PilotBench: A Benchmark for General Aviation Agents with Safety Constraints

1st Yalun Wu*
School of Computing
National University of Singapore
China
aaaronwww@outlook.com

2nd Haotian Liu*
School of Informatics
Xiamen University
China
liuhaotian@stu.xmu.edu.cn

3rd Zhoujun Li
CESS
Beihang University
China
lizj@buaa.edu.cn

4th Boyang Wang†
CESS
Beihang University
China
wangboyang@buaa.edu.cn

***Abstract*—As Large Language Models (LLMs) advance toward embodied AI agents operating in physical environments, a fundamental question emerges: can models trained on text corpora reliably reason about complex physics while adhering to safety constraints? We address this through PILOTBENCH, a benchmark evaluating LLMs on safety-critical flight trajectory and attitude prediction. Built from 708 real-world general aviation trajectories spanning nine operationally distinct flight phases with synchronized 34-channel telemetry, PILOTBENCH systematically probes the intersection of semantic understanding and physics-governed prediction through comparative analysis of LLMs and traditional forecasters. We introduce PILOT-SCORE, a composite metric balancing 60% regression accuracy with 40% instruction adherence and safety compliance. Comparative evaluation across 41 models uncovers a Precision-Controllability Dichotomy: traditional forecasters achieve superior MAE of 7.01 but lack semantic reasoning capabilities, while LLMs gain controllability with 86–89% instruction-following at the cost of 11–14 MAE precision. Phase-stratified analysis further exposes a Dynamic Complexity Gap—LLM performance degrades sharply in high-workload phases such as Climb and Approach, suggesting brittle implicit physics models. These empirical discoveries motivate hybrid architectures combining LLMs' symbolic reasoning with specialized forecasters' numerical precision. PILOTBENCH provides a rigorous foundation for advancing embodied AI in safety-constrained domains.**

***Index Terms*—Large Language Models, Embodied AI, Safety-Critical Systems, Aviation Agents, Flight Trajectory Prediction, Benchmark Evaluation**

I. INTRODUCTION

The rise of Large Language Models (LLMs) has sparked ambitious visions of autonomous AI agents operating in physical environments—from robotic manipulation [1] to autonomous driving [2]. Yet a fundamental question remains unanswered: *Can language models, trained primarily on text corpora, reliably reason about the complex physics governing real-world systems while adhering to strict safety constraints?*

Aviation presents an ideal testbed for this question. Flight trajectory prediction demands not only numerical accuracy but also semantic understanding of instructions, phase-aware reasoning across diverse flight regimes such as takeoff, cruise, and approach, and unwavering compliance with certified safety envelopes. Recent work demonstrates that incorporating Air Traffic Control (ATC) instructions can improve trajectory forecasting [3], [4], suggesting LLMs' potential for aviation tasks. However, existing benchmarks either focus solely on regression accuracy [5] or evaluate general reasoning without

Fixed-Wing Aircraft

Navigation & Positioning

Parameter	Value	Unit
UTC Date	2024/12/4	—
UTC Time	01:03:08	—
Latitude (WGS84)	40.3881607	°
Longitude (WGS84)	115.9206848	°
GPS Altitude	2347	ft
Barometric Altitude	735	ft
Pressure Altitude	2136	ft
GPS Ground Track	311	° (true)
Magnetic Heading	313.6	°
Magnetic Variation	7.7	°
Position Note	Good GPS Fix	—

Attitude & Flight Dynamics

Parameter	Value	Unit
Roll Angle	-1.45	°
Pitch Angle	9.79	°
Turn Rate	0.42	°/sec
Slip/Skid	-0.24	—
Vertical Speed (Baro)	899	fpm
Indicated Airspeed	72.8	kt
Normal Acceleration	1.218	G
Lateral Acceleration	-0.034	G
Attitude Status Code	280017FF N98 N100	—
Attitude Deviation Index	121	—
Flight Note	Stable Cruise	—

System & Environmental

Parameter	Value	Unit
Supply Voltage	13.5	V
Battery Status Code	00000004	—
Battery Charge	100	%
Internal Temperature	15	°C
Barometric Setting	28.44	inHg
GPS HDOP	0.8	—
GPS VDOP	1.2	—
GPS Satellite Count	10	—
GPS Fix Status	3DDiff	—
GPS Ground Speed	67.1	kt
System Note	Power Normal	—

Fig. 1. Synchronized flight-state snapshot from PILOTBENCH during cruise.

*Equal contribution. †Corresponding author.

[1]https://github.com/haotian-io/PilotBench-A-Benchmark-for-General-Aviation-Agents-with-Safety-Constraints

physical constraints [6], leaving a critical gap: *how do LLMs perform when language understanding meets physics-governed prediction in safety-critical settings?*

We present PILOTBENCH, a benchmark explicitly designed to probe this intersection through empirical comparative analysis of 41 LLMs and traditional forecasting baselines on 708 real general-aviation trajectories spanning nine flight phases. We uncover a fundamental **Precision-Controllability Dichotomy**: traditional models such as FlightPatchNet achieve superior MAE of 7.01 but lack semantic reasoning, while LLMs gain 86–89% instruction-following at the cost of 11–14 MAE precision. Our phase-stratified analysis further reveals a **Dynamic Complexity Gap**—LLMs maintain acceptable performance in steady-state phases but degrade sharply in high-dynamic regimes like climb and approach, exposing brittleness in their implicit physics models.

We formalize these findings through PILOT-SCORE, a composite metric balancing 60% regression accuracy with 40% instruction adherence and safety compliance, enabling holistic evaluation of aviation agents.

Our contributions: (1) PILOTBENCH, a benchmark with 708 phase-annotated trajectories and 34-channel telemetry; (2) PILOT-SCORE, a safety-aware composite metric; and (3) empirical evidence of the Precision-Controllability Dichotomy and Dynamic Complexity Gap, motivating hybrid architectures for aviation AI.

Dataset and code are available at[1].

II. RELATED WORK

Embodied AI [1], [7] requires agents to ground reasoning in physical interaction. While LLMs show promise for robotic control [8] and navigation, their capacity for physics-governed prediction under safety constraints remains underexplored. In flight prediction, architectures have evolved from RNNs/LSTMs to transformer-based models such as FlightBERT++ [9] and CNN–Bi-LSTM hybrids [10], [11], with multimodal ATC integration reducing errors by over 20% [4]. However, these methods treat prediction as homogeneous regression, overlooking phase differentiation and certified flight envelopes [12]. On the benchmark side, datasets like OpenSky [13] and TartanAviation [14] provide rich telemetry, while autonomous driving benchmarks such as LaMPilot [2] evaluate safety-aware agents—yet aviation lacks a dedicated LLM benchmark with safety constraints. General benchmarks like AGIEval [6] and ToolBench [15] also lack physical constraints. We address these gaps with PILOTBENCH, evaluating LLMs across nine FAA-aligned flight phases, and PILOT-SCORE, quantifying adherence to certified flight envelopes.

III. PILOTBENCH

The benchmark is designed to evaluate LLMs within aviation safety-critical scenarios, emphasizing structured flight segmentation, rigorous quality control, and detailed statistical evaluation of flight dynamics understanding.

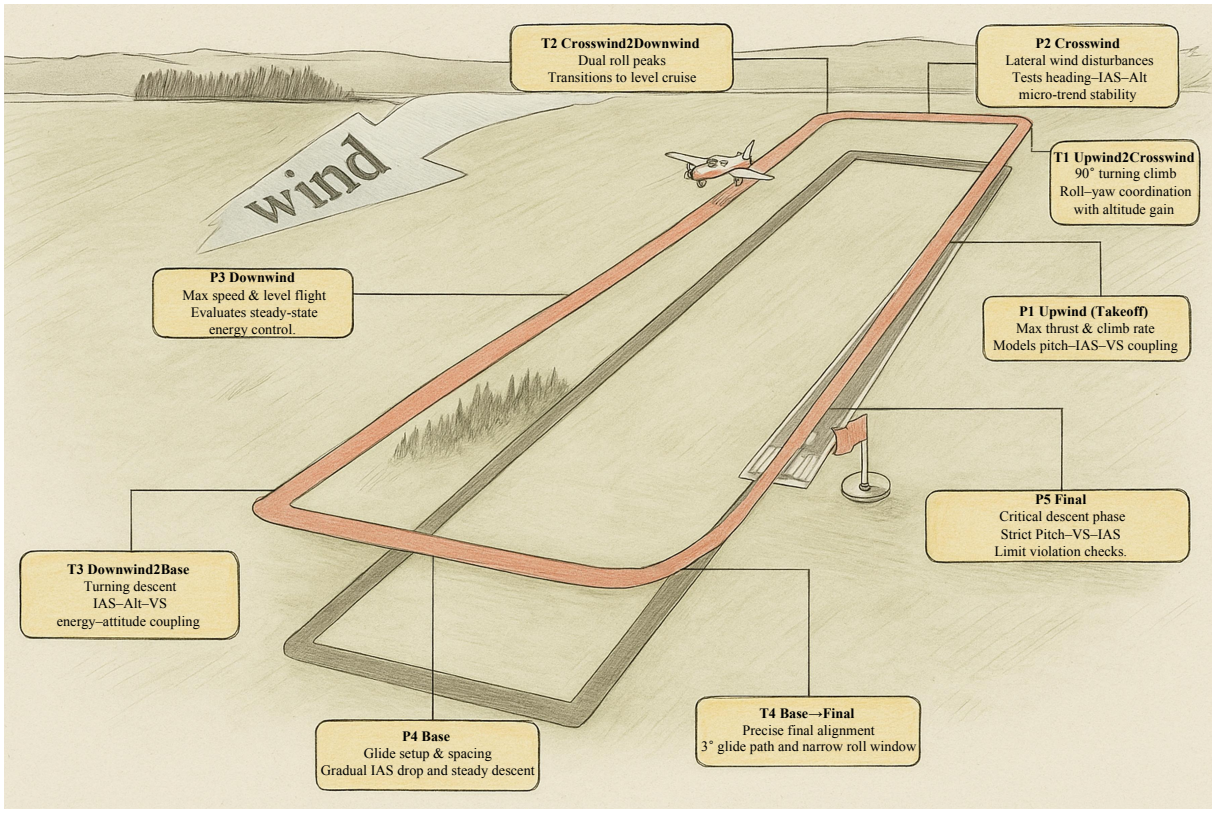

Fig. 2. Nine-phase segmentation in PILOTBENCH based on standard traffic pattern.

A. Task Definition

PILOTBENCH segments flights into nine precise phases, derived from a standard aviation traffic pattern as shown in Figure 2, consisting of five straight segments P1–P5 and four transition segments T1–T4. Phase boundaries are defined by heading change rates, roll angles, and vertical speeds, validated by annotated data.

Upwind P1 sets thrust, climb rate, VS, and IAS. T1 evaluates roll–yaw coordination through a 90° climbing turn. Crosswind P2 probes heading–IAS stability under lateral wind; T2 stresses dual-peak roll and altitude stabilisation. Downwind P3 maintains stable, level, high-speed flight parallel to the runway; T3 tests simultaneous throttle reduction, flap deployment, and descent initiation. Base P4 handles glide setup, descent rate, and IAS control; T4 demands runway alignment and glide-path interception. Final P5 enforces strict pitch, roll, and IAS limits during the critical landing approach.

B. Building PILOTBENCH

Inspired by existing benchmarks [6], [16] and tailored to safety-critical aviation, PILOTBENCH is built via an audited eight-stage pipeline, as shown in Figure 3, ensuring compliance with standards.

Data Collection. We recorded **38 h 18 min** of synchronised flight-sensor and avionics data from 31 visual-flight-rules (VFR) circuits on a DA40 and 22 dual-instruction sorties on a C172N. Each aircraft carried dual-frequency RTK-GNSS, an air-data computer (ADC), and an inertial reference unit (IRU) at 20 Hz, plus ARINC 429 streams and vendor annotations.

Data Processing. Channels were aligned to GNSS 1-PPS, resampled to 10 Hz, and consolidated into 34 standardised variables. ADC gaps greater than 0.8 s were forward-filled. Frames with HDOP greater than 2 or inertial residuals greater than 2.5σ were discarded, removing **1.7%** of the corpus.

Flight-Phase Annotation. With 45% missing labels, we apply rule-based seeding, contrastive boundary refinement, and expert vetting, achieving Cohen's kappa of 0.93.

Quality Assurance. We apply **34** sanity checks plus automatic screening [17]; low-quality or POH-violating segments are removed, keeping MI leakage less than 0.7%.

Expert Safety Review. An avionics engineer and former LOSA auditor reviewed all segments, anonymised identifiers, and verified extreme-attitude accuracy.

C. Statistics

The PILOTBENCH corpus contains 708 trajectory segments from a ten-day campaign, spanning altitudes 1,514–4,633 ft with a mean of 2,485 ft MSL, ground speeds 0–106.4 kt averaging 58.5 kt, with 34 synchronised sensor features per record. Figure 4 summarises key distributions: sorties trace a 4 km oval pattern with two dominant altitude bands at 1.7–2.1 kft and 2.4–3.2 kft covering 70%, three speed clusters, bimodal distance-to-airport peaks at 1.3 and 2.8 km, and headings concentrated at 270°–315° consistent with counter-clockwise circuits.

D. PILOT-SCORE

The metric $\mathcal{F} : \mathbb{R}^+ \times [0,1]^3 \to [0,100]$ combines regression accuracy and instruction-following:

$$\mathcal{F}(\text{MAE}, \text{RMSE}, \mathbf{c}) = 0.6 \cdot \mathcal{R}(\text{MAE}, \text{RMSE}) + 0.4 \cdot \mathcal{I}(\mathbf{c}) \quad (1)$$

where $\mathcal{R} = 0.6 \cdot S_{\text{MAE}} + 0.4 \cdot S_{\text{RMSE}}$ via piecewise linear scoring and $\mathcal{I}(\mathbf{c}) = 0.5c_1 + 0.3c_2 + 0.2c_3$ over Field Completeness, Validity, and Format. Instruction-following is computed by automated parsers; a floor score of 5 prevents complete failure classification. The metric is monotonically decreasing in error and increasing in compliance, bounded in $[5, 100]$, and aligned with Required Navigation Performance (RNP) standards: $\mathcal{F} \geq 90$ maps to RNP 0.1, ≥ 70 to RNP 0.3–1.0, ≥ 50 to RNP 2–4 [18]. The 60/40 regression-instruction weighting follows aviation safety assessments [19]–[21].

IV. EXPERIMENT

We evaluate 41 LLMs on PILOTBENCH. Each model ingests structured historical waypoints and outputs one-step

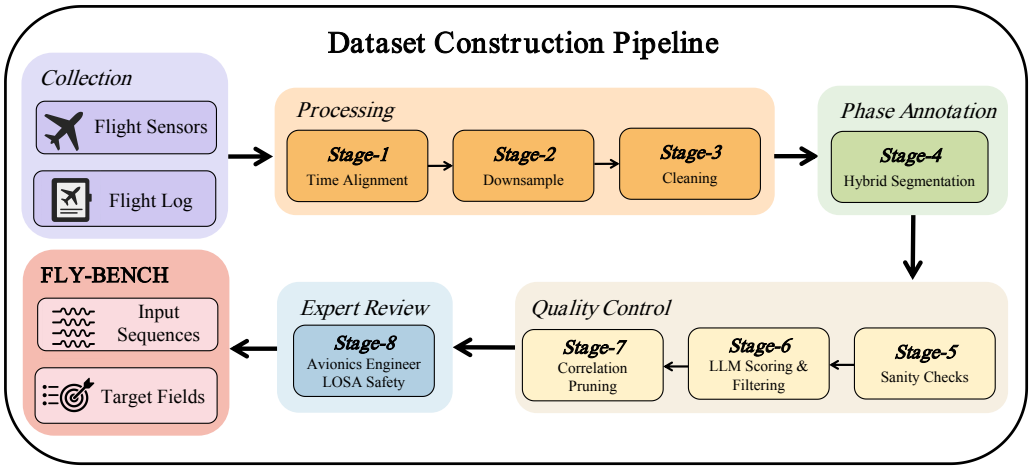

Fig. 3. Eight-stage pipeline for building PILOTBENCH.

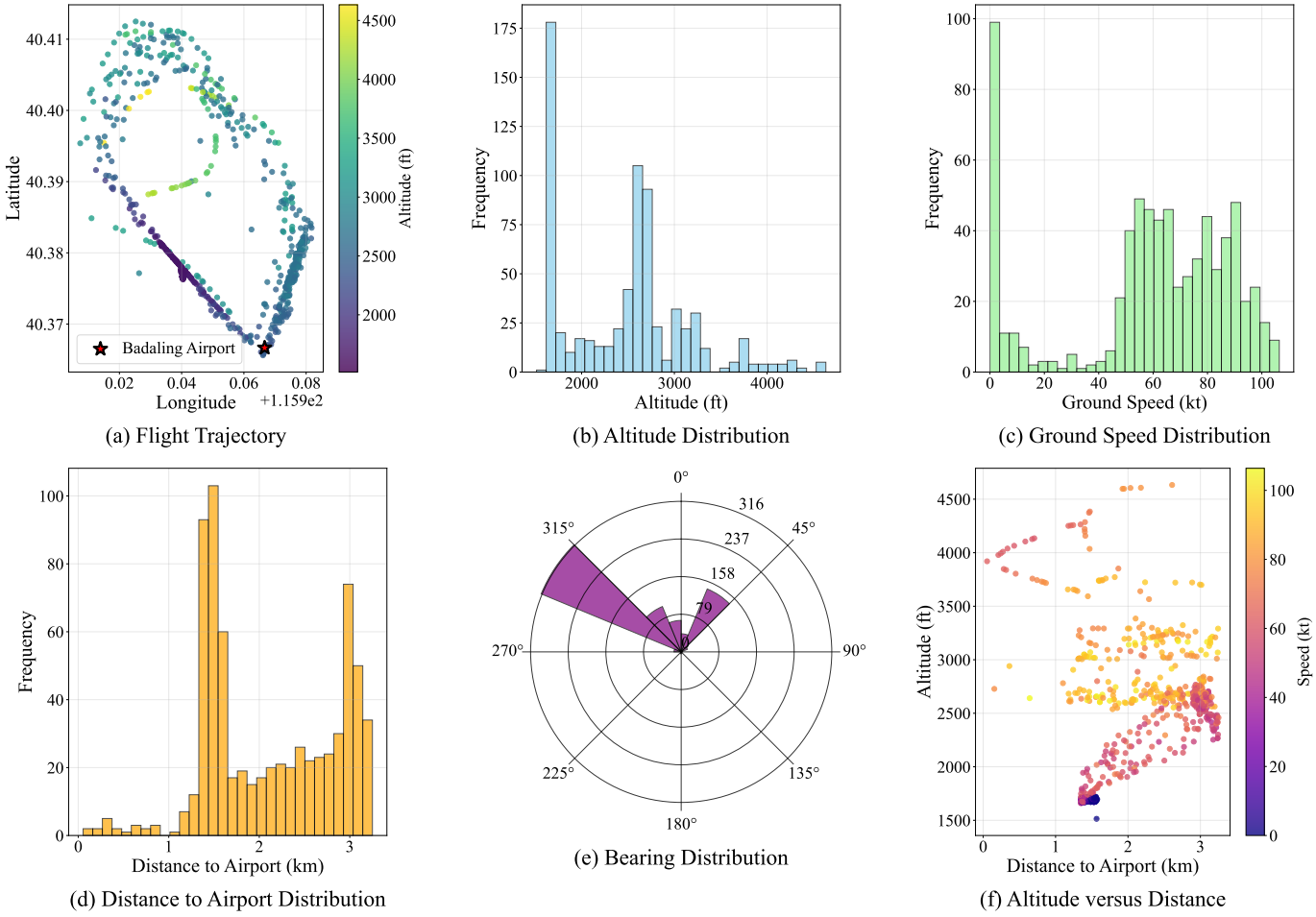

Fig. 4. Flight and spatial statistics of PILOTBENCH.

trajectory and attitude predictions in a fixed schema. Decoding is deterministic with temperature 0 and p=1.0. Experiments ran on eight NVIDIA A100 80GB GPUs; the model roster spans Qwen, DeepSeek, GLM, InternLM, GPT, and Doubao.

Results (Table I) show **Qwen3-32B** achieves the best MAE of 9.54. Scaling helps but non-linearly; architecture and training matter more than size. QVQ-72B-PREVIEW is numerically accurate yet often refuses instructions, limiting utility. Large models like Qwen2.5-72B reach 88.8% field validity; small models such as GLM-4-9B-Chat maintain format but attain only 0.7% validity, decoupling syntax from semantics. Errors are lowest in cruise, moderate in descent, and highest in climb, motivating phase-aware modeling.

Key failure patterns include numerical anomalies, flight-logic errors, structural issues, refusal responses, and temporal discontinuities—informing hybrid designs where LLMs handle intent while specialized modules ensure physics consistency.

V. PERFORMANCE EVALUATION ACROSS METHODS AND FLIGHT PHASES

To complement the 41-model screening, we conduct an ablation on 9 configurations: 3 traditional baselines, namely FlightPatchNet, DLinear, and PatchTST, and 6 LLM variants covering GPT-4o and Qwen3-32B under varied prompting strategies.

A. The Precision-Controllability Dichotomy

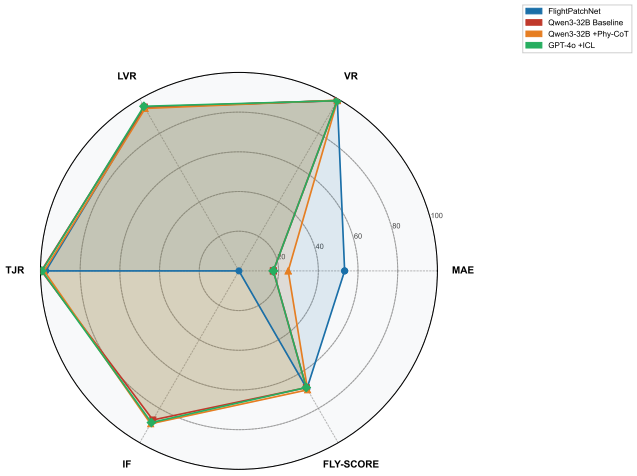

Fig. 5. Performance radar: traditional models shown in blue dominate MAE/VR; LLMs shown in orange, green, and purple gain IF controllability. Qwen3-32B +Phy-CoT achieves the best balance.

TABLE I
MAIN RESULTS ON PILOTBENCH

Model	Regression Error			Instruction Following				PILOT-SCORE
	MAE ↓	RMSE ↓	Overall ↑	Field Comp.	Field Valid	Format	Overall	
Qwen3-32B	**9.5409**	14.7005	85.43	100.0	91.5	100.0	**97.43**	**90.23**
Qwen2.5-72B-Instruct	11.9144	16.3591	83.20	100.0	88.8	100.0	96.64	88.58
DeepSeek-V3	11.9410	16.3006	83.19	100.0	86.7	100.0	96.00	88.31
GPT-4o-mini	12.1116	16.5832	82.99	99.8	86.3	99.8	95.74	88.09
GPT-o3-mini	10.9403	15.0377	84.24	100.0	76.8	100.0	93.05	87.76
Doubao-1.5Pro-32k	11.6238	15.8508	83.53	100.0	78.3	100.0	93.50	87.52
LoRA-Qwen2.5-14B-Instruct	13.1038	31.9880	79.12	99.8	89.4	99.8	96.67	86.14
Qwen2.5-32B-Instruct	13.0683	32.7015	79.01	100.0	82.2	100.0	94.66	85.27
QwQ-32B	9.7241	13.4456	85.53	100.0	46.5	100.0	83.95	84.90
Qwen2.5-14B-Instruct	14.0535	42.2590	76.31	99.8	89.3	99.8	96.65	84.44
Qwen2.5-VL-72B-Instruct	20.3666	76.1834	65.66	100.0	83.0	100.0	94.89	77.35
Qwen2.5-7B-Instruct	35.5291	110.1615	54.98	100.0	84.4	100.0	95.32	71.11
GLM-4-9B	41.1249	134.9971	50.75	100.0	88.2	100.0	96.45	69.03
DeepSeek-R1-Distill-Llama-70B	37.1562	129.6232	52.77	99.9	44.6	100.0	83.34	65.00
QVQ-72B-Preview	9.7267	**11.2431**	**85.97**	2.7	**97.7**	2.7	30.67	63.85
DeepSeek-R1-Distill-Qwen-32B	68.8524	190.3993	38.24	100.0	70.3	100.0	91.07	59.37
Qwen3-14B	150.4770	293.6560	19.84	100.0	80.3	100.0	94.09	49.54
DeepSeek-R1-Distill-Qwen-7B	273.8185	422.7081	6.57	99.5	82.6	99.7	94.42	41.71
GLM-Z1-Rumination-32B	84.5391	224.6919	32.23	29.8	25.2	99.0	42.03	36.15
GLM-4-9B-Chat	483.7880	553.1207	5.00	100.0	0.7	100.0	70.21	31.08

Figure 5 illustrates a fundamental trade-off in current modeling paradigms. Traditional forecasting baselines excel in pure regression with FlightPatchNet achieving MAE=7.01, leveraging inductive biases tailored for time-series extrapolation. However, their inability to process natural language renders them inert in human-on-the-loop scenarios requiring semantic guidance.

Conversely, LLMs sacrifice profound numerical precision with MAE ranging 11.28–13.59 to gain semantic controllability. This **Precision-Controllability Dichotomy** forces a choice: traditional models for fixed-trajectory prediction versus LLMs for adaptive, instruction-guided maneuvering. Notably, Qwen3-32B with Phy-CoT achieves the most effective compromise at PILOT-SCORE=69.21, maintaining acceptable regression error while enabling high-level safety constraints with VR below 0.8% and instruction adherence IF=88.9.

B. Safety and Accuracy as Orthogonal Targets

Figure 6 dissects the mechanisms of prompting improvements, revealing that accuracy and safety are improved via distinct pathways.

- **In-Context Learning (+ICL)** primarily acts as a *syntax stabilizer*. By providing exemplars, it reduces MAE—for instance, achieving -1.19 improvement for GPT-4o—and improves format compliance, but offers limited gains in physical safety.
- **Physics-CoT (+Phy-CoT)** functions as a *semantic constraint injector*. It explicitly grounds the reasoning process in flight dynamics, yielding dramatic safety improvements with VR reduction of approximately 25% that ICL alone cannot achieve.

This suggests that while few-shot prompting suffices for pattern matching, reliable safety-critical behavior requires explicit reasoning chains that verify physical feasibility before output generation.

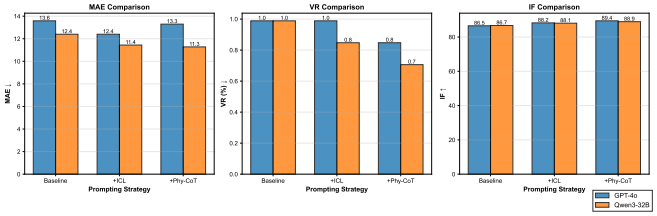

Fig. 6. Decoupling prompting effects: +ICL benefits MAE precision; +Phy-CoT drives VR safety gains, suggesting distinct mechanisms for pattern matching vs. constraint satisfaction.

C. The Dynamic Complexity Gap

Phase-stratified analysis (Figure 7) exposes a **Dynamic Complexity Gap**. While all models perform robustly in low-dynamic phases like Cruise and Descent with MAE of 7–10, LLM performance degrades disproportionately in high-workload phases.

In Climb and Approach—characterized by coupled changes in altitude, airspeed, and configuration—LLM error rates spike above MAE 13, whereas specialized forecasters like FlightPatchNet maintain stability below MAE 8. This sensitivity suggests that LLMs' implicit physics models are brittle: adequate for steady-state extrapolation but fragile under the coupled non-linear dynamics of aggressive maneuvering.

VI. CONCLUSION

PILOTBENCH reveals that LLMs demonstrate emergent capabilities in semantic interpretation and instruction following, yet face intrinsic limitations in high-precision physics modeling. The **Precision-Controllability Dichotomy** exposes that traditional forecasters excel at numerical prediction but remain inert to natural language guidance, while LLMs sacrifice regression fidelity to gain semantic controllability. The **Dynamic Complexity Gap** reveals that LLMs' acceptable

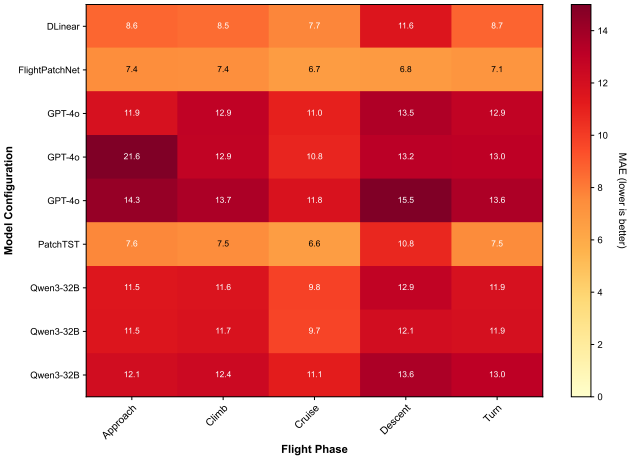

Fig. 7. MAE heatmap across flight phases. LLM errors spike in Climb and Approach, marked in red, while traditional baselines remain stable in yellow, exposing the Dynamic Complexity Gap.

performance in steady-state regimes collapses under coupled, high-workload maneuvers.

These findings motivate a **Hybrid Architectural Paradigm**: leveraging LLMs for intent understanding while delegating high-frequency control to specialized forecasters, generalizing beyond aviation to any safety-critical domain requiring both semantic reasoning and physical prediction.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, 62406033, U1636211, 61672081), and the State Key Laboratory of Complex & Critical Software Environment (Grant No. SKLCCSE-2024ZX-18). This study used exclusively anonymized device-level flight sensor data; no human subjects or personally identifiable information were involved.

AI Usage Disclosure. In accordance with IEEE policy, we disclose that AI-assisted writing tools (e.g., GPT-5) were used for language refinement and editing in the preparation of this manuscript. The use of AI-generated text is limited to improving clarity and readability. No AI tools were used in the design of the study, data collection, data analysis, or the derivation of scientific conclusions.

REFERENCES

[1] R. Mon-Williams, G. Li, R. Long, W. Du, and C. G. Lucas, "Embodied large language models enable robots to complete complex tasks in unpredictable environments," *Nature Machine Intelligence*, pp. 1–10, 2025.

[2] Y. Ma, C. Cui, X. Cao, W. Ye, P. Liu, J. Lu, A. Abdelraouf, R. Gupta, K. Han, A. Bera *et al.*, "Lampilot: An open benchmark dataset for autonomous driving with language model programs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 141–15 151.

[3] S. Abdulhak, W. Hubbard, K. Gopalakrishnan, and M. Z. Li, "Chatatc: Large language model-driven conversational agents for supporting strategic air traffic flow management," *arXiv preprint arXiv:2402.14850*, 2024.

[4] D. Guo, Z. Zhang, B. Yang, J. Zhang, H. Yang, and Y. Lin, "Integrating spoken instructions into flight trajectory prediction to optimize automation in air traffic control," *Nature Communications*, vol. 15, no. 1, p. 9662, 2024.

[5] H. Liu, F. Shen, F. Gao *et al.*, "Research on flight accidents prediction based back propagation neural network," *arXiv preprint arXiv:2406.13954*, 2024.

[6] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, "Agieval: A human-centric benchmark for evaluating foundation models," in *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, K. Duh, H. Gómez-Adorno, and S. Bethard, Eds. Association for Computational Linguistics, 2024, pp. 2299–2314.

[7] S. An, Z. Ma, Z. Lin, N. Zheng, J.-G. Lou, and W. Chen, "Make your llm fully utilize the context," *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 160–62 188, 2024.

[8] E. Q. Wu, Y. Gao, W. Tong, Y. Hou, R. Law, and G. Zhu, "Cognitive state detection in task context based on graph attention network during flight," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[9] D. Guo, Z. Zhang, Z. Yan, J. Zhang, and Y. Lin, "Flightbert++: A non-autoregressive multi-horizon flight trajectory prediction framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 127–134.

[10] Y. Xu, Q. Pan, Z. Wang, and B. Hu, "A novel trajectory prediction method based on cnn, bilstm, and multi-head attention mechanism," *Aerospace*, vol. 11, no. 10, p. 822, 2024.

[11] X. Dong, Y. Tian, K. Niu, M. Sun, and J. Li, "Research on flight trajectory prediction method based on transformer," in *International Conference on Smart Transportation and City Engineering (STCE 2023)*, vol. 13018. SPIE, 2024, pp. 1403–1409.

[12] N. Schimpf, Z. Wang, S. Li, E. J. Knoblock, H. Li, and R. D. Apaza, "A generalized approach to aircraft trajectory prediction via supervised deep learning," *IEEE Access*, vol. 11, pp. 116 183–116 195, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3325053

[13] J. Sun, X. Olive, M. Strohmeier, and V. Lenders, "Opensky report 2025: Improving crowdsourced flight trajectories with ads-c data," in *2025 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE, 2025, pp. 1–8.

[14] J. Patrikar, J. P. A. Dantas, B. G. Moon, M. M. Hamidi, S. Ghosh, N. V. Keetha, I. Higgins, A. Chandak, T. Yoneyama, and S. A. Scherer, "Tartanaviation: Image, speech, and ADS-B trajectory datasets for terminal airspace operations," *CoRR*, vol. abs/2403.03372, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.03372

[15] Y. Qin, S. Liang, Y. Ye, K. Zhu *et al.*, "Toolllm: Facilitating large language models to master 16000+ real-world apis," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=dHng2O0Jjr

[16] Z. Guo, S. Cheng, H. Wang, S. Liang *et al.*, "Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 11 143–11 156. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-acl.664

[17] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[18] F. A. Administration, *Airplane Flying Handbook: FAA-H-8083-3C (2025)*. Simon and Schuster, 2022.

[19] T. B. Spence, R. O. Fanjoy, C.-t. Lu, and S. W. Schreckengast, "International standardization compliance in aviation," *Journal of air transport management*, vol. 49, pp. 1–8, 2015.

[20] D. D. Boyd, "A review of general aviation safety (1984–2017)," *Aerospace medicine and human performance*, vol. 88, no. 7, pp. 657–664, 2017.

[21] T. Puranik, H. Jimenez, and D. Mavris, "Energy-based metrics for safety analysis of general aviation operations," *Journal of Aircraft*, vol. 54, no. 6, pp. 2285–2297, 2017.