MathBode: Understanding LLM Reasoning with Dynamical Systems

Charles L. Wang
Department of Computer Science
Columbia University
charles.w@columbia.edu

Abstract

We present **MathBode**, a *dynamic diagnostic* for mathematical reasoning in large language models (LLMs). Instead of one-shot accuracy, MathBode treats each parametric problem as a system: we drive a single parameter sinusoidally and fit first-harmonic responses of model outputs and exact solutions. This yields interpretable, frequency-resolved metrics—*gain* (amplitude tracking) and *phase* (lag)—that form Bode-style fingerprints. Across five closed-form families (linear solve, ratio/saturation, compound interest, 2×2 linear systems, similar triangles), the diagnostic surfaces systematic *low-pass* behavior and growing phase lag that accuracy alone obscures. We compare several models against a symbolic baseline that calibrates the instrument ($G \approx 1$, $\phi \approx 0$). Results separate frontier from mid-tier models on dynamics, providing a compact, reproducible protocol that complements standard benchmarks with actionable measurements of reasoning fidelity and consistency. We open-source the dataset and code to enable further research and adoption. Code | Dataset

1 Introduction

Large language models (LLMs) now score highly on math benchmarks, but final–answer accuracy obscures *how* they reason and whether behavior is stable under controlled changes. We propose a *dynamic* evaluation: treat each parametric problem as a system, drive one parameter sinusoidally, and summarize the model's response by *gain* (amplitude tracking) and *phase* (lag) over frequency. **MathBode** implements this across five closed-form families, fitting first-harmonic responses to produce Bode-style fingerprints that reveal low-pass behavior and growing phase lag even when static accuracy ties. The protocol is simple (short prompts, deterministic decoding) and includes a symbolic baseline to calibrate the instrument (ideal $G\approx1$, $\phi\approx0$). We report $G(\omega)$, $|\phi(\omega)|$, mid-band aggregates, residual autocorrelation, and first-harmonic fit quality (R^2), providing a complementary lens on reasoning fidelity, consistency, and prompt sensitivity that accuracy alone cannot capture.

Context. Progress in mathematical reasoning is typically reported on static, final-answer datasets such as GSM8K and MATH, with domain-tuned systems (e.g., Minerva) pushing scores higher [Cobbe et al., 2021, Hendrycks et al., 2021, Lewkowycz et al., 2022]. Newer suites emphasize expert difficulty and recency—OlympiadBench, Omni-MATH, FrontierMath—yet still follow the one-input/one-answer paradigm [He et al., 2024, Gao et al., 2024, Glazer et al., 2024]. A parallel thread probes robustness: small semantic edits can flip answers (SVAMP; MATH-Perturb), while sampling strategies like self-consistency improve end accuracy without *measuring* stability [Patel et al., 2021, Huang et al., 2025, Wang et al., 2022]. Meta-reasoning probes and repeated-trial consistency likewise show models can be correct once yet unreliable across paraphrases or restarts [Zeng et al., 2023]. Together, these observations motivate metrics that capture reliability and invariance, not just correctness.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

Why a frequency/phase view? Interpretability results suggest a principled bridge to the frequency domain: transformers trained on arithmetic learn sinusoidal/rotational internal codes; modular addition emerges via Fourier-like features and rotations; recent work describes clock-like number embeddings and trigonometric operations [Nanda et al., 2023, Kantamneni and Tegmark, 2025, Li et al., 2024]. If numeric reasoning is expressed in amplitude and phase, then frequency-response style probing is natural rather than metaphorical.

What MathBode measures. For each family, we generate a parameter trajectory $p_t = p_0 + \epsilon \sin(\omega t)$, decode a single numeric line with temperature 0, and fit $\{1, \sin(\omega t), \cos(\omega t)\}$ to both ground truth and model outputs. From the fitted coefficients we recover amplitude and phase and compute $G(\omega) = \mathrm{amp}(\hat{y})/\mathrm{amp}(y^*)$ and $\phi(\omega) = \mathrm{wrap}\big(\phi(\hat{y}) - \phi(y^*)\big)$. We sweep $\omega \in \{1, 2, 4, 8, 16\}$ (64 steps), optionally vary start phase to assess phase stability, and include a symbolic baseline that realizes the ideal response. The resulting frequency-resolved curves and aggregates expose amplitude fidelity, timing lag, and prompt-surface sensitivity—even when static accuracy saturates or training-data familiarity blurs the line between recall and robust computation. Deterministic decoding and strict numeric parsing ensure we compare numeric sequences, not templates. A generic pattern or echo policy would typically yield incorrect amplitude/timing (non-unity G, shifted ϕ) and elevated residual autocorrelation, even if surface formatting looked consistent.

2 Benchmark

Instrument. We probe *dynamic* mathematical reasoning by driving one problem parameter with a sinusoid and fitting first-harmonic responses of model outputs against exact solutions. For a sweep of length T and angular frequency ω we instantiate prompts with

$$p_t = p_0 + \epsilon \sin(\omega t + \phi_0), \quad t = 1, \ldots, T,$$

decode deterministically (temperature 0) to a single numeric line (`FINAL: <number>`), and parse the model series $\hat{y}_t$ alongside the exact series y_t^*. Each series is regressed onto $\{\sin(\omega t), \cos(\omega t), 1\}$; from the fitted coefficients (a, b, c) we recover amplitude and phase

$$\mathrm{amp}(y) = \sqrt{a^2 + b^2}, \qquad \phi(y) = \mathrm{atan2}(b, a).$$

We then report

$$G(\omega) = \frac{\mathrm{amp}(\hat{y})}{\mathrm{amp}(y^*)}, \qquad \phi(\omega) = \mathrm{wrap}_{(-\pi,\pi]}\big(\phi(\hat{y}) - \phi(y^*)\big),$$

along with first-harmonic R^2 (fit quality), residual RMS (normalized), residual ACF(1), and a nonlinearity proxy H_2/H_1 from a joint fit at ω and 2ω. A symbolic solver baseline runs through the identical pipeline, providing the ideal reference ($G \approx 1$, $\phi \approx 0$).

Although gain and phase originate in linear systems, we do *not* assume linear time–invariant behavior. The sinusoid is used purely as a controlled probe: we project both exact and model series onto the first harmonic to summarize amplitude fidelity (gain) and timing (phase), while residual diagnostics and H_2/H_1 explicitly capture departures from a single-tone (e.g., nonlinearity and memory). Mechanistic findings of sinusoidal/rotational number codes [Nanda et al., 2023, Kantamneni and Tegmark, 2025, Li et al., 2024] motivate this descriptive frequency lens rather than a modeling assumption.

Families. We evaluate five closed-form families with fixed domains and three question variants each: *Linear Solve* ($a{=}p$: solve x in $ax{+}b{=}c$), *Ratio Saturation* ($p/(p{+}k)$), *Exponential Interest* ($A(1{+}p)^t$), *Linear System* (solve x in a $2{\times}2$ system with $a{=}p$), and *Similar Triangles* (scaling $s' = s\,p$). Families expose $(p_{\text{range}}, p_0, \epsilon)$ via code, and inputs are clipped in-range.

Frequency grid and phases. We choose $T{=}64$ and sweep $\Omega = \{1, 2, 4, 8, 16\}$ cycles per 64 steps. To assess phase robustness we use start phases $\{0^\circ, 120^\circ, 240^\circ\}$. Defaults set ϵ to roughly 10% of the family's half-range.

All experiments use temperature 0 (deterministic decoding) and strict numeric parsing with compliance filtering. At $\omega{=}16$ (16 cycles over $T{=}64$), the drive approaches the Nyquist limit; small dips in R^2 or phase swings can include aliasing artefacts, so we emphasize the mid-band $\{4, 8\}$ region for ranking.

Why this design? Gain and phase isolate amplitude tracking and lag—two core behaviors that final-answer accuracy obscures—while R^2 and residual diagnostics validate the first-harmonic approximation and expose structure left unexplained by it. The frequency grid (with tri-phase repeats) yields stability bands rather than single-shot outcomes, and the symbolic baseline calibrates the measurement end-to-end. The result is an inexpensive, reproducible instrument that complements static accuracy with a frequency-domain lens on reasoning fidelity and consistency.

3 Dataset Details

Cardinality. MATHBODE contains **9,408 rows per family** and **47,040 rows total** across five families.

Table 1: **Dataset rows by family.**

Family	Rows
Exponential Interest	9,408
Linear Solve	9,408
Linear System	9,408
Ratio Saturation	9,408
Similar Triangles	9,408
Total	**47,040**

Attribute	Type	Description
`family`	string	One of {*exponential_interest*, *linear_solve*, *linear_system*, *ratio_saturation*, *similar_triangles*}.
`question_id`	int	Variant index within a family.
`signal_type`	string	Drive label: {*sinusoid*, *chirp*, *step*}.
`amplitude_scale`	float	Relative amplitude (e.g., 0.5, 1.0, 2.5).
`frequency_cycles`	float	Frequency label (cycles per 64 steps).
`phase_deg`	float	Start phase (degrees).
`time_step`	int	Index within the rendered sequence.
`p_value`	float	Concrete parameter value used to render the prompt.
`prompt`	string	Fully-rendered natural-language question for the instance.
`ground_truth`	float	Exact numerical answer.

4 Evaluation

Scores. For each family and frequency we compute $G(\omega) = \mathrm{amp}(\hat{y})/\mathrm{amp}(y^*)$ and $\phi(\omega) = \mathrm{wrap}\big(\phi(\hat{y}) - \phi(y^*)\big)$ from the first-harmonic fit. **MB-Core** aggregates mid-band $\{4, 8\}$ deviations via a normalized combination of $|G-1|$ and $|\phi|$ across families. **MB-Plus** applies multiplicative down-weights derived from first-harmonic R^2, residual RMS/ACF(1), and H_2/H_1, penalizing responses that are poorly explained or exhibit nonlinear distortion. (Implementation details and ranges are in code; the same normalization is used for all models.)

Why these views? Final-answer accuracy hides *how* a model tracks controlled variation. We therefore summarize each family's response along four complementary axes: **(i) gain** (amplitude tracking), **(ii) phase error** (timing/lag), **(iii) residual autocorrelation** ACF(1) (leftover temporal structure not captured by the first harmonic), and **(iv) first-harmonic fit quality** R^2. Together these expose low-pass behavior, timing slippage, and prompt-surface sensitivity even when accuracy ties. Additional diagnostics (H2/H1 nonlinearity, compliance, phase-stability across start phases) appear in the appendix.

Takeaway (Gain). Most models are *low-pass*: gain declines with frequency in *Linear Solve* and *Exponential Interest*; *Similar Triangles* stays near $G{\approx}1$ (instrument check). *Linear System* amplifies between-model differences.

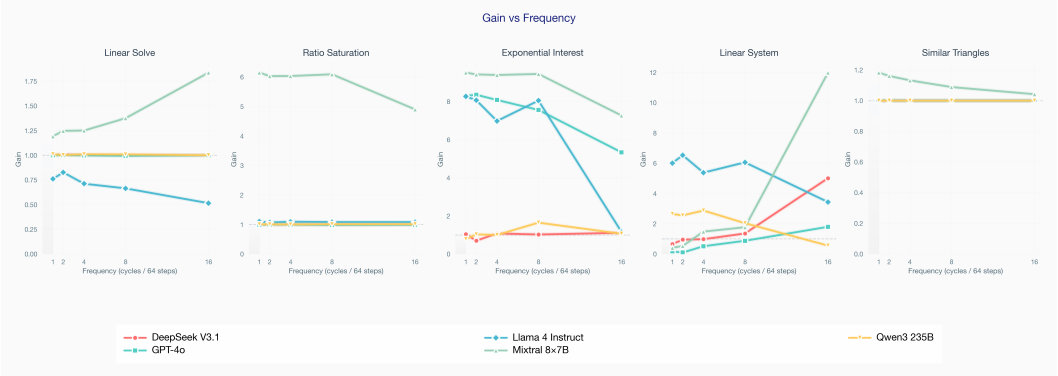

Figure 1: **Gain vs. frequency.** Panels are families; curves overlay models (unity $G{=}1$ dashed). Mid-band ({4,8}) deviations indicate under/over-reaction despite identical ground truth.

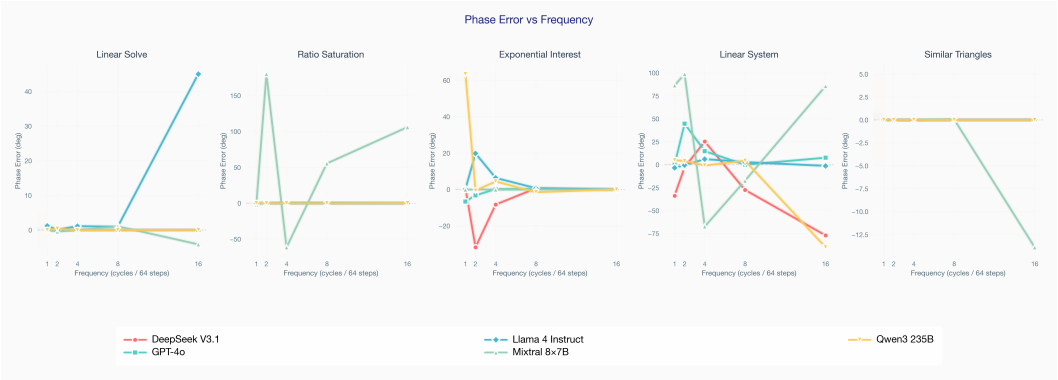

Figure 2: **Phase error vs. frequency.** Signed model–truth phase (rad), wrapped to $(-\pi, \pi]$; $0°$ implies perfect timing.

Takeaway (Phase). Phase lag typically grows with frequency (delayed tracking). Closed-form proportional families (e.g., *Similar Triangles*) remain near $0°$; *Linear System* shows the largest swings (coupling sensitivity).

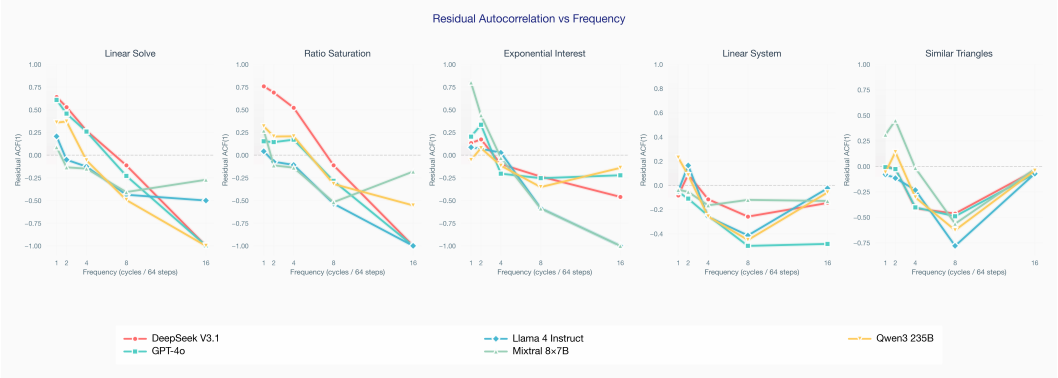

Figure 3: **Residual ACF(1) vs. frequency.** Near-zero ACF(1) means little temporal structure remains after the harmonic fit; negative values align with alternating over/undershoots at higher frequencies.

Takeaway (Residuals). Residual ACF(1) trends toward 0 or negative with frequency, indicating the first harmonic explains most structure and that remaining errors alternate rather than drift. Residual RMS and H2/H1 curves are provided in the appendix.

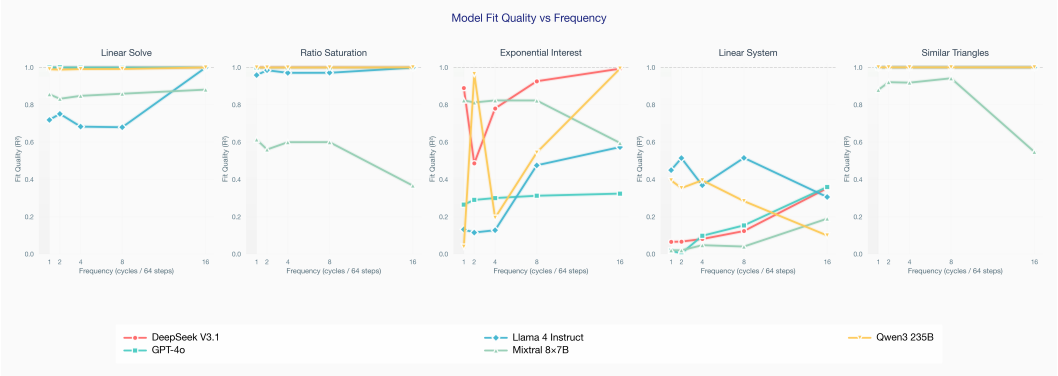

Figure 4: **First-harmonic fit quality (R^2) vs. frequency.** High R^2 validates a single-sinusoid description; dips signal nonlinear distortion or prompt-surface effects.

Takeaway (R^2). R^2 is near 1 for *Similar Triangles* and in the mid-band elsewhere; drops in *Exponential Interest* and *Linear System* co-locate with the largest gain/phase deviations, pointing to emergent nonlinearities rather than random noise.

Table 2: Overall MathBode scores. MB-Core aggregates mid-band gain/phase deviations; MB-Plus additionally downweights responses with poor fit quality (R^2), high residual structure (RMS/ACF), or nonlinearity (H_2/H_1). DeepSeek V3.1 leads overall on both MB-Core and MB-Plus.

Model	MB-Core	MB-Plus
DeepSeek V3.1	**0.834**	**0.656**
Qwen3 235B Instruct	0.782	0.576
GPT-4o	0.778	0.566
Llama 4 Instruct	0.644	0.433
Mixtral 8×7B	0.360	0.281

Table 3: Per-family MB-Core (mean mid-band performance).

Model	Exponential Interest	Linear Solve	Linear System	Ratio Saturation	Similar Triangles
DeepSeek V3.1	**0.848**	**0.995**	0.331	**0.997**	**1.000**
GPT-4o	0.497	0.993	0.418	0.980	**1.000**
Llama 4 Instruct	0.461	0.489	0.450	0.821	**1.000**
Mixtral 8×7B	0.500	0.494	0.029	0.000	0.779
Qwen3 235B Instruct	0.467	0.982	**0.471**	0.990	**1.000**

5 Conclusion.

MathBode reframes mathematical evaluation as a dynamic, frequency–domain probe, yielding interpretable gain/phase curves rather than only final answers, moving evaluations towards more reliable mathematical reasoning. Across five closed-form families, models consistently exhibit low-pass behavior and growing phase lag, while the symbolic baseline and our MB-Core/MB-Plus scores summarize these dynamics in a comparable and robust way. The results indicate that strong static accuracy can mask systematic amplitude and timing errors that degrade stability and consistency of reasoning. Practically, the frequency fingerprints provide a compact diagnostic for model selection and ablation studies, complementing standard benchmarks with measurements that are reproducible and easy to interpret. We release the dataset and reference code to support transparent replication and extension. Our use of a sinusoidal drive is an analytical probe rather than an LTI assumption; MB-Core captures mid-band amplitude/timing fidelity, while MB-Plus incorporates explicit penalties for unexplained structure and nonlinearity. Limitations include the small number of families and single-tone drives; future work will expand the task set, add richer inputs (chirps, steps), and link frequency fingerprints to internal mechanisms (e.g., attention dynamics, layer-wise delays).

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Łukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024. URL https://arxiv.org/abs/2410.07985.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024. URL https://arxiv.org/abs/2411.04872.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025. URL https://arxiv.org/abs/2502.06453.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025. URL https://arxiv.org/abs/2502.00873.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Junlin Li, Zhan Sun, Jiahao Ma, Qipeng He, Qizhe Huang, Huanzhang Xu, and Yan Li. Mechanistic interpretability of binary and ternary modular addition in transformers. *arXiv preprint arXiv:2405.17703*, 2024. URL https://arxiv.org/abs/2405.17703.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. URL https://aclanthology.org/2021.naacl-main.168/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*, 2023. URL https://arxiv.org/abs/2312.17080.

A Appendix

B Presets

Table 4: **Inference presets.** Tri-phase indicates whether phases {0,120,240} are used.

Preset	Frequencies	Phases	Tri-phase coverage	K (base keys/family)
SMOKE	{4, 8}	{0}	none	2
MVP	{4, 8, 16}	{0}	none	2
MVP_PLUS	{1, 2, 4, 8, 16}	{0}*	only for {4, 8}	2
FULL	{1, 2, 4, 8, 16}	{0, 120, 240}	all frequencies	2

Note* In **MVP_PLUS**, phases {0,120,240} are applied only at mid-band frequencies {4, 8}; other frequencies use phase {0}.

C Answer Format & Strict Parsing

Models output `[answer_start] X.YYYYYY [answer_end]` where the payload is a fixed-precision decimal with exactly six places.

Parsing. From the raw response we (i) find the *last complete* `[answer_start] ... [answer_end]` pair, (ii) scan inside for decimal literals (ASCII digits only; no scientific notation, separators, or units), (iii) take the *last* literal found, and (iv) *truncate* to exactly six decimals (pad with zeros if fewer; cut off if more). Non-finite values (NaN/Inf) or missing tags are non-compliant.

Compliance. Rows that pass this pipeline count as compliant; only compliant rows are used for harmonic fitting and residual diagnostics. Non-compliant rows still contribute to compliance statistics.

D Figures & Tables

Table 5: **A.1 Mean $|G-1|$ at mid-frequencies (4 & 8 cycles).** *Lower is better. EI and LS dominate amplitude error; DeepSeek is best on EI gain, while Mixtral is worst on RS.*

	DeepSeek V3.1	GPT-4o	Llama 4 Instruct	Mixtral 8×7B	Qwen3 235B
Exponential Interest	0.051	6.819	6.512	8.418	0.323
Linear Solve	0.002	0.003	0.312	0.313	0.009
Linear System	0.188	0.308	4.714	0.622	1.453
Ratio Saturation	0.002	0.010	0.087	5.059	0.005
Similar Triangles	0.000	0.000	0.000	0.110	0.000

Implications. Mid-band amplitude fidelity matters for stability: *EI* exposes large magnitude distortions in GPT-4o/Llama/Mixtral, so downstream pipelines that depend on accurate scaling (e.g., compounding, normalization, controller gains) will drift unless corrected. DeepSeek's best-in-class EI gain suggests safer use when amplitude tracking dominates, whereas Mixtral's large RS error flags sensitivity to saturating transforms. Family-level selection thus changes which model is "best" for a given deployment.

Table 6: **A.2 Mean |Phase Error| (deg) at mid-frequencies (4 & 8 cycles).** *Lower is better. LS is the timing bottleneck (largest lags/leads); Qwen is best on LS, while Mixtral collapses on RS.*

	DeepSeek V3.1	GPT-4o	Llama 4 Instruct	Mixtral 8×7B	Qwen3 235B
Exponential Interest	4.47	0.24	3.54	0.04	2.97
Linear Solve	0.02	0.01	1.02	0.56	0.03
Linear System	26.38	7.38	4.49	42.40	2.61
Ratio Saturation	0.01	0.01	0.38	58.42	0.01
Similar Triangles	0.00	0.00	0.00	0.05	0.00

Implications. Phase governs *timing consistency*: large LS phase errors (Mixtral, DeepSeek) imply lag/lead that can destabilize iterative procedures (solvers, rollouts) and corrupt ablations that assume time alignment. Qwen's low LS phase is attractive for timing-sensitive use cases even if its gain is not always best. When choosing models for pipelines with feedback or chaining, prioritize low phase on the relevant family.

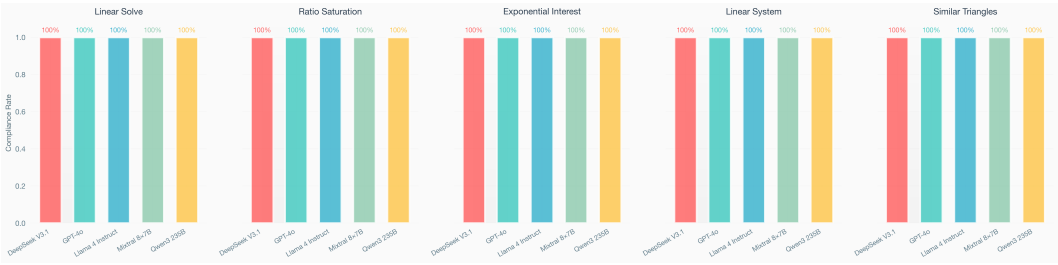

Figure 5: **A3. Compliance by family.** Compliance is perfect overall.

Implications. Near-perfect compliance removes formatting as a confound: observed dynamics (gain/phase/residuals) reflect model behavior rather than parse failures. This also means MB-Plus penalties primarily capture quality, not I/O brittleness, and reproductions should match our curves given the same row IDs.

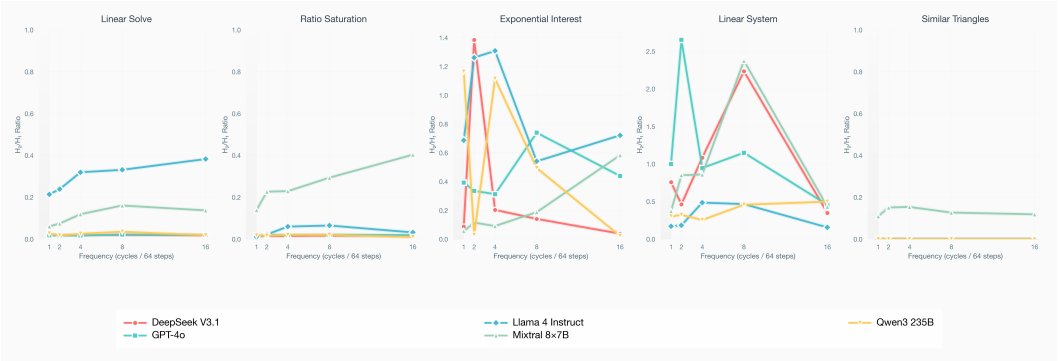

Figure 6: **A4.** H_2/H_1 **vs. frequency.** Nonlinearity concentrates in EI and LS; Similar Triangles stays near zero.

Implications. Elevated H_2/H_1 indicates distortion rather than pure linear gain/phase behavior. Peaks in EI/LS suggest that prompts with compounding or coupled relations will exhibit waveform deformation under parameter sweeps—use multi-tone tests or chirps to separate memory effects from static nonlinearity, and avoid using single-sinusoid fingerprints alone to claim linearity.

Implications. High residuals mean a first-harmonic model is insufficient: EI/LS retain structure after removing the main tone, so downstream diagnostics should include richer inputs (chirps, steps,

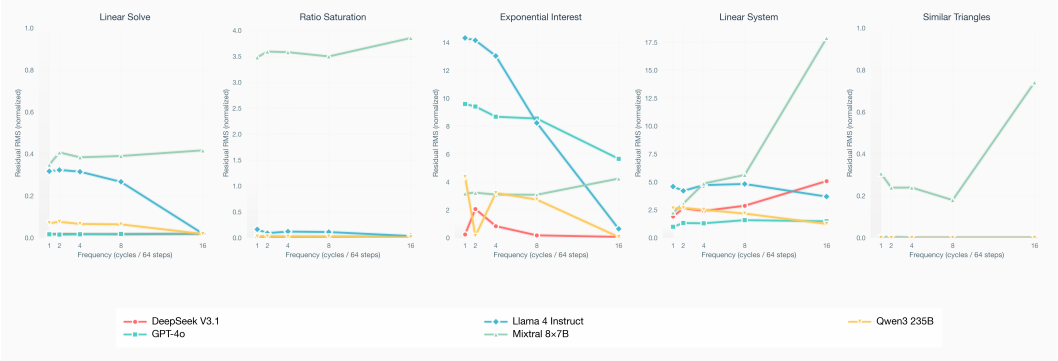

Figure 7: **A5. Residual RMS (normalized).** Single-sinusoid fits leave the largest residuals in EI and LS; simpler families fit tightly.

two-tone mixtures) before attributing errors solely to amplitude or timing. Low residuals on simpler families justify using mid-band summaries (MB-Core/MB-Plus) as compact, reliable proxies there.

E API Settings

For all model calls (Together and OpenAI), we used the following fixed decoding settings:

- **Temperature:** 0.0
- **Max tokens:** 1028

To ensure stable throughput and reproducibility, we applied simple rate limiters:

- **Together:** 600 requests per minute (RPM)
- **OpenAI:** 20,000 tokens per minute (TPM)

These settings were held constant across all experiments unless explicitly noted elsewhere.