7	6	7.13
Alison Gopnik	hard_science	8	5.7	7	7.3	7.7	6.7	7	7	7.09
David Deutsch	hard_science	8	5.3	6.7	7.3	7.7	7	7.3	6.7	7.03
Steven Pinker	hard_science	7.3	6	6	7.3	7.7	7	7.3	7.3	6.91
Paul Bloom	hard_science	7.3	6	6	7	7.7	7	6	6.3	6.86
Philip Ball	hard_science	7.3	6	6	6.7	7.7	7	6.7	6.7	6.81
Michael Nielsen	hard_science	7.7	5.7	6	6.3	7.3	7	6.3	6.3	6.70
Atul Gawande	hard_science	7.7	5.7	6	6.7	7	7	6.7	6.7	6.69
Ezekiel Emanuel	hard_science	7	5.7	6	6.7	7	6.7	6.3	6.3	6.52
Jess Wade	hard_science	7	4.3	6	6.3	7.3	7	6.7	6.7	6.37
Helen Castor	history	7.7	6	6	7.7	8	7	7	7	7.09
Adam Tooze	history	8	6.3	6.3	7	7.7	7	8	7.3	7.09
Roy Foster	history	7.7	6	6	7.3	7.7	6.7	7.3	7.3	6.92
Paul Gillingham	history	8	6.3	6	7	7	7	7.3	7	6.91
Ada Palmer	history	7.3	6	6.7	7	7.3	6.7	7.3	7.3	6.85
Jill Lepore	history	7.3	5.3	6.3	7	7.7	7	6.3	7	6.81
Diarmaid Macculloch	history	7.3	6	6.3	7	7	6.3	7.7	7	6.69
Jennifer Burns	history	7	5.3	6	7	7.3	6.3	6.3	6.3	6.53
Niall Ferguson	history	7	6	6	6.7	7	6.3	7	6.7	6.52
Patricia Fara	history	7	5.7	6	7	7	6.3	6.3	6.7	6.52
Reza Aslan	history	7.3	5.7	6	6.7	7	6	6.7	6	6.47

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Nate Silver	journalism_public	8	6	6	7	8	7	6	6	7.04
Barkha Dutt	journalism_public	7.3	6	6	7	7.7	7	6.7	6.7	6.86
Malcolm Gladwell	journalism_public	7.3	6	6	7	7	7.3	6.7	7	6.79
Ben Thompson	journalism_public	7	6.3	6	7	7.3	6	6.3	6	6.63
Ezra Klein	journalism_public	7.3	5.7	6	6.7	7.3	6.3	6	6	6.59
Andrew Sullivan	journalism_public	7.3	6	6	6.3	6.7	7	6.7	6.7	6.57
Andrew Ross Sorkin	journalism_public	6.7	5.7	6	6.7	7	6.7	6	6	6.46
David Brooks	journalism_public	7.3	4.7	6	6.3	7	7	6.3	6	6.42
Larissa Macfarquhar	journalism_public	7	4.7	6	6.7	7	7	7	7	6.41
Ben Westhoff	journalism_public	6.7	5.7	6	6.7	7	6.3	6	6	6.41
Annie Jacobsen	journalism_public	7	4.7	6	6.3	7.3	6.7	6	6.3	6.37
Jamal Greene	law_policy	7.7	6	6	7	8	7	6.7	6.3	6.98
Rachel Harmon	law_policy	7.7	6	5.7	7	7.7	6.7	6.7	6.3	6.81
Cass Sunstein	law_policy	7	6	6	7	7.3	7	6.7	6.3	6.74
Bruno Macaes	law_policy	7.3	6	6	7	7	6.3	6.3	6	6.63
Ben Sasse	law_policy	7	5.7	6	6.7	7	6.7	6	6.3	6.52
Samantha Power	law_policy	7.7	5	5.3	6.7	7	7	6.7	6.7	6.48
Jennifer Pahlka	law_policy	7	6	6	6.7	6.7	6.3	6	6	6.46
Stanley Mcchrystal	law_policy	7.3	5.7	6	6.3	6.7	6	6	6	6.36
Tom Tugendhat	law_policy	7	5.3	5.3	6.7	7	6.3	6.7	6.3	6.31
Leopoldo Lopez	law_policy	7	6	5	6.7	6	6	6	6	6.13
John O Brennan	law_policy	7	5	4.3	6	7.3	6	6.3	6	5.99
Alex Ross	lit_arts	7.7	6	6	7	7	6.7	7.7	7	6.75
Andy Weir	lit_arts	7	6	6	7	7	7	6.3	6.3	6.68
Margaret Atwood	lit_arts	7	5.7	6.7	6.3	7.3	6.7	7	7	6.63
Fuchsia Dunlop	lit_arts	7	6	6.3	7	7	6	7.3	6.7	6.57
Cynthia Haven	lit_arts	7	5.7	5.7	6.7	7.3	6.3	6.7	6	6.47
Brian Koppelman	lit_arts	7	5	6	6.7	7.3	6.7	7.3	7	6.47
Dana Gioia	lit_arts	7.7	4.7	6	7	6.7	6.3	7.3	7	6.42
Benjamin Moser	lit_arts	7	5.3	6	6	6.7	6.3	6.3	6	6.25
Emily St John Mandel	lit_arts	6	4.3	5.7	6	7	6.3	6	6.3	5.91
Camille Paglia	lit_arts	7	3.3	6.7	6	4.7	5.7	7.3	6.7	5.57

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Ana Vidovic	lit_arts	6	3.3	4	6	6	6	6	6	5.25
Agnes Callard	philosophy	7.7	6	6.7	7	7.7	7	6.7	6	7.03
Rabbi David Wolpe	philosophy	8	5.3	6.3	7	8	7	7.7	7	6.99
John Gray	philosophy	8	6	6	7	7.7	7	7.7	6.7	6.98
Amia Srinivasan	philosophy	8	6	6.3	7	7.7	6.7	7	6.7	6.98
Rebecca Kukla	philosophy	7.7	5.7	6.7	7	7.7	7	6.3	6.7	6.97
Elijah Millgram	philosophy	7.7	6	6.7	6.7	7.3	7.3	7	6.3	6.97
Noam Chomsky	philosophy	8	6.3	6.3	7.3	7	6.7	7	7	6.97
William Macaskill	philosophy	7.7	6	5.7	7	8	7	6.7	6.3	6.93
David Bentley Hart	philosophy	7.7	6	6	7	7.7	7	7.7	7.3	6.92
Kwame Anthony Appiah	philosophy	7.7	6	6	7	7.3	6.7	6.3	7	6.81
Slavoj Zizek	philosophy	7	5.3	6.7	6.7	7	6.3	6.7	6	6.52
Peter Singer	philosophy	7.7	5.7	5.3	6.7	7.3	6	6.7	6.3	6.49
Daniel Kahneman	social_science	8	6	7	7	8	7.3	6.3	6.3	7.25
Henry Farrell	social_science	8	6	6.3	7	8	7	7.3	7	7.09
Joseph Henrich	social_science	7.3	6	6.7	7	8	7	6.7	6	7.03
Daniel Carpenter	social_science	7.7	6	6	7	8	7	6.7	6.7	6.98
Philip E Tetlock	social_science	7.7	6	6.3	7	7.3	7.3	6.7	6	6.97
Jonathan Haidt	social_science	7.3	6	6	7	7.3	7	6.3	6.7	6.80
Arthur Brooks	social_science	7.7	5	6.3	7	7	7	7	6.3	6.69
Chris Blattman	social_science	7.3	5.3	6	7	7.3	6.3	6	6	6.59
Harvey Mansfield	social_science	7.7	6	6	6.3	7	6.3	7	6	6.59
Eric Kaufmann	social_science	7.3	6	5.7	6.7	7	6	6	6	6.47
Ashley Mears	social_science	7.3	5	5.7	7	7	6	6	6	6.37
Vitalik Buterin	tech_entrepreneurship	8	6	7	7	8	7	7	6.3	7.20
Audrey Tang	tech_entrepreneurship	7.7	5.7	7.3	7.3	7.3	6.3	7	7	6.97
Balaji Srinivasan	tech_entrepreneurship	7.7	6	7	6.7	7	7	7	6.3	6.91
Daniel Gross	tech_entrepreneurship	7	6	6	7	8	7	6	6.3	6.86
Chris Dixon	tech_entrepreneurship	7.3	6	6	7	7.7	6.7	7.3	6.3	6.81
Blake Scholl	tech_entrepreneurship	7.3	6	7	7	6.7	6.7	6	6	6.79
Marc Andreessen	tech_entrepreneurship	7.3	6	6	6.7	7.7	6.7	6.3	6	6.75
Sam Altman	tech_entrepreneurship	7.3	6	6	6.7	7.3	7	6	6.3	6.75

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Brian Armstrong	tech_entrepreneurship	7	5.7	6	6.7	7.7	6.7	6	6.3	6.64
Patrick Collison	tech_entrepreneurship	7	6	6	6.3	6.7	6	6	6	6.35
David Rubenstein	tech_entrepreneurship	7	4.7	4.3	6	7	6	6	6	5.88

Appendix A.3: Mistral Large scores (n = 99)

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Daron Acemoglu	economics	8	7	8	9	8	8	8	7	8.00
Russ Roberts	economics	8	7	8	8	8	7.7	8	7	7.79
Nassim Nicholas Taleb	economics	8	7	8.3	8	7.7	7.3	9	7	7.73
Alain Bertaud	economics	8	7.3	8	8.3	7.3	7.3	8	7	7.72
Abhijit Banerjee	economics	8	7	7.3	8	8.3	7.3	8	7	7.69
Paul Krugman	economics	8	7	7.7	8	8	7	8	7	7.63
Larry Summers	economics	8	7.3	7	8	8	7	8	7	7.57
Raj Chetty	economics	8	7	7	8	8	7	8	7	7.52
Simon Johnson	economics	7.7	6.7	7	8	7.7	7	8	7	7.35
Alan Taylor	economics	7	6	6	7.3	7	6.3	7.3	6.7	6.63
Michelle Dawson	hard_science	9	8	9	9	9	8.3	9	8	8.73
David Deutsch	hard_science	9	8	9	9	9	8	9	8	8.68
Alison Gopnik	hard_science	9	8	9	9	8.3	8	8.7	7.7	8.56
Steven Pinker	hard_science	9	8	8	9	8.7	7.7	9	8	8.41
Ed Boyden	hard_science	8	7	9	8.3	8	8	8	7	8.05
Philip Ball	hard_science	8	7	8	8.3	8.7	8	8	7	8.01
Michael Nielsen	hard_science	8	7	8	8	8.3	8	8	7	7.90
Paul Bloom	hard_science	8	7	8	8	8	8	8	7	7.84
Jess Wade	hard_science	8	7	8	8.7	8	7.3	8.3	7.3	7.84
Ezekiel Emanuel	hard_science	8	7	8	8	7.3	7	8	7	7.56
Atul Gawande	hard_science	8	7	8	8	7	7.3	8	7	7.55
Helen Castor	history	8	7	8	9	8	8	9	8	8.00
Ada Palmer	history	8	7	8.3	8.7	8	7.7	9	8	7.95
Roy Foster	history	8	7	8	9	8	7.7	9	8	7.95
Reza Aslan	history	8	7	8	9	8	7.3	9	8	7.89
Diarmaid Macculloch	history	8	7	8	9	8	7	9	8	7.84
Adam Tooze	history	8	7	8	8	8	7	9	8	7.68
Paul Gillingham	history	8	7	8	8.3	7.7	7	8.3	7.3	7.67
Jill Lepore	history	8	7	8	8	7	7	8	7	7.50
Niall Ferguson	history	8	7	8	8	7	7	8.7	7.7	7.50
Jennifer Burns	history	7	6	7	8	7	7	7	7	7.00
Patricia Fara	history	7	6	7	8	7	6	8	7	6.84

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Nate Silver	journalism_public	8	7	8	8	8	8	8	7	7.84
Larissa Macfarquhar	journalism_public	8	7	8	8.3	8	7.7	8	7	7.84
Ezra Klein	journalism_public	8	7	8	8	8.3	7.3	8	7	7.79
Andrew Sullivan	journalism_public	8	7	8	8	7.7	8	8.3	7.3	7.78
Annie Jacobsen	journalism_public	8	7	8	8.7	7.7	7.3	8.3	7	7.78
Barkha Dutt	journalism_public	8	7	7.3	8.3	8	7.3	8.3	7.3	7.68
Malcolm Gladwell	journalism_public	8	7	8	8	7	8	8	7	7.66
Ben Thompson	journalism_public	8	7	7.7	8.3	7.7	7	8	7	7.62
David Brooks	journalism_public	8	7	8	8	7.3	7.3	8	7	7.61
Andrew Ross Sorkin	journalism_public	7	6	7	8	7	7.3	7.7	7	7.05
Ben Westhoff	journalism_public	7	6	7	8	7	7	7	7	7.00
Cass Sunstein	law_policy	8	7	8	8.3	7.7	8	8	7	7.83
Jennifer Pahlka	law_policy	8	7	8	8.7	7.3	8	8	7	7.83
Leopoldo Lopez	law_policy	8	7	8	8.3	8	7.3	8	7	7.79
Samantha Power	law_policy	8	7	8	8	8	7.3	8	7	7.73
Ben Sasse	law_policy	8	7	8	8.3	7.7	7.3	8	7	7.73
Bruno Macaes	law_policy	8	7	8	8	7	7.3	8	7	7.55
Jamal Greene	law_policy	8	7	7	8	8	7	8	7	7.52
Rachel Harmon	law_policy	8	7	7	8	8	7	7.7	7	7.52
Tom Tugendhat	law_policy	7.3	6.7	7	8	7.7	7	8	7.3	7.29
Stanley Mcchrystal	law_policy	7.3	6.3	7	8	7.3	7	7	6.7	7.17
John O Brennan	law_policy	7	6	6	7	7.3	7	8	7	6.74
Brian Koppelman	lit_arts	8	7	8	8.7	8	8	8	7	7.95
Camille Paglia	lit_arts	8.3	7.3	9	8.3	7.3	7.3	9	8	7.94
Andy Weir	lit_arts	8	7	8	8.7	7.7	7.3	8	7	7.78
Margaret Atwood	lit_arts	8	7	8	8	8	7	8.3	7.3	7.68
Dana Gioia	lit_arts	8	7	8	8.3	7.7	7	8.7	7.7	7.67
Alex Ross	lit_arts	8	7	8	8	7.7	7	8.7	7.3	7.62
Benjamin Moser	lit_arts	8	7	8	8	7	7	8.3	7.3	7.50
Cynthia Haven	lit_arts	8	7	8	8	7	7	8	7	7.50
Fuchsia Dunlop	lit_arts	7.3	6.3	7.3	8.3	7.3	7	8.3	7	7.28
Emily St John Mandel	lit_arts	7	6	7	7.7	7	7	7	7	6.95

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Ana Vidovic	lit_arts	7	6	6	7	7	7	7	7	6.68
Elijah Millgram	philosophy	9	8	9	9	9	9	9	8	8.84
John Gray	philosophy	9	8	9	9	9	8	9	8	8.68
David Bentley Hart	philosophy	9	8	8.7	9	9	8	9	8	8.63
Amia Srinivasan	philosophy	9	8	9	9	8.3	8	9	8	8.56
Noam Chomsky	philosophy	9	8	9	9	8	8	9	8	8.50
Rebecca Kukla	philosophy	8.7	7.7	8.7	8.7	7.7	8	8.7	7.7	8.22
Kwame Anthony Appiah	philosophy	8	7	8	8.3	9	8	8.3	7.3	8.07
William Macaskill	philosophy	8	7	8	8.3	8	8	8	7	7.89
Rabbi David Wolpe	philosophy	8	7	8	8	8	7	8	7	7.68
Peter Singer	philosophy	8	7	7.3	8.3	8	7.3	8	7	7.68
Slavoj Zizek	philosophy	8	7	8.3	8	7	7.7	8.7	7	7.66
Agnes Callard	philosophy	8	7	8	8	7.7	7	8	7	7.62
Daniel Kahneman	social_science	8.3	7.3	8.3	8.3	9	8	8.3	7.3	8.24
Arthur Brooks	social_science	8.7	7.7	8	9	8	7.7	9	8	8.17
Harvey Mansfield	social_science	8.3	7.3	8	9	7.3	8	8.3	7.3	7.99
Philip E Tetlock	social_science	8	7	8	8.7	8.3	7.7	8	7	7.95
Daniel Carpenter	social_science	8	7	7.7	8	8.3	7.7	8.3	7	7.79
Jonathan Haidt	social_science	8	7	8	8.3	8	7	8	7	7.73
Henry Farrell	social_science	8	7	8	8	7.7	7.7	8.3	7.3	7.73
Chris Blattman	social_science	8	7	7.3	8	8.3	7.3	8	7	7.69
Joseph Henrich	social_science	8	7	8	8.3	7.3	7	8	7	7.61
Ashley Mears	social_science	7.7	6.7	7.7	8	7	7	8	7	7.33
Eric Kaufmann	social_science	7.3	6.3	7	8	7.3	7	7.7	6.7	7.17
Audrey Tang	tech_entrepreneurship	8.3	7.7	8.7	8.7	8	8.3	9	8	8.27
Blake Scholl	tech_entrepreneurship	8	7.7	9	9	8	8	8	7	8.27
Marc Andreessen	tech_entrepreneurship	8	7	8	9	8	8	8	7	8.00
Vitalik Buterin	tech_entrepreneurship	8.3	7.3	8	8.3	8	7.3	8.3	7.3	7.90
Sam Altman	tech_entrepreneurship	8	7	8	8.3	7.7	7.3	8	7	7.73
Patrick Collison	tech_entrepreneurship	8	7	8	8.7	7.7	7	8	7	7.73
Chris Dixon	tech_entrepreneurship	8	7	8	8.3	7.3	7	8	7	7.61
Daniel Gross	tech_entrepreneurship	8	7	8	8	7	7.3	8	7	7.55

GUEST	CELL	ABS	CMP	ORI	CC	EC	GSM	VOC	SYN	VRI
Balaji Srinivasan	tech_entrepreneurship	8	7	8	8	7	7	8	7	7.50
Brian Armstrong	tech_entrepreneurship	8	7	8	8	7	7	7.3	7	7.50
David Rubenstein	tech_entrepreneurship	7	6	6.3	7.3	7	7	7.7	7	6.79