Fine-Tuning Large Language Models for Cooperative Tactical Deconfliction of Small Unmanned Aerial Systems

Iman Sharifi* Alex Zongo* Peng Wei
George Washington University
{i.sharifi,a.zongo,pwei}@gwu.edu
*Equal contribution

Abstract

The growing deployment of small Unmanned Aerial Systems (sUASs) in low-altitude airspaces has increased the need for reliable tactical deconfliction under safety-critical constraints. Tactical deconfliction involves short-horizon decision-making in dense, partially observable, and heterogeneous multi-agent environments, where both cooperative separation assurance and operational efficiency must be maintained. While Large Language Models (LLMs) exhibit strong reasoning capabilities, their direct application to air traffic control remains limited by insufficient domain grounding and unpredictable output inconsistency. This paper investigates LLMs as decision-makers in cooperative multi-agent tactical deconfliction using fine-tuning strategies that align model outputs to human operator heuristics. We propose a simulation-to-language data generation pipeline based on the BlueSky air traffic simulator that produces rule-consistent deconfliction datasets reflecting established safety practices. A pretrained Qwen-Math-7B model is fine-tuned using two parameter-efficient strategies: supervised fine-tuning with Low-Rank Adaptation (LoRA) and preference-based fine-tuning combining LoRA with Group-Relative Policy Optimization (GRPO). Experimental results on validation datasets and closed-loop simulations demonstrate that supervised LoRA fine-tuning substantially improves decision accuracy, consistency, and separation performance compared to the pre-trained LLM, with significant reductions in near mid-air collisions. GRPO provides additional coordination benefits but exhibits reduced robustness when interacting with heterogeneous agent policies.

1. Introduction

The rapid growth in civil and commercial deployment of small Unmanned Aerial Systems (sUASs), including package delivery, infrastructure inspection, and emergency response, has intensified the demand for safe and efficient operations in low-altitude, shared airspaces [10, 25]. As traffic density increases, conflicts between vehicles become inevitable, particularly near intersections, merging corridors, and other constrained airspace regions. Tactical deconfliction, which involves real-time, short-horizon decision-making that maintains safe separation while preserving operational efficiency, has therefore emerged as a central challenge in UAS traffic management (UTM) ecosystems. Unlike strategic planning [22] or trajectory optimization, tactical deconfliction must operate under strict time constraints, partial observability, and complex multi-agent interactions, where delayed or overly conservative decisions can significantly degrade both safety and traffic throughput [29]. Rule-based approaches lack flexibility and scalability [8, 16], while optimization-based and learning-based methods often struggle with latency, robustness, or interpretability under safety-critical constraints [6, 27].

Recent advances in Large Language Models (LLMs) have shown strong capabilities in reasoning [18, 24], contextual understanding [20], and sequential decision-making [30], making them a promising candidate for tactical deconfliction in dense, uncertain multi-agent airspaces [7]. Yet, general-purpose LLMs are not designed for safety-critical aviation [33]: zero-shot or prompt-based use can yield inconsistent and prompt-sensitive outputs [9, 14], misaligned with human safety norms [19], and uninformed about domain-specific trade-offs [19]. These limitations motivate systematic alignment of LLM behavior with human tactical reasoning in sUAS operations.

Human experts (e.g., air traffic controllers and experienced pilots) resolve conflicts by applying implicit safety principles, i.e., prioritizing separation, anticipating others' intent and reasoning over short horizons, rather than optimizing explicit reward functions [23]. We therefore advocate leveraging human-aligned datasets that encode expert knowledge as logical rules, and fine-tuning LLMs to transfer these judgments and preferences into inference-time behavior [26]. Compared to trial-and-error multi-agent rein-

forcement learning [5], human-aligned fine-tuning can inject domain-appropriate reasoning priors while promoting interpretability and behavioral consistency [2].

In this paper, we present a simulation-to-language dataset generation pipeline that enables systematic learning of human-aligned cooperative tactical deconfliction behaviors from high-fidelity air traffic simulations. The proposed pipeline generates diverse multi-agent scenarios, encodes human tactical knowledge through logical rules, and transforms raw simulation data into structured prompt–response pairs suitable for training LLMs. Using this dataset, we study two complementary fine-tuning strategies for adapting pre-trained LLMs to tactical deconfliction in multi-agent sUAS environments. To the best of our knowledge, this work constitutes the first systematic investigation of fine-tuned LLMs for tactical deconfliction evaluated both on held-out datasets and in closed-loop air traffic simulations. The main contributions of this work are as follows:

- We develop a simulation-to-language dataset generation pipeline based on the BlueSky air traffic simulator [15] that enables rapid construction of large-scale, rule-consistent tactical deconfliction datasets, allowing LLMs to internalize human safety heuristics and operational preferences.
- We demonstrate that parameter-efficient Supervised Fine-Tuning (SFT) with Low-Rank Adaptation (LoRA) [17] substantially improves LLM decision accuracy, behavioral consistency, and separation safety compared to a pretrained baseline, as validated through both offline evaluation and closed-loop simulation.
- We evaluate the performance of the preference-based fine-tuning using Group-Relative Policy Optimization (GRPO) compared to SFT, providing insight into the strengths and limitations of reinforcement-style alignment for tactical deconfliction.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 formulates the tactical deconfliction problem and the fine-tuning strategies. Section 4 describes the simulation-to-language dataset generator pipeline. Section 5 elaborates on the two fine-tuning strategies. Section 6 reports experimental results and comparative evaluations. Finally, Section 7 draws conclusions.

2. Related Work

Recent research has explored the application of LLMs to air traffic control. These efforts have primarily positioned LLMs as high-level reasoning, interface, or knowledge-support components rather than direct low-level controllers. Several studies employ LLMs as natural-language interfaces integrated with existing conflict resolution solvers, allowing air traffic controllers to express preferences and constraints while preserving safety guarantees through restricted LLM outputs limited to filtering or ranking candidate solutions [21]. Other work investigates LLMs as embodied or tool-augmented agents capable of directly issuing control commands in simulation environments [3], often augmented with role decomposition or experience libraries to improve reasoning consistency. Complementary efforts leverage LLMs for air traffic scenario generation [13], aviation-domain knowledge modeling [31], and systematic evaluation of LLM reliability, recall, and reasoning performance in aviation contexts [12].

These studies reveal recurring limitations, including sensitivity to prompt structure, hallucinations, limited recall, inference latency, and the absence of explicit alignment with human operational preferences, that pose significant challenges for real-time, safety-critical tactical deconfliction. Unlike prior LLM-based ATC approaches that rely on zero-shot prompting or prompt engineering with function-calling at inference time [3], this paper adopts a systematic fine-tuning strategy grounded in human-aligned data and achieves near-real-time performance. By adapting pre-trained LLMs through parameter-efficient fine-tuning and preference-aware optimization on rule-consistent, simulator-generated datasets, our approach positions LLMs as human-aligned tactical decision-makers rather than free-form reasoning agents. This directly addresses the reliability and consistency concerns highlighted in existing LLM-based ATC research.

3. Problem Formulation and Methodology

3.1. Tactical Deconfliction with LLM-based Policies

We consider a tactical deconfliction problem in a shared low-altitude airspace populated by multiple sUASs with heterogeneous configurations and decision-making policies. Agents may differ in kinematic limits, sensing capabilities, maneuverability, and onboard autonomy architecture. The objective of tactical deconfliction is to maintain, cooperatively, safe separation while minimizing unnecessary deviations from nominal mission trajectories. Decisions must be made in real time under partial observability and amid complex multi-agent interactions.

Rather than addressing deconfliction through continuous control or trajectory optimization, we formulate the problem at the policy level. At each decision step, an agent observes a structured representation of the surrounding environment, including its own state, nearby traffic information, and safety constraints. Based on this context, the agent selects a discrete tactical action, such as accelerating, maintaining speed, or decelerating. The LLM serves as a high-level policy that maps structured agent state descriptions to tactical decisions. The LLM outputs abstract actions that are subsequently executed by UAS flight control modules, allowing the model to reason over heterogeneous agent interactions and implicit safety priorities with-

out requiring access to explicit models of low-level dynamics. To align the LLM behavior with domain-specific operational requirements, we fine-tuned the model on a large-scale dataset spanning diverse traffic scenarios.

3.2. Fine-Tuning Strategies

3.2.1. Supervised Fine-Tuning (SFT)

This first strategy adapts a pre-trained LLM to tactical deconfliction through supervised learning on human-aligned, rule-consistent datasets. Each training sample consists of a structured description of the ownship's local traffic context paired with a target tactical action derived from human-designed safety rules. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the ownship context and y_i the corresponding target action, the objective is to maximize the conditional likelihood of human-aligned decisions under the fine-tuned model.

Formally, SFT minimizes the negative log-likelihood loss $\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\log p_\theta(y \mid x)\right]$, where p_θ denotes the LLM parameterized by θ. Minimizing this loss transfers human decision heuristics into the model's inference behavior, encouraging consistent reproduction of safety-oriented tactical actions across similar agent state configurations.

To enable efficient domain adaptation without updating the full parameter set of the LLM, we employ Low-Rank Adaptation (LoRA) [17], as shown in Figure 1, which injects trainable low-rank updates into selected projection layers while keeping the pretrained weights frozen. This parameter-efficient design enables scalable adaptation while preserving the general reasoning capabilities of the base model.

3.2.2. Group-Relative Policy Optimization (GRPO)

The second fine-tuning strategy employs GRPO, a preference-based alignment method that refines LLM behavior using sampled candidate actions and scalar reward feedback. For a given agent context x, the pretrained LLM generates a set of candidate tactical responses $\{y^{(1)}, \ldots, y^{(K)}\}$ via high-temperature sampling, promoting decision exploration beyond deterministic imitation.

Each candidate response is evaluated using a task-specific reward function $R(x, y)$ that encodes human-aligned safety rules and operational preferences, assigning higher scores to actions that maintain separation, respect right-of-way precedence, and favor conservative, interpretable maneuvers. These rewards are used to compute a group-relative advantage

$$\hat{A}^{(k)} = R(x, y^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} R(x, y^{(j)}), \tag{1}$$

which measures the relative quality of each response within the sampled group.

Model parameters are then updated using an objective based on Proximal Policy Optimization (PPO) loss that increases the likelihood of higher-advantage responses while constraining policy updates for stability. The resulting GRPO loss is given by

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{x,y^{(k)}}[\,&\min(\rho^{(k)}\hat{A}^{(k)}, \\ &\text{clip}(\rho^{(k)}, 1-\epsilon, 1+\epsilon)\hat{A}^{(k)})],\end{aligned} \tag{2}$$

where $\rho^{(k)} = \frac{p_\theta(y^{(k)}|x)}{p_{\theta_{\text{old}}}(y^{(k)}|x)}$ denotes the likelihood ratio between the updated and previous policies, and ϵ is a clipping parameter. As in SFT, GRPO updates are applied exclusively through LoRA parameters, leaving the base model unchanged.

By combining stochastic exploration, rule-based reward evaluation, and PPO-style optimization, GRPO enables preference-driven refinement of LLM decision-making beyond direct imitation. Unlike SFT, which enforces alignment through supervised reproduction of human actions, GRPO encourages relative improvement among competing candidate responses. This distinction enables a principled comparison between imitation-based and preference-based alignment for safety-critical tactical deconfliction.

4. Dataset Generation Pipeline

As major companies increasingly deploy sUAS fleets in shared airspace, safety- and privacy-related constraints have become central considerations to their operational frameworks. Due to proprietary concerns and regulatory sensitivities, high-fidelity operational data relevant to tactical deconfliction is rarely publicized, limiting the availability of real-world datasets for learning-based methods. This lack of accessible data poses a fundamental barrier to the development and evaluation of data-driven deconfliction policies, which typically rely on large-scale, representative training corpora. To address this challenge, we design a simulation-based dataset generation pipeline that enables systematic, privacy-preserving collection of human-aligned tactical decision data, while remaining extensible to future integration with real-world observations as such data become available.

Thus, to collect trainable datasets including pairs of prompts and rule-based responses, we designed a simulation-to-language pipeline that generates scenarios and converts them to trainable prompt-answer pairs. Figure 1 illustrates an overview of the pipeline, in which we initially collect a series of high-fidelity multi-agent simulations using the BlueSky Air Traffic Simulator [15]. The simulation environment was configured to emulate low-altitude airspace over the city of Frisco, Texas, a representative urban hub for drone delivery operations. The dataset generation pipeline includes the following stages:

Scenario Configurations: We generated diverse multi-agent flight scenarios to capture the traffic complexity and

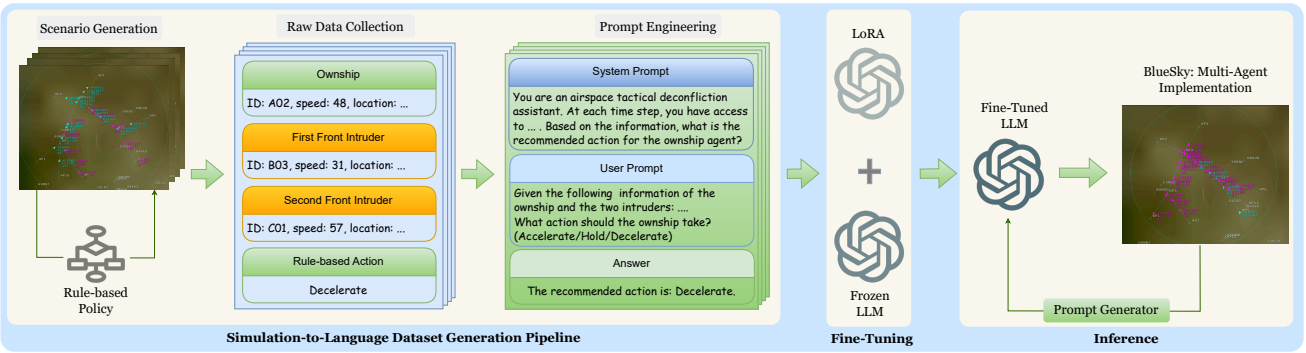


Figure 1. **Architecture overview.** The figure illustrates the end-to-end system architecture and the role of the proposed simulation-to-language dataset generation pipeline. Multi-agent traffic scenarios are generated in the BlueSky simulator, from which raw state data are extracted and converted into structured natural-language prompts using rule-based supervision. The resulting prompt–response pairs constitute the training dataset for LoRA-based fine-tuning. At deployment, the fine-tuned LLM generates tactical actions for multiple agents, which are executed in BlueSky, closing the simulation loop.

interaction patterns characteristic of urban low-altitude operations. Each scenario involves 20–30 sUASs operating concurrently in shared airspace and includes two merging points and one intersection, reflecting common bottlenecks in drone delivery corridors. To introduce variability in traffic density and agent state geometry, the number of active flight routes per scenario was randomly varied between four and six, producing heterogeneous traffic flows with intersecting and merging trajectories.

To model realistic fleet diversity, we defined two distinct agent configurations characterized by different speed limits, acceleration capabilities, and sensing ranges. These configurations represent heterogeneous vehicle capabilities commonly observed across different drone operators and enable systematic evaluation of an LLM's ability to generalize across agents with varying dynamics. Specifically, we consider configurations `X` and `Y`, where configuration `X` exhibits stronger kinematic and sensing capabilities than configuration `Y`. The speed and acceleration limits for configurations `X` and `Y` are selected based on the performance specifications of the Google Wing Hummingbird drone [32] and the Amazon MK30 drone [11], respectively. Sensing ranges reflect current technological constraints associated with Remote ID-based communication or radar-based detection systems, ensuring realistic perception asymmetry among agents.

Table 1 summarizes the kinematic and sensing specifications for each configuration. By incorporating heterogeneous vehicle capabilities and structurally complex airspace layouts, the proposed scenario design yields a challenging and representative testbed for learning and evaluating cooperative tactical deconfliction policies under realistic drone delivery operations.

Rule-Based Policy Design: To generate human-aligned supervisory signals for tactical deconfliction, we designed a deterministic rule-based policy that enforces safe separation across all simulated scenarios. The policy is intended to emulate human pilot or controller reasoning by prescribing actions through interpretable if–then rules derived from operational heuristics.

Table 1. Kinematic and sensing specifications for UAS configurations `X` (strong) and `Y` (weak).

Parameter	Notation	Configuration	
		`X` (strong)	**`Y` (weak)**
Speed Range (m/s)	$[v_{\min}, v_{\max}]$	$[0, 44.88]$	$[0, 30.12]$
Acceleration (m/s^2)	$\Delta v/\Delta t$	$\{-1.71, 0, 1.71\}$	$\{-1.02, 0, 1.02\}$
Sensing Range (m)	$\mathcal{R}$	1000	750

At each decision step, the policy evaluates the local traffic context of a given agent (referred to as the ownship) and selects an appropriate tactical action based on multiple state-dependent factors. These include the ownship's current and desired speeds, distance to the next waypoint, the number of nearby intruders, and relative spatial relationships with those intruders. To balance computational efficiency with behavioral fidelity, only the two closest front intruders are considered, as they typically represent the most critical conflict threats in dense airspace configurations.

The policy further distinguishes between intruders operating on the same route and those on intersecting or merging routes. The policy is then enabled to modulate maneuver aggressiveness based on conflict geometry. Based on the evaluated conditions, the rule engine outputs one of three discrete tactical actions: *Accelerate*, *Hold*, or *Decelerate*. These actions serve as human-aligned supervisory labels

for dataset generation rather than as optimized control commands.

The complete rule hierarchy, decision thresholds, and tie-breaking logic are detailed in Supplementary Material (Appendix A).

Raw Data Collection: For each simulation episode, state information was recorded for every ownship at discrete time steps. The collected data include the ownship's position, velocity, heading, route identifier, and distance to the next waypoint, along with detailed information about the two closest front intruder agents, such as their relative positions, velocities, and distances to their respective waypoints.

To support flexible prompt construction and preserve contextual richness, both essential and supplementary attributes were retained during data logging. This design choice ensures that no potentially relevant information is lost during post-processing and allows multiple prompt formulations to be explored without re-running simulations. The resulting dataset captures dynamic multi-agent interactions across thousands of time steps and diverse traffic configurations. An example of the raw observation record is provided in Supplementary Material (Appendix B). In total, over 38K state–action samples were collected in under 10 minutes. The data collection pipeline is fully modular, enabling additional scenarios and samples to be generated as needed.

Prompt Engineering: As illustrated in Figure 1, following data collection, the raw numerical and categorical state information was transformed into structured natural-language prompts suitable for LLM training. Each prompt consists of two components: a *system prompt*, which defines the model's operational role and high-level objectives (e.g., ensuring safe separation in shared airspace), and a *user prompt*, which describes the current local traffic situation of the ownship and nearby intruders in natural language. An illustrative example of the prompt format is presented in Supplementary Material (Appendix C).

This translation process converts low-level simulator states into human-readable descriptions that emphasize relative relationships, safety-relevant constraints, and decision context. As a result, the LLM is encouraged to infer tactical reasoning patterns rather than merely learning numerical correlations. The prompt format is kept consistent across training and inference to ensure behavioral stability.

The resulting pipeline produces a large-scale, context-rich dataset that embeds human tactical reasoning through interpretable rule-based supervision. The pipeline is computationally efficient, enabling rapid generation of training data and straightforward scaling to larger datasets as needed. Moreover, the pipeline's modular architecture allows both the rule-based policy and prompt engineering strategy to be replaced without modifying the underlying simulation infrastructure. By grounding LLM training data in high-fidelity simulations while maintaining flexibility and scalability, the pipeline provides a principled and extensible foundation for aligning LLM inference behavior with safety-critical deconfliction objectives.

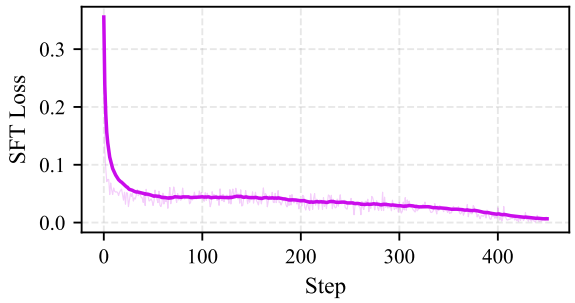
5. LLM Selection and Fine-Tuning

For this study, we selected Qwen-Math-7B [1, 4, 28] as the pretrained backbone for all fine-tuning experiments. Qwen-Math-7B is a member of the Qwen-2.5 family of transformer-based language models and is optimized for enhanced reasoning, mathematical comprehension, and logical consistency. Unlike general-purpose instruction-tuned models, Qwen-Math-7B incorporates domain-focused pretraining on scientific and quantitative corpora, enabling robust structured reasoning and symbolic manipulation. These characteristics make it well suited for tactical deconfliction tasks, which require reasoning over spatial relationships, safety margins, and action consequences under uncertainty. Throughout this paper, we refer to the pretrained model as the *Base* model.

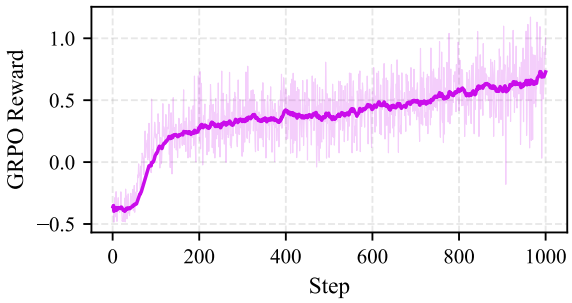
LoRA Configuration: As illustrated in Figure 1, we adapt the Base (Frozen) LLM to the tactical deconfliction domain via LoRA-based fine-tuning, implemented using the `transformers` library with a PyTorch backend. LoRA adapters were applied to the feed-forward projection layers (`up_proj`, `down_proj`, and `gate_proj`) as well as attention projection layers (`q_proj`, `k_proj`, and `v_proj`) to enhance contextual reasoning. For both SFT and GRPO, the LoRA rank, scaling factor, and dropout were set to 8, 32, 0.05, respectively. The learning rates for SFT and GRPO are set to 10^{-4} and 5×10^{-6}, respectively. The rest of the parameters are set to default values in the corresponding Python packages. These hyperparameters were chosen to balance adaptation capacity, training stability, and computational efficiency.

Due to memory limitations, we restricted output generation to 10 tokens to reduce inference time when serving multiple agents. Similarly, GRPO fine-tuning sampled four candidate responses per prompt to compute the advantage function following Eq (1) and using maximum temperature to encourage exploration. Both SFT and GRPO training were conducted for a single epoch, requiring approximately 6 and 14 hours, respectively. Optimization was performed using the AdamW optimizer with a cosine learning-rate schedule and warm-up steps to ensure stable convergence.

Reward Function in GRPO: The reward signal guiding GRPO optimization combines two complementary components: a format reward and an action reward. The format reward, denoted as r_{format}, encourages adherence to the desired response structure by quantifying normalized textual similarity between the generated response $\hat{y}$ and the ground-truth response y via Levenshtein similar-

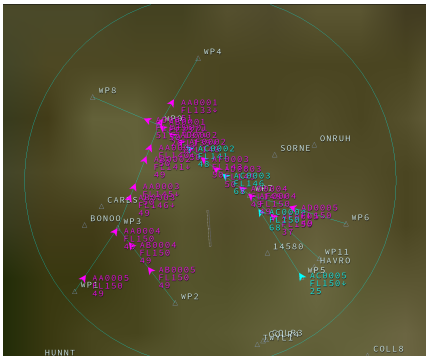


(a) Loss curve during SFT training.

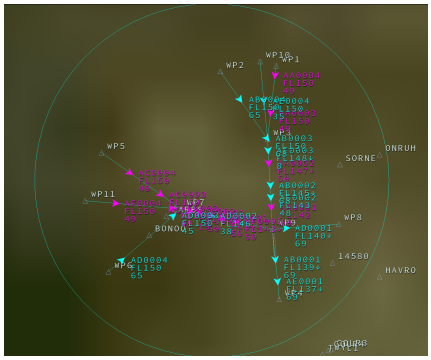


(b) Reward progression during GRPO fine-tuning.

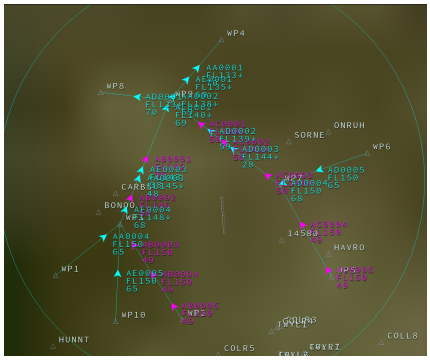
Figure 2. Training effectiveness of fine-tuning methods. (a) shows the supervised learning progress through loss reduction, hence accuracy increase, while (b) shows the GRPO reward evolution across training iterations.



(a) Scenario A



(b) Scenario B



(c) Scenario C

Figure 3. Traffic snapshots for the three scenarios (A, B, C) used in Table 3. The LLM agents and the Rule-based agents are colored in pink and green, respectively. Each scenario has 5-6 routes, each of which hosts 5 agents with random spawning times. Throughout all scenarios, we considered 10 LLM agents, and the rest are Rule-based agents.

ity: $r_{\text{format}} = \left(1 - \frac{\Gamma(\hat{y},y)}{\max(|\hat{y}|,|y|)}\right)^{\gamma}$, where $\Gamma(\cdot)$ denotes the Levenshtein distance and $\gamma \in [1,\infty)$ controls sensitivity to formatting deviations. The action reward, denoted as r_{action}, enforces decision correctness by verifying whether the action specified in the generated response matches the ground-truth action label: $r_{\text{action}} = \mathbb{I}[\text{action}(\hat{y}) = \text{action}(y)] - 0.5$, where $\mathbb{I}[\cdot]$ is the indicator function. The offset of -0.5 centers the reward around zero, penalizing incorrect actions while rewarding correct ones. The overall reward is computed as $r(y_k, x) = \lambda_f r_{\text{format}}(y_k, x) + \lambda_a r_{\text{action}}(y_k, x)$, with weighting coefficients λ_f and λ_a balancing structural compliance and decision accuracy.

All experiments were conducted on two NVIDIA RTX 3090 GPUs using mixed-precision training to reduce memory consumption and improve throughput. GRPO training was implemented using the TRL framework. Through this fine-tuning process, Qwen-Math-7B internalizes both rule-based decision logic and context-dependent tactical reasoning, yielding interpretable and safety-aligned decision policies suitable for cooperative multi-agent tactical deconfliction.

Table 2. Performance comparison on the evaluation dataset. All numbers are reported in percent (%).

Model	Accuracy	Precision	Recall	F1-score
Base	27	**75**	20	31
SFT	**88**	**75**	**66**	**69**
GRPO	53	**75**	40	50

6. Experimental Results and Discussions

Figure 2 illustrates the training dynamics of the two fine-tuning approaches. The SFT loss curve exhibits stable convergence, indicating effective supervised alignment with human-labeled actions, while the GRPO reward trajectory reflects gradual preference-based policy refinement. These trends suggest that both methods effectively incorporate training signals, albeit through different learning mechanisms. To comprehensively assess the fine-tuned models,

we use two strategies:

6.1. Evaluation with Datasets

We first assess the effectiveness of the proposed fine-tuning strategies on a held-out dataset of prompt–response pairs different from the training data. Each sample consists of a natural-language description of a local traffic situation and a corresponding ground-truth tactical action. This evaluation assesses how accurately each model reproduces the desired decision given identical inputs.

During testing, all models were prompted with the same evaluation set, and their generated responses were compared against the reference labels. A prediction was deemed *correct* if the response contained the target action (*Accelerate*, *Hold*, or *Decelerate*); otherwise, it was classified as incorrect. This criterion enables a consistent comparison among the pretrained Base model, the SFT model, and the GRPO fine-tuned model.

Quantitative results on the evaluation dataset are reported in Table 2 using standard classification metrics. The Base model achieves an accuracy of 27%, underscoring the challenge posed by tactical deconfliction for general-purpose LLMs without domain adaptation. In contrast, SFT with LoRA substantially improves performance, achieving an accuracy of 88% and an F1-score of 69%, indicating effective alignment with the structured decision patterns encoded in the dataset. The improvement in recall indicates that the SFT model generalizes more reliably across diverse conflict geometries.

The GRPO fine-tuned model attains moderate gains over the Base model, with an accuracy of 53% and an F1-score of 50%. While preference-based optimization improves response structure and consistency, its performance remains below that of SFT under the current reward formulation. This outcome suggests that, for this task, direct supervised alignment with human-labeled actions provides a stronger learning signal than relative preference optimization alone.

6.2. Evaluation with BlueSky Simulations

We next evaluate the fine-tuned LLM policies in closed-loop multi-agent simulations using the BlueSky simulator. Figure 1 illustrates the *Inference* loop. At each simulation time step, the state information of every LLM-controlled agent is transformed into a structured prompt by the *prompt generator* and passed to the fine-tuned LLM, which outputs the corresponding tactical actions. These actions are then applied to the simulator to update the environment. The process repeats iteratively until all agents exit the scenario. Unlike the dataset-level evaluation, which assesses single-step decision accuracy, this experiment examines emergent system-level behavior, including safety, coordination, and operational efficiency, under realistic multi-agent interactions in unseen scenarios. Table 3 summarizes safety and performance metrics across three representative traffic scenarios depicted in Figure 3.

Across all scenarios, the pretrained Base model exhibits poor safety and reliability, with high near mid-air collision (NMAC) rates and very low success rates. Success rate is defined as the fraction of LLM agents that complete the scenario without any collision event. These results indicate that zero-shot LLM reasoning, without domain-specific alignment, is insufficient for tactical deconfliction in dense and heterogeneous airspace. In contrast, both fine-tuning strategies substantially improve safety and mission completion, confirming the necessity of domain adaptation for closed-loop deployment.

The SFT model consistently achieves the strongest overall performance across scenarios A, B, and C. It yields the lowest total NMAC rates and the highest success rates, while maintaining reasonable flight times among successful episodes. This behavior suggests that supervised alignment with human-labeled tactical decisions enables the model to internalize safety-oriented heuristics that generalize across diverse conflict geometries. Notably, SFT reduces both LLM–LLM (L–L) and LLM–Rule-based (L–R) NMACs, indicating improved coordination not only among learning agents but also in mixed-policy environments.

The GRPO model demonstrates intermediate performance, consistently improving over the Base model but falling short of SFT in overall safety and reliability. While GRPO reduces NMAC rates and increases success rates relative to the pretrained baseline, its performance varies more strongly across scenarios. In particular, GRPO achieves the lowest L–L NMAC rate in Scenario C, suggesting that preference-based optimization can enhance coordination among LLM agents in dense traffic. However, this benefit is accompanied by higher L–R NMAC rates and lower success rates compared to SFT, highlighting a trade-off between relative coordination and global safety consistency.

Flight time analysis further illustrates this trade-off. The Base model's shorter average flight times primarily reflect early episode termination due to NMACs. In contrast, the longer flight times observed for SFT and GRPO correspond to successful mission completion and more conservative deconfliction behavior. Among methods achieving comparable success rates, SFT attains the lowest average flight time, indicating a favorable balance between safety and operational efficiency.

The BlueSky evaluation demonstrates that supervised fine-tuning with human-aligned labels yields the most consistent and reliable closed-loop behavior across heterogeneous scenarios. Preference-based optimization via GRPO offers complementary benefits in specific coordination settings but exhibits reduced robustness under mixed-policy interactions. These results reinforce the importance of human-aligned supervision for deploying LLM-based

Table 3. Safety and efficiency across configurations and LLM models (mean ± std, for 10 episodes). Rates are NMACs/episode. Abbreviations: *L–L* = NMACs between two LLM agents; *L–R* = NMACs between an LLM agent and a Rule-based agent; *All* = *L–L* + *L–R*; *SR* = success rate of LLM agents (fraction completing without NMACs); *Time* = average flight time of successful LLM agents. **Bold** indicates best values: lowest NMAC for (All, L–L, L–R), highest SR, and lowest Time among methods with SR $\geq 0.9\times$ SR_{best} for the scenario.

	Base					**SFT**					**GRPO**				
Scen.	All	L–L	L–R	SR	Time	All	L–L	L–R	SR	Time	All	L–L	L–R	SR	Time
A	3.5±1.1	2.7±0.5	0.8±1.0	0.12±0.38	3.7±4.1	**1.0±0.8**	**0.7±0.7**	**0.3±0.5**	**0.77±0.29**	**5.7±0.6**	1.7±0.5	1.1±0.7	0.6±0.5	0.57±0.31	5.2±0.1
B	3.4±1.1	1.8±1.4	1.6±0.9	0.20±0.56	3.3±7.0	**1.9±1.2**	**0.9±0.7**	**1.0±0.8**	**0.62±0.36**	**8.1±0.9**	3.0±0.7	1.3±0.4	1.7±0.5	0.27±0.22	6.9±0.3
C	4.0±0.9	2.5±1.2	1.5±1.2	0.05±0.58	1.6±5.0	**1.9±0.7**	0.8±0.6	**1.1±0.9**	**0.52±0.37**	**7.5±0.7**	2.3±0.8	**0.6±0.5**	1.7±0.7	0.42±0.29	6.6±0.1

tactical deconfliction policies in safety-critical airspace environments.

6.3. Limitations and Broader Impacts

Despite the encouraging results, several limitations currently constrain the deployment of LLM-based policies in real-time coordinated multi-agent sUAS operations. A primary challenge is inference latency. Even under optimized inference settings, the Base model requires approximately 0.2 seconds to generate a short response for a single agent, with latency scaling linearly with the number of agents and output length. This overhead limits scalability in dense traffic scenarios and restricts the use of more computationally intensive reasoning techniques, such as chain-of-thought prompting or retrieval-augmented generation, which could otherwise enhance decision transparency.

A second limitation concerns prompt sensitivity and stability. LLM behavior is highly dependent on prompt structure, and deviations between the formats used during fine-tuning and inference can lead to degraded performance or partial reversion to pretrained behavior. While structured prompt design mitigates this effect, longer and more descriptive prompts further increase inference time, introducing a trade-off between reasoning richness and real-time responsiveness.

Moreover, reinforcement-based fine-tuning introduces practical constraints. GRPO requires sampling multiple candidate responses per query to estimate relative advantages, resulting in significant computational and memory demands. In this study, hardware limitations constrained the number of sampled responses, likely reducing exploration diversity and training stability. Scaling preference-based optimization for large LLMs therefore remains an open challenge requiring more efficient training strategies and distributed infrastructure.

Despite these constraints, this work demonstrates the potential of large language models to support human-aligned and interpretable decision-making in autonomous air traffic coordination, particularly in heterogeneous and dynamic environments. At the same time, the identified limitations underscore the need for careful system-level integration, emphasizing latency-aware design, resource efficiency, and safety assurance. From a broader perspective, the computational cost associated with large-scale fine-tuning motivates continued exploration of lightweight architectures and hybrid symbolic–neural approaches. This study contributes to a growing body of evidence that LLMs can augment, but not yet replace, established decision-making frameworks in real-time, safety-critical applications such as aircraft tactical deconfliction.

7. Conclusion

This study examined fine-tuned Large Language Models (LLMs) as high-level decision-making policies for tactical deconfliction in dense, heterogeneous, cooperative multi-agent air traffic environments.

By introducing a "simulation-to-language" dataset generation pipeline grounded in interpretable rule-based human decision heuristics, we showed that LLMs, specifically Qwen-Math-7B, can acquire structured, safety-oriented reasoning capabilities for sUAS tactical deconfliction. Using this dataset, we evaluated two complementary parameter-efficient alignment strategies: Supervised Fine-Tuning (SFT) and Group-Relative Policy Optimization (GRPO).

Evaluations on held-out datasets and closed-loop BlueSky simulations demonstrate that SFT provides the most consistent improvements over the baseline LLM in decision accuracy, behavioral stability, and separation safety relative to the pretrained baseline. In contrast, GRPO enables preference-based refinement that improves coordination among LLM agents in certain traffic configurations but exhibits reduced robustness in mixed-policy environments. Despite these advances, challenges remain, including inference latency, sensitivity to prompt structure, and the computational demands of reinforcement-style fine-tuning. Overcoming these constraints will be essential for deploying LLM-based tactical deconfliction policies to real-time, large-scale sUAS operations.

Future work should further benchmark LLM-based approaches against established rule-based and reinforcement-learning-based methods.

References

[1] Alibaba Group AI Team. Qwen-Math: Mathematical Rea-

soning Models from Alibaba Cloud AI. Technical report, Alibaba Group, 2024. 5

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, 2016. 2

[3] Justas Andriuškevičius and Junzi Sun. Automatic Control With Human-Like Reasoning: Exploring Language Model Embodied Air Traffic Agents. In *14th SESAR Innovation Days, SIDS 2024*, 2024. 2

[4] Yuhang Bai, Zhihong Deng, Wei Liu, et al. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*, 2023. 5

[5] Marc Brittain and Peng Wei. Autonomous separation assurance in an high-density en route sector: A deep multi-agent reinforcement learning approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3256–3262, 2019. 2

[6] Fabio Suim Chagas, Neno Ruseno, and Aurilla Aurelie Arntzen Bechina. Artificial Intelligence Approaches for UAV Deconfliction: A Comparative Review and Framework Proposal. *Automation*, 6(4), 2025. 1

[7] Long Cheng, Bowen Zhou, and Xinyi Zhang. From Language to Action: A Review of Large Language Models as Autonomous Agents and Tool Users. *Artificial Intelligence Review*, 59:71, 2026. 1

[8] Stijn Van Dam, Max Mulder, and René Paassen. The Use of Intent Information in an Airborne Self-Separation Assistance Display Design. In *AIAA Guidance, Navigation, and Control Conference*, 2009. 1

[9] Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 1

[10] Federal Aviation Administration. FAA Makes Drone History in Dallas Area, 2024. 1

[11] Federal Aviation Administration. Amazon Prime Air Amendment to Operations Specifications (OpSpecs). Technical report, U.S. Department of Transportation, 2025. 4

[12] Kathleen Ge and William Coupe. Aviation-Specific Large Language Model Fine-Tuning and LLM-as-a-Judge Evaluation. In *AIAA AVIATION FORUM AND ASCEND 2025*, page 3712, 2025. 2

[13] Dewi Gould, George De Ath, Ben Carvell, and Nick Pepper. AirTrafficGen: Configurable Air Traffic Scenario Generation with Large Language Models. *ArXiv*, abs/2508.02269, 2025. 2

[14] Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The Order Effect: Investigating Prompt Sensitivity to Input Order in LLMs. *arXiv preprint arXiv:2502.04134*, 2025. 1

[15] Jacco Hoekstra and Joost Ellerbroek. BlueSky ATC Simulator Project: an Open Data and Open Source Approach. 2016. 2, 3

[16] J.M Hoekstra, R.N.H.W van Gent, and R.C.J Ruigrok. Designing for safety: the 'free flight' air traffic management concept. *Reliability Engineering & System Safety*, 75(2): 215–232, 2002. 1

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[18] Shima Imani, Liang Du, and Harsh Shrivastava. MathPrompter: Mathematical Reasoning using Large Language Models, 2023. 1

[19] Hantao Jiang et al. Training Large Language Models on Narrow Tasks Can Lead to Broad Misalignment. *Nature*, 649: 584–589, 2026. 1

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 1

[21] Yucheng Liu. Large language models for air transportation: A critical review. *Journal of the Air Transport Research Society*, 2:100024, 2024. 2

[22] Yanchao Liu and Timothy C. Henderson. Strategic Deconfliction of Unmanned Aircraft Based on Hexagonal Tessellation and Integer Programming. *Journal of Guidance, Control, and Dynamics*, 46(8):1–14, 2023. 1

[23] Shayne Loft, Penelope Sanderson, Andrew Neal, and Mark Mooij. Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications. *Human Factors*, 49(3):376–399, 2007. 1

[24] Francesco Manigrasso, Stefan Schouten, Lia Morra, and Peter Bloem. Probing LLMs for Logical Reasoning. In *Neural-Symbolic Learning and Reasoning: 18th International Conference, NeSy 2024, Proceedings, Part I*, page 257–278, Berlin, Heidelberg, 2024. Springer-Verlag. 1

[25] Y. L. Marquand. FAA Authorises Zipline and Wing for BVLOS Operations in Dallas, 2024. 1

[26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 1

[27] Bizhao Pang, Kin Huat Low, and Chen Lv. Adaptive conflict resolution for multi-UAV 4D routes optimization using stochastic fractal search algorithm. *Transportation Research Part C: Emerging Technologies*, 139:103666, 2022. 1

[28] Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2410.13848*, 2024. 5

[29] Marta Ribeiro, Joost Ellerbroek, and Jacco Hoekstra. Review of Conflict Resolution Methods for Manned and Unmanned Aviation. *Aerospace*, 7(6):79, 2020. 1

[30] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*, 2024. 1

[31] Liya Wang, Jason Chou, Xin Zhou, Alex Tien, and Diane M. Baumgartner. AviationGPT: A Large Language Model for the Aviation Domain. *ArXiv*, abs/2311.17686, 2023. 2

[32] Wing. Meet the drones taking delivery to new heights. https://wing.com/technology, 2024. Accessed: January 2026. 4

[33] Liangqi Yuan, Chuhao Deng, Dong-Jun Han, Inseok Hwang, Sabine Brunswicker, and Christopher G. Brinton. Next-Generation LLM for UAV: From Natural Language to Autonomous Flight. *arXiv preprint arXiv:2510.21739*, 2025. 1