# Internet Appendix to "Measuring Monetary Policy Uncertainty Through an Agentic AI Framework: The Virtual Federal Reserve"

**Chi Wang**
G. Brint Ryan College of Business, University of North Texas*

September 2026

This Internet Appendix reports supplementary tables for the paper. Table, figure, equation, and section numbers without the "IA" prefix refer to the main text.

## IA.I The Information Ladder

Tables IA.1–IA.4 document the M0–M4 information ladder behind the objection probabilities: model performance across the rungs, the persona rung by persona source, the dissent-class PR-AUC deltas along the ladder, and the frontier-versus-local comparison discussed in Section 8.3 of the paper.

*Address: 1155 Union Circle #305339, Denton, TX 76203. Email: Chi.Wang@unt.edu. An Internet Appendix with supplementary tables (Tables IA.1–IA.12) is available at http://www.chiwangfinance.com/asset/paper/essay1_internet_appendix.pdf.

Table IA.1: **Model Performance Across the M0–M4 Stacking Ladder.** Each column is the same simulation given one more piece of information than the column before it, so the change from one column to the next isolates what that piece adds. Panel A measures how well each layer identifies the rare dissent; Panel B reports the two indices the paper constructs; Panel C the name ablation; Panel D the benchmarks the ladder is judged against. Sample: 240 scored meetings, 2,517 voter-meeting cells, a 4.49% dissent base rate. Each panel is summarized in the note below; every number is defined and discussed in the text (Section 6).

| **Metric** | **M0** random | **M1** +economy | **M2** +name | **M3** +evidence | **M4** +debate |
|---|---|---|---|---|---|
| *Panel A: How well the layer spots dissents (higher is better)* | | | | | |
| PR-AUC | 0.045 | 0.050 | 0.080 | **0.144** | 0.174 |
| Dissent recall | 1.000 | 0.938 | 0.212 | 0.513 | 0.513 |
| Dissent F1 | 0.086 | 0.091 | 0.153 | 0.272 | 0.337 |
| Balanced accuracy | 0.500 | 0.528 | 0.569 | 0.704 | 0.721 |
| *Panel B: The two indices the paper builds* | | | | | |
| Agentic MPU ($\text{MPU}_{\text{dissent}}$, inferred dissent rate) | — | — | — | 0.161 | — |
| Suppressed-dissent gap | — | — | — | 0.116 | — |
| Mean-prob. MPU (robustness twin) | — | — | — | 0.357 | — |
| Herding $\theta$ (0 ignore, 1 follow) | — | — | — | — | 0.529 |
| Policy-rate gap (pct. points) | — | — | — | — | −0.020 |
| *Panel C: Robustness (dissent PR-AUC, full panel)* | | | | | |
| M3′: evidence kept, name hidden | | | 0.068 | | |
| *Panel D: Benchmarks (dissent PR-AUC)* | | | | | |
| Always-agree (never calls a dissent) | | | 0.045 | | |
| Logistic (stance + macro regression) | | | 0.168 | | |
| Member fixed effects (own past dissents) | | | 0.315 | | |
| $N$ meetings / voter-meetings | | | 240 / 2,517 | | |

Columns are the five ladder rungs, each adding one input to the one before: M0 random; M1 macro briefing (identity-blind); M2 adds the member's name; M3 adds date-gated dossier evidence (the persona rung the Agentic AI MPU is built from); M4 adds a chair-first deliberation round. Panel A: dissent-class detection metrics (base rate 0.045; higher is better). Panel B: the two constructed indices. Panel C: name ablation of M3. Panel D: the four baselines. All metrics are dissent-class; each number is defined and discussed in the text (Section 6).

Table IA.2: **M3 Persona Rung by Persona Source.** Cells are split by the resolved persona source (Section 3.7): speech for a live composite speech signal, dossier for the dated-episode gap-fill, inferred for the vote-history fallback, none for the empty persona. Dissent-class metrics. The dossier tier carries 103 of the panel's 114 dissents and the speech tier, though larger, only 8; the tiers scored here hold 113 in all, the remaining dissent falling in a cell dropped for missing inputs (the full-panel dossier channel carries 104).

| Persona source | $n$ | Dissents | PR-AUC | F1 | Recall |
|---|---|---|---|---|---|
| speech | 1226 | 8 | 0.015 | 0.043 | 0.375 |
| dossier | 1197 | 103 | 0.206 | 0.357 | 0.553 |
| inferred | 70 | 2 | 0.083 | 0.125 | 1.000 |
| none | 24 | 0 | (no dissents) | | |

Table IA.3: **Channel Decomposition of the Persona Ladder.** Change in mean objection probability at each rung contrast on the headline (calibrated) run. M2−M1 is the memorized-reputation channel, M3−M2 the marginal contribution of date-gated dossier retrieval over reputation, M3−M3′ the name channel, and M4−M3 the deliberation (herding) channel.

| Channel (rung contrast) | Δ mean objection |
|---|---|
| M2−M1 (memorized reputation) | -0.0386 |
| M3−M2 (dossier RAG over reputation) | +0.0177 |
| M3−M3′ (name channel) | -0.0046 |
| M4−M3 (deliberation / herding) | -0.0437 |

Table IA.4: **Reasoning-Capacity Robustness: Local 7B vs. Frontier Reasoner (DeepSeek V4-Flash).** Meeting-level correlation of the two index-bearing rungs; the post-assumed-cutoff column is the contamination sensitivity split (the vendor publishes no training cutoff; June 2024 is an estimate). The frontier series is the average of four independent API runs.

| Metric | All meetings ($n = 240$) | Pre-cutoff ($n = 227$) | Post-assumed-cutoff ($n = 13$) |
|---|---|---|---|
| $\text{MPU}_{\text{dissent}}$ (M3) correlation | 0.42 | 0.39 | 0.55 |
| Mean-prob. MPU (robustness twin) | 0.44 | 0.40 | 0.69 |
| PRG (M4) correlation | 0.76 | 0.76 | 0.44[†] |
| $\text{MPU}_{\text{dissent}}$ level: 7B / V4-Flash | 0.16 / 0.12 | (frontier trims the inferred dissent rate) | |
| Mean-prob. MPU level: 7B / V4-Flash | 0.36 / 0.20 | (frontier trims the objection floor) | |
| PRG level: 7B / V4-Flash | −0.02 / −0.05 | (near-identical) | |

[†] Post-cutoff PRG correlation is noisy ($n = 13$). Each frontier run cost approximately $16 (M3 + M4, one draw per cell); the reported index averages four runs.

## IA.II Robustness and Contamination Audit

Table IA.5 collects the robustness checks and the hard contamination checks of the pre-registered protocol.

Table IA.5: **Robustness and Contamination Audit.** Three panels collect the paper's robustness and contamination checks. Panel A: dissent-class PR-AUC on the pre- and post-cutoff windows, split at the model's empirically bracketed training cutoff; on the memorization-exposed name-only rung the holdout is not below the training window, contrary to what memorization would produce, and the evidence rung's noisier split is shown alongside. Panel B: the memorization probe (name and date only), whose vote recall stays below the trivial always-agree rule and whose reconstructed dissenter set is almost entirely wrong. Panel C: the eight pre-registered validator checks, all passing, plus the M1 sampling-band diagnostic.

| **Robustness / audit check** | **Result** | **Detail** |
|---|---|---|
| *Panel A: Out-of-sample holdout (dissent PR-AUC, name-only rung M2, cutoff 2023-12-01)* | | |
| Pre-cutoff ($n$ = 2,338, 103 dissents) | 0.086 | training-window meetings |
| Post-cutoff holdout ($n$ = 203, 11 dissents) | 0.143 | holdout not below pre-cutoff |
| Evidence rung M3, pre / post | 0.158 / 0.074 | post-window noisy (11 dissents) |
| *Panel B: Memorization probe (name + date only, all 2,541 cells)* | | |
| Vote recall, full / best decade | 37.4% / 58.2% | below the 95.5% always-agree rule |
| Vote recall, pre / post cutoff | 40.2% / 5.4% | default-NO on unfamiliar cells |
| Dissenter precision / recall | 0.054 / 0.075 | 1,577 fabricated dissenter names |
| *Panel C: Mechanical validator gate (pre-registered, every scored run)* | | |
| Date gate; evidence date gate | PASS | no input dated on/after meeting $T$ |
| No outcome stats; no district leak | PASS | no dissent rate or district in prompt |
| Prompt hash; identity-blind rungs | PASS | sha256 matches; no name in M1/M3′ |
| Framework behaves; indep. recompute | PASS | 72.1% persona-direction; metrics reproduced |
| M1 sampling-band diagnostic | clean | band $> 0.4$ on 0/240 meetings |

## IA.III Replication of the FOMC-Cycle Premium

Table IA.6 replicates Cieslak et al. (2019), Table 1, on their sample and extends it; the construction underlies Tables 5 and 6 of the paper.

Table IA.6: **Replication of Cieslak et al. (2019), Table 1.** Daily excess stock returns on FOMC-cycle dummies. Column (1) is the published estimate; column (2) the replication on the same 1994–2013 sample with the construction used in Table 5; columns (3) and (4) extend the sample.

| | (1) Published 1994–2013 | (2) Replication 1994–2013 | (3) Extended 1994–2016 | (4) This paper 1996–2025 |
|---|---|---|---|---|
| *Panel A: Week 0 and weeks 2, 4, 6* | | | | |
| Week 0 | 0.136*** | 0.1362*** | 0.1411*** | 0.0751* |
| | (2.75) | (2.76) | (3.17) | (1.78) |
| Weeks 2, 4, 6 | 0.0995*** | 0.0994*** | 0.1094*** | 0.0680** |
| | (2.65) | (2.65) | (3.25) | (2.29) |
| Constant | −0.021 | −0.0209 | −0.0245 | 0.0037 |
| | (-0.96) | (-0.96) | (-1.24) | (0.20) |
| *Panel B: Separate even-week dummies* | | | | |
| Week 0 | 0.136*** | 0.1362*** | 0.1411*** | 0.0751* |
| | (2.75) | (2.76) | (3.17) | (1.78) |
| Week 2 | 0.081* | 0.0810* | 0.0895** | 0.0546 |
| | (1.70) | (1.70) | (2.10) | (1.44) |
| Week 4 | 0.108** | 0.1076** | 0.1203** | 0.0726* |
| | (2.00) | (1.99) | (2.52) | (1.82) |
| Week 6 | 0.178** | 0.1778** | 0.1875** | 0.1521* |
| | (1.99) | (1.99) | (2.08) | (1.76) |
| Constant | −0.021 | −0.0209 | −0.0245 | 0.0037 |
| | (-0.96) | (-0.96) | (-1.24) | (0.20) |
| Observations (days) | 5,214 | 5,214 | 5,997 | 7,825 |

Dependent variable: daily excess return on the U.S. stock market over the one-month T-bill rate, in percent per day (Fama–French market factor), on weekdays in FOMC-cycle time with holiday returns set to zero; odd weeks are the omitted category. Column (1) reproduces the published estimates of Cieslak et al. (2019), Table 1 (working-paper version, 1994–2013); columns (2)–(4) are estimated here with the same construction on the stated samples. Heteroskedasticity-robust $t$-statistics in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

## IA.IV Voting Panel and Herding

Table IA.7 summarizes the audited voting panel and its persona-source composition; Tables IA.8 and IA.9 report the herding measurements of the M4 rung.

Table IA.7: **Audited Voting Panel and Persona-Source Composition.** The audited panel spans 240 meetings; persona source gives the share of voter–meeting cells resolved to a live speech signal, a dated-evidence dossier, a vote-history inference, or an empty persona. Every dossier claim carries a source URL, an archived snapshot, and a publication date.

| Statistic | Value |
|---|---|
| FOMC meetings (1996–2025) | 240 |
| Directive votes | 2,541 |
| Dissents | 114 |
| Dissent base rate | 4.49% |
| Voter–meeting cells scored | 2,517 |
| Dual-coded dossiers | 74 |
| Persona source (voter–meeting cells) | |
| speech | 48.7% |
| dossier | 47.6% |
| inferred | 2.8% |
| none | 1.0% |
| Mean persona stance | −0.002 |
| Std. persona stance | 0.211 |

Table IA.8: **Herd Pressure by Agentic AI MPU Quartile.** The 240 scored meetings (1996–2025) are partitioned into quartiles of the Agentic AI MPU ($\text{MPU}_{\text{dissent}}$, local 7B arm). Official dissent is the recorded NO-vote share; the suppressed-dissent gap is the mean $\text{MPU}_{\text{dissent}}$ minus the official dissent rate; the last column gives the mean-probability twin over the same meetings. Official dissent stays near the base rate across quartiles while the suppressed-dissent gap widens with the MPU, so latent disagreement is increasingly masked where it is highest.

| **Quartile** | $N$ | **$\text{MPU}_{\text{dissent}}$ range** | **Mean $\text{MPU}_{\text{dissent}}$** | **Official dissent** | **Supp.-dissent gap** | **Mean-prob. MPU** |
|---|---|---|---|---|---|---|
| Q1 (Low) | 60 | [0.000, 0.091] | 0.035 | 2.6% | +0.009 | 0.298 |
| Q2 | 60 | [0.091, 0.100] | 0.098 | 3.8% | +0.060 | 0.332 |
| Q3 | 60 | [0.100, 0.200] | 0.182 | 4.4% | +0.138 | 0.370 |
| Q4 (High) | 60 | [0.200, 0.800] | 0.330 | 7.3% | +0.256 | 0.430 |
| Full sample | 240 | [0.000, 0.800] | 0.161 | 4.5% | +0.116 | 0.357 |

Table IA.9: **Herding: Full-Panel Suppression and Deliberation $\theta$.** The upper block is the full-panel gap between the inferred dissent rate (the MPU index) and the recorded dissent rate, with the mean pre-herding objection probability as a robustness row. The lower block is the chair-first deliberation estimate: the empirical herding coefficient, the calibration-based benchmark of 0.80 for contrast, and the mean objection before (M3) and after (M4) the deliberation round.

| Herding statistic | Value |
|---|---|
| M3 (persona) $\rightarrow$ recorded vote (full panel, 7B arm) | |
| $\text{MPU}_{\text{dissent}}$ (inferred dissenters / voters; $\tau = 0.473$) | 0.161 |
| Official dissent rate | 0.045 |
| Suppressed-dissent gap | 0.116 |
| Robustness: mean objection probability (mean-prob. MPU) | 0.357 |
| Robustness: gap on the mean-prob. MPU | 0.312 |
| M4 chair-first deliberation (full panel) | |
| Empirical herding $\theta$ | 0.529 |
| Calibration-based benchmark $\theta$ | 0.80 |
| Mean objection, M3 (pre) | 0.357 |
| Mean objection, M4 (post) | 0.314 |

## IA.V Capability Gradient at the M3R Rung

Table IA.10 reports latent rate dispersion across the four-reasoner capability gradient at the M3R rung.

Table IA.10: **Latent Rate Dispersion Across the Capability Gradient (M3R Rung).** Within-meeting population standard deviation of independently elicited preferred rates (the M3R rung: the persona prompt with the directive anchor removed and a preferred rate asked instead), computed over 240 meetings for the three arms that ran the rung. Paired mean increases: V4-Flash − 7B = +0.86bp ($t = 2.29$); Sonnet-5 − V4-Flash = +0.53bp ($t = 1.19$); Sonnet-5 − 7B = +1.38bp ($t = 3.40$). Pairwise correlations of the dispersion series across arms are low (0.11–0.24): the arms agree on the level, not the meeting-by-meeting pattern, so dispersion is model-specific and is reported as a capability diagnostic rather than a measure. The Sonnet-5 column carries the memorization caveat of Section 8.3; the clean capability comparison is 7B vs. V4-Flash.

| | 7B | V4-Flash | Sonnet-5 |
|---|---|---|---|
| Mean dispersion (bp) | 7.6 | 8.4 | 9.0 |
| Median dispersion (bp) | 7.4 | 8.3 | 8.3 |
| Correlation with own-arm $\text{MPU}_{\text{dissent}}$ | 0.18 | 0.28 | 0.46 |
| Correlation with own-arm mean-prob. MPU | 0.20 | 0.36 | 0.56 |
| Correlation with official dissent rate | 0.04 | 0.06 | 0.37 |

## IA.VI Pre-Meeting Economic Briefings

Table IA.11 reports the source coverage of the date-gated pre-meeting briefings across the 240 meetings; Table IA.12 reproduces one meeting's condensed briefing in full.

Table IA.11: Briefing source coverage, 240 FOMC meetings, 1996–2025.

| Briefing source | Meetings | Coverage |
|---|---|---|
| Beige Book + Tealbook/Greenbook | 200 | 1996–2019 (staff forecast, 5-yr FRASER lag) |
| Beige Book + SEP macro | 20 | 2012–2025 quarterly (SEP macro only, dots excluded) |
| Beige Book only | 20 | 2020–2025 non-SEP (realized-macro fallback) |
| Total | 240 | all 1996–2025 meetings |

Table IA.12: Example condensed pre-meeting briefing (FOMC meeting of March 20, 2019).

| Field | Content (FOMC meeting 2019-03-20) |
|---|---|
| National conditions | Economic activity continued to expand in late January and February; ten Districts reported slight-to-moderate growth while Philadelphia and St. Louis were flat. About half noted the government shutdown slowed some sectors. Consumer spending mixed; manufacturing strengthened but contacts cited weaker global demand, tariffs, and trade uncertainty. Employment rose amid tight labor markets and worker shortages; wages moderately higher. Prices increased modest-to-moderate, with input costs outpacing selling prices. Staff estimates current-quarter GDP growth of about 1%. |
| Forward outlook | Staff: Q1 GDP 1% rebounding to 2.6% Q2, 1.8% for 2019 then 2.0% (2020); unemployment bottoming 3.6% late 2019; core PCE 1.9%, reaching 2% H2 2019. SEP 2019 central tendencies: real GDP 1.9-2.2%, unemployment 3.6-3.8%, PCE 1.8-1.9%, core PCE 1.9-2.0%. |
| District (Dallas) | Expanded moderately with slight demand pickup in manufacturing, services, housing. Drilling dipped. Hiring moderate, employment outlooks bullish. Input-price pressures moderated but wage pressures elevated. Outlooks more optimistic. |

## References


Apel, M. and Blix Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series*, No. 92.

Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4):1593–1636.

Bauer, M. D. and Swanson, E. T. (2023). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37:87–155.

Bernanke, B. S. (1983). Irreversibility, uncertainty, and cyclical investment. *Quarterly Journal of Economics*, 98(1):85–106.

Bernanke, B. S. and Kuttner, K. N. (2005). What explains the stock market's reaction to Federal Reserve policy? *Journal of Finance*, 60(3):1221–1257.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–945.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3):623–685.

Chappell, H. W., McGregor, R. R., and Vermilyea, T. (2005). *Committee Decisions on Monetary Policy*. MIT Press, Cambridge, MA.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82.

Cieslak, A., Morse, A., and Vissing-Jørgensen, A. (2019). Stock returns over the FOMC cycle. *Journal of Finance*, 74(5):2201–2248.

Ehrmann, M. and Fratzscher, M. (2007). Communication by central bank committee members: Different strategies, same effectiveness? *Journal of Money, Credit and Banking*, 39(2–3):509–541.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Gürkaynak, R. S., Sack, B., and Swanson, E. T. (2005). Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1(1):55–93.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, 133(2):801–870.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper* No. 31122.

Husted, L., Rogers, J., and Sun, B. (2020). Monetary policy uncertainty. *Journal of Monetary Economics*, 115:20–36.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working Paper*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Lucca, D. O. and Trebbi, F. (2009). Measuring central bank communication: An automated approach with application to FOMC statements. *NBER Working Paper* No. 15367.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Malmendier, U., Nagel, S., and Yan, Z. (2021). The making of hawks and doves. *Journal of Monetary Economics*, 117:19–42.

Meade, E. E. and Sheets, D. N. (2005). Regional influences on FOMC voting patterns. *Journal of Money, Credit and Banking*, 37(4):661–677.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST 2023*.

Riboni, A. and Ruge-Murcia, F. J. (2010). Monetary policy by committee: Consensus, chairman dominance, or simple majority? *Quarterly Journal of Economics*, 125(1):363–416.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2019). Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper* 2017-01.

Swanson, E. T. (2021). Measuring the effects of federal funds rate changes, forward guidance, and asset purchases on financial markets. *Journal of Monetary Economics*, 118:32–53.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214.

Clarida, R., Galí, J., and Gertler, M. (2000). Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics*, 115(1):147–180.

Judd, J. P. and Rudebusch, G. D. (1998). Taylor's rule and the Fed: 1970–1997. *Federal Reserve Bank of San Francisco Economic Review*, 1998(3):3–16.

Thornton, D. L. and Wheelock, D. C. (2014). Making sense of dissents: A history of FOMC dissents. *Federal Reserve Bank of St. Louis Review*, 96(3):213–227.

Hansen, A. L. and Kazinnik, S. (2023). Can ChatGPT decipher Fedspeak? *SSRN Working Paper*.

Bobrov, A. E., Kamdar, R., Paulson, C. M., Poduri, A., and Ulate, M. (2025). Do local economic conditions influence FOMC votes? *FRBSF Economic Letter*, 2025(13):1–5.

Bordo, M. and Istrefi, K. (2023). Perceived FOMC: The making of hawks, doves and swingers. *Journal of Monetary Economics*, 136:125–143.

Seok, S., Wen, S., Yang, Q., Feng, J., and Yang, W. (2025). MiniFed: LLMs-based agentic-workflow for simulating FOMC meetings. *PACIS 2025 Proceedings*, 21.

Kazinnik, S. and Sinclair, T. M. (2026). FOMC in silico: A multi-agent system for monetary policy decision modeling. *Working Paper*.

Istrefi, K. (2019). In Fed watchers' eyes: Hawks, doves and monetary policy. *Banque de France Working Paper* No. 725.

Zhu, Y., Yang, S., Zhu, J., and Jiang, J. (2026). Agentic framework for political biography extraction. *Working Paper* (SSRN).

Meade, E. E. (2005). The FOMC: Preferences, voting, and consensus. *Federal Reserve Bank of St. Louis Review*, 87(2):93–101.

Blinder, A. S. and Morgan, J. (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit and Banking*, 37(5):789–811.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Brown, T. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Qwen Team (2025). Qwen2.5 technical report. arXiv:2412.15115.

DeepSeek-AI (2024). DeepSeek-V3 technical report. arXiv:2412.19437.

DeepSeek-AI (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948.

DeepSeek-AI (2026). DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv:2606.19348.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv:2411.10109.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. (2023). Quantifying memorization across neural language models. *International Conference on Learning Representations*.