Hierarchical Grammar-informed Optimization for Domain-Specific Language Models

Anonymous ACL submission

Abstract

LLMs struggle with generating structured responses compliant with standardized protocols and structured phraseology in specialized safety critical domains (e.g. standardized air traffic control communication, maritime communication, railway communication etc.). In such settings, LLMs not only have to comply with the use of a specialized vocabulary but also need to generate natural language responses that follow formal grammatical rules. Domain-specific supervised fine-tuning (SFT) of language models does not provide a remedy to this challenge, and fails to establish a formal compliance bound. Further, inference-time grammar-constrained decoding also has its drawback: It distorts the model's output distribution, potentially lowering natural language quality, and requires a complete context-free grammar (CFG) at every decoding step. However, domain-specific communication protocols may not always require or align with such context-free production rules. To overcome these challenges, we introduce a novel hierarchical grammar-informed optimization framework that regularizes a model's output distribution based on a constraint set derived from a standardized domain-specific vocabulary, phraseology and CFG. This framework **SCOPE** (**S**tructured-**C**onstraint **OP**timization for Grammar-informed g**E**neration) regulates compliance at three levels jointly: lexical, phraseological, and syntactical. Theoretically, SCOPE admits a *formal vocabulary compliance bound* at optimality, and employs a Group Relative Policy Optimization (GRPO) to enforce phraseological and syntactical compliance. We evaluate SCOPE across three open-weight models (Llama-3.1-8B, Qwen3-8B, GPT-2 Large) on Air Traffic Control communication (ICAO Doc. 4444; FAA JO 7110.65BB), with cross-domain transfer to maritime communication (IMO SMCP). Extensive experiments show that SCOPE outperforms instruction-tuned SFT, direct preference optimization (DPO), and grammar-constrained decoding (GCD) baselines across its three-level compliance, with each hierarchical level confirmed to contribute incrementally. We release all code, data, and trained models publicly.[1]

1 Introduction

LLMs struggle with generating structured responses compliant with standardized protocols and structured phraseology in specialized safety critical domains. In such settings, LLMs not only have to comply with the use of a specialized vocabulary but also need to generate responses that follow formal grammatical rules. For example, in the context of air traffic communication between pilots and air traffic controllers as well as operators of autonomous drones, strict adherence to the radiotelephony phraseology and syntactic rules standardized by the International Civil Aviation Organization (ICAO) is required (International Civil Aviation Organization, 2007; SKYbrary Aviation Safety, 2025). Such compliance implies more than just consistency with a specific set of words: According to ICAO, a phraseology refers to the standardized set of words and expressions to ensure communication is "clear, concise, unambiguous, and internationally consistent" (International Civil Aviation Organization, 2007). The failure to adhere to such standards is a causal factor of incidents and accidents (International Civil Aviation Organization, 2007; International Civil Aviation Organisation, 2016; International Civil Aviation Organization, 2010). Recent events like the LaGuardia accident (Sisak et al., 2026) illustrate this. ATC is not the only domain were such standards matter: Similiar standards have been developed for marine communication (Standard Marine Communication Phrases (SMCP) (IMO SMCP (International Maritime Organization, 2001))), railway (EIRENE (European

[1] https://anonymous.4open.science/r/ATC_Chatbot-6F51/README.md

Railway Agency, 2019)), emergency response/firefighting (EMS), healthcare and software development (e.g. DevOps) and adherence to such domain-specific grammars is essential for LLMs to assist or automate communication.

Researchers and industry experts tackle this challenge in multiple ways. The most widely adopted approach is to opt for standard fine-tuning (Kumar et al., 2025; Gururangan et al., 2020) of pre-trained models using domain-specific datasets for domain-specific language generation tasks using a cross-entropy loss. In addition, researchers also develop advanced grammar-informed prompting methods (Wang et al., 2023) However, both approaches fail to ensure that the generated text (the next tokens) appear in any constraint set. To ensure a more rigid enforcement, researchers have turned to formal grammars, such as context-free grammars (CFG), and propose grammar-constrained (or aligned decoding) that mask or filter outputs. Such inference-time constrained decoding such as manipulating beam search (Hokamp and Liu, 2017; Post and Vilar, 2018) or applying grammar masks during token generation (Geng et al., 2023; Park et al., 2024, 2025; Banerjee et al., 2025) can enforce hard constraints. However, such approaches have significant limitations: They complete formal context free grammar but cannot handle situations where the grammar is domain specific (rather than context-free) and production rules are probabilistic rather than deterministic. Further, they distorts the model's learned distribution (Park et al., 2024), and also add per-step parser overhead infeasible in edge deployment.

To address this gap, we introduce **SCOPE** (**S**tructured-**C**onstraint **OP**timization for Grammar-informed g**E**neration), described in Figure 1, a hierarchical multi-objective optimization framework for grammar-informed language generation. SCOPE advances standard cross-entropy training with three grammar regularization objectives: $\mathcal{L}_{\text{tok}}$ for lexical compliance, $\mathcal{L}_{\text{phr}}^{\text{GRPO}}$ for phrase-level compliance, and $\mathcal{L}_{\text{cfg}}^{\text{GRPO}}$ for CFG-informed syntactic compliance with a domain-specific grammar. The framework applies to domains where structured constraint sets $(\mathcal{V}, \mathcal{P}, \mathcal{G})$ can be defined and introduces negligible training overhead.

We evaluate SCOPE across three open-weight models (774M–8B) Llama-3-8B, Qwen-3-8B and GPT-2 Large-774M on Air Traffic Control (ATC) communication and maritime VHF communication (IMO SMCP, see Section B) to provide empirical evidence for cross-domain support . SCOPE achieves consistent gains on all three compliance metrics (C_{tok}, C_{phr}, C_{cfg}) over instruction-tuned SFT, DPO, and inference-time GCD baselines at every scale while preserving semantics, with ablation confirming each hierarchical level contributes independently.

Key Contributions. The key contributions of our paper are:

- **SCOPE framework.** The developed **S**tructured-**C**onstraint **OP**timization for Grammar-informed g**E**neration is a new multi-objective optimization framework that constraints text generation through a generalizable framework for domain-specific communication rules and standards while being consistent with theory of formal grammar (e.g. chomsky's formal CFG).
- **Formal compliance bound.** We prove that the grammar-regularised objective yields a hard bound on expected out-of-whitelist generation probability
- **Grammar Artefacts - ATC/Maritime CFG.** We construct and release $\mathcal{V}_{\text{ATC}}$ (vocabulary whitelist), $\mathcal{P}_{\text{ATC}}$ (phrase whitelist), and a CFG-grounded $\mathcal{G}_{\text{ATC}}$ based on ICAO Doc 4444 and FAA JO 7110.65, as well as $\mathcal{V}_{\text{maritime}}$ (vocabulary whitelist), $\mathcal{P}_{\text{maritime}}$ (phrase whitelist), and a CFG-grounded $\mathcal{G}_{\text{maritime}}$ based on IMO SMCP.
- **Evaluation metrics.** In addition to established metrics for LLM evaluation, we introduce a three-metric compliance hierarchy.
- **Empirical evaluation.** Across three open-weight models (SLM 774M to LLM 8B), compared against instruction-tuned SFT, DPO, inference-time GCD (Park et al., 2025; Banerjee et al., 2025), SCOPE achieves the highest compliance at every scale across all three metrics.
- **Preliminary cross-domain validation.** We instantiate SCOPE's hierarchical grammar regularisation on maritime VHF communication (IMO SMCP (International Maritime Organization, 2001)), constructing $\mathcal{V}_{\text{maritime}}$, $\mathcal{P}_{\text{maritime}}$, and $\mathcal{G}_{\text{maritime}}$ from the SMCP regulatory document. Positive gains on C_{tok} and C_{cfg} provide empirical cross-domain generalization.

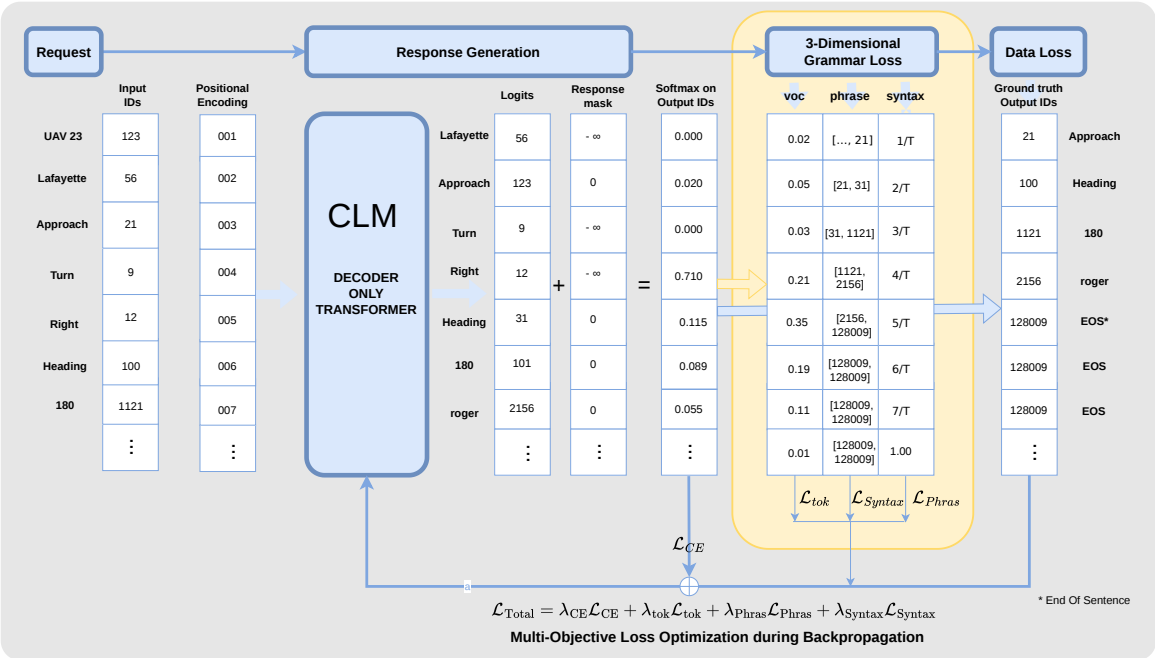

Figure 1: Overview of the proposed SCOPE framework for grammar-informed language generation. During training, a decoder-only transformer is jointly optimized using standard cross-entropy loss together with hierarchical grammar regularization objectives operating at the vocabulary, phrase, and syntax levels.

2 SCOPE: Hierarchical Grammar-Informed Multi-Objective Optimization Framework

SCOPE operates through hierarchical grammar regularization at three complementary levels: *lexical*, *phraseological*, and *syntactic* (Fig. 1). The framework applies wherever a target domain can be specified through a vocabulary whitelist $\mathcal{V}$, phraseological constraints $\mathcal{P}$, and a domain-specific context-free grammar $\mathcal{G}$, in sublanguages and domains like aviation, maritime or railway. Unlike inference-time constrained decoding approaches (Park et al., 2025; Banerjee et al., 2025), which enforce constraints post hoc, SCOPE jointly optimizes all three compliance levels during training through co-designed grammar regularization objectives.

2.1 Preliminaries: Formal and Context-free Grammar

A context-free grammar (Chomsky, 1965; Wang et al., 2023) is a four-tuple (N, Σ, R, S) of non-terminals, terminals, productions, and a start symbol, with productions of the form $A \rightarrow \alpha$ that rewrite a single non-terminal A as a string α over the combined alphabet. CFGs sit at the second level of the Chomsky hierarchy, expressive enough to capture nested structure but tractable enough to admit polynomial-time parsing. Importantly, many core syntactic structures of natural language, including hierarchical phrase structure, recursion, and nested clause composition, can be approximated effectively using CFGs. The formal definition of our proposed CFG for ATC and Maritime Communication are presented in the Annex (see Section A and B)

2.2 Domain Compliance Formalisation

In structured-vocabulary domains, an output sequence $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ must satisfy three nested conditions to be compliant: every token must belong to the domain vocabulary $\mathcal{V}$; every local subsequence must match an approved phraseological pattern in $\mathcal{P}$; and the full sequence must be a well-formed utterance under the domain grammar $\mathcal{G}$. We formalise each condition as a scalar score in $[0, 1]$.

2.2.1 Token Whitelist $\mathcal{V}$ and Lexical Compliance

The vocabulary $\mathcal{V}$ is the set of all word types sanctioned for the domain. It is constructed via a **specification-first, corpus-validated** procedure: the authoritative domain specification (a regulatory standard, a language reference, a schema defini-

tion) provides the initial candidate set; corpus occurrences below a frequency threshold k are then discarded as operationally obsolete or corpus artefacts. This ordering makes $\mathcal{V}$ a compliance artefact rather than a domain vocabulary list: the specification defines what is permissible; the corpus confirms what is operationally current.

Definition 1 (Token compliance)**.** *Given a generated sequence $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ and a domain vocabulary $\mathcal{V}$, the token compliance score is:*

$$C_{tok}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \in \mathcal{V}]. \tag{1}$$

$C_{\text{tok}} = 1.0$ is necessary but not sufficient for compliance: a sequence consisting entirely of valid tokens in an invalid arrangement—such as *"descend climb altitude heading"* in ATC, or `return def if` in Python—violates domain structure despite perfect token membership. The phrase-level constraint in Section 2.2.2 addresses this gap.

2.2.2 Phrase Whitelist $\mathcal{P}$ and Phraseological Compliance

The phrase whitelist $\mathcal{P}$ captures valid local token combinations at multiple n-gram orders:

$$\mathcal{P} = \mathcal{P}^{(2)} \cup \mathcal{P}^{(3)} \cup \mathcal{P}^{(4)}, \tag{2}$$

where $\mathcal{P}^{(n)}$ is the set of domain-sanctioned n-grams. Construction parallels $\mathcal{V}$: the authoritative specification provides canonical phrase templates; corpus-mined n-grams below threshold k are discarded; and corpus entries not traceable to a specification template are removed to prevent corpus-specific patterns from entering $\mathcal{P}$ as pseudo-valid phraseology.

Definition 2 (Phrase compliance)**.** *The phrase compliance score over order n is:*

$$C_{phr}^{(n)}(\hat{y}) = \frac{1}{T-n+1} \sum_{t=n}^{T} \mathbb{1}[(\hat{y}_{t-n+1}, \ldots, \hat{y}_t) \in \mathcal{P}^{(n)}]. \tag{3}$$

The aggregate phrase compliance averages uniformly across orders:

$$C_{phr}(\hat{y}) = \frac{1}{3} \sum_{n \in \{2,3,4\}} C_{phr}^{(n)}(\hat{y}). \tag{4}$$

2.2.3 Context-Free Grammar $\mathcal{G}$ and Syntactic Compliance

The grammar $\mathcal{G}$ encodes utterance-level structural constraints as a CFG. Production rules are derived from the authoritative domain specification; regex terminals handle open-ended identifiers and numerical ranges without requiring full terminal enumeration, making partial CFG specification tractable for domains with variable-length identifiers.

Definition 3 (Grammar compliance)**.** *Let $\mathcal{L}(\mathcal{G})$ denote the language of $\mathcal{G}$ and $\mathcal{L}_{\text{prefix}}(\mathcal{G})$ its prefix language. The grammar compliance score is:*

$$C_{cfg}(\hat{y}) = \begin{cases} 1 & \text{if } \hat{y} \in \mathcal{L}(\mathcal{G}), \\ \max_k \frac{k}{T} \text{ s.t. } \hat{y}_{1:k} \in \mathcal{L}_{\text{prefix}}(\mathcal{G}) & \text{otherwise.} \end{cases} \tag{5}$$

Partial credit rewards sequences that begin correctly but fail at a single non-compliant token, providing a graded evaluation signal rather than a binary pass/fail. The aggregate compliance score used for checkpoint selection during training is:

$$\bar{C}(\hat{y}) = \tfrac{1}{3}\big(C_{\text{tok}}(\hat{y}) + C_{\text{phr}}(\hat{y}) + C_{\text{cfg}}(\hat{y})\big). \tag{6}$$

2.3 Hierarchical Multi-objective Losses

2.3.1 Level 1: Lexical Constraint Loss ($\mathcal{L}_{\text{tok}}$)

The base optimization objective is cross-entropy over response tokens:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathbf{1}_{\text{resp}}|} \sum_{t \in \mathbf{1}_{\text{resp}}} \log P_\theta(\hat{y}_t^* \mid \hat{y}_{<t}^*, x), \tag{7}$$

where $\mathbf{1}_{\text{resp}}$ is a binary mask selecting response tokens only, excluding the system prompt and instruction header positions. $\mathcal{L}_{\text{CE}}$ imposes no vocabulary constraint: the model can satisfy it by producing fluent, on-topic text in non-standard phrasing, a failure mode acute for large pre-trained models whose broad prior vocabulary resists lexical specialisation through data alone.

We introduce a differentiable **lexical constraint loss** $\mathcal{L}_{\text{tok}}$ that directly penalises the probability mass the model assigns to out-of-whitelist tokens at each response step. Let $p_t(v)$ be the softmax probability of token v at step t. Then:

$$\mathcal{L}_{\text{tok}} = \frac{1}{|\mathbf{1}_{\text{resp}}|} \sum_{t \in \mathbf{1}_{\text{resp}}} \sum_{v \notin \mathcal{V}(s_t)} p_t(v), \tag{8}$$

where $\mathcal{V}(s_t) \subseteq \mathcal{V}$ is a state-conditional whitelist determined by the current discourse state s_t. Restricting the sum to $\mathbf{1}_{\text{resp}}$ positions ensures that system prompt tokens—where domain vocabulary need not apply—do not contribute to the compliance gradient. When $\mathcal{V}(s_t) = \mathcal{V}$ for all s_t, (8) recovers the global whitelist penalty.

A key insight motivates this formulation: vocabulary compliance cannot be enforced by penalising sampled token *identities*, since discrete sampling is not differentiable with respect to model parameters. $\mathcal{L}_{\text{tok}}$ instead penalises *probability mass* over the full softmax distribution—a continuous function of θ—making training-time compliance enforcement tractable without approximation, and admitting the formal bound below.

Connectives and shared vocabulary. $\mathcal{V}$ includes tokens shared between the domain and general language— connectives, directional modifiers, and numeric words. $\mathcal{L}_{\text{tok}}$ does not penalise these regardless of context; their positional validity is the responsibility of $\mathcal{L}_{\text{phr}}$. This is by design: the two losses are co-designed, and this shared-token limitation of $\mathcal{L}_{\text{tok}}$ is the principal motivation for requiring both components jointly.

Proposition 1 (Compliance bound)**.** *Let θ^* minimise $\mathcal{L} = \mathcal{L}_{CE} + \lambda_{\text{tok}}\mathcal{L}_{tok} + \lambda_{\text{phr}}\mathcal{L}_{phr} + \lambda_{\text{cfg}}\mathcal{L}_{cfg}$ with $\lambda_{\text{tok}} > 0$. Then:*

$$\mathbb{E}_t\left[\sum_{v \notin \mathcal{V}} p_t(v)\right] \leq \frac{\mathcal{L}_{CE}(\theta^*)}{\lambda_{\text{tok}}}. \tag{9}$$

In contrast, any minimiser of $\mathcal{L}_{CE}$ alone satisfies no such bound.

Proof sketch. At optimality, any further reduction of $\mathcal{L}_{\text{tok}}$ below $\mathcal{L}_{\text{CE}}(\theta^*)/\lambda_{\text{tok}}$ would allow a compensating increase in $\mathcal{L}_{\text{CE}}$ while reducing total loss, contradicting optimality. The strict lower bound on $\mathcal{L}_{\text{CE}}$ follows from non-whitelist tokens present in reference sequences. Full proof in Section C. □

2.3.2 Level 2: Phraseological Constraint Loss ($\mathcal{L}_{\text{phr}}$)

Lexical compliance does not constrain token order: a sequence of valid tokens in an invalid arrangement achieves $C_{\text{tok}} = 1.0$ while violating domain structure. We address this with a phrase-level reward signal optimised via GRPO (Shao et al., 2024).

Reward and estimator. We define $R_{\text{phr}}(\hat{y}) = C_{\text{phr}}(\hat{y}) \in [0, 1]$ over a fully decoded sequence. Because R_{phr} depends on discrete token samples it is non-differentiable; we optimise it with GRPO, which draws M sequences from the model and normalises rewards within the group to produce low-variance advantage estimates without a separate baseline model or value function. Given M samples $\{\hat{y}^{(m)}\}_{m=1}^{M} \sim P_\theta(\cdot \mid x)$, the group-normalised advantage is:

$$A^{(m)} = \frac{R_{\text{phr}}(\hat{y}^{(m)}) - \mu_R}{\sigma_R + \varepsilon}, \tag{10}$$

where

$$\mu_R = \tfrac{1}{M}\textstyle\sum_m R_{\text{phr}}(\hat{y}^{(m)}), \qquad \sigma_R = \text{std}\{R_{\text{phr}}(\hat{y}^{(m)})\},$$

yielding the gradient:

$$\nabla_\theta \mathcal{L}_{\text{phr}} \approx -\frac{1}{M}\sum_{m=1}^{M} A^{(m)} \cdot \nabla_\theta \log P_\theta(\hat{y}^{(m)} \mid x). \tag{11}$$

GRPO is preferred over Proximal Policy Optimization (PPO) and REINFORCE with a self-critical baseline for two reasons: (i) it does not require an additional greedy decoding pass, halving inference overhead during training; and (ii) group normalisation produces non-trivial advantages even when all samples score similarly, maintaining a meaningful gradient signal throughout training rather than collapsing to near-zero updates in low-variance reward regimes. Gradient clipping at norm 1.0 stabilises training. Full derivation is in Section C.

2.3.3 Level 3: Syntactic Structural Loss ($\mathcal{L}_{\text{cfg}}$)

A sequence of valid n-grams can still form an invalid utterance structure: a required opening element missing, clauses in the wrong order, or a mandatory closing token absent. The grammar loss $\mathcal{L}_{\text{cfg}}$ enforces utterance-level structural validity.

Soft partial-parse reward. Binary grammar validity is sparse during early training when the model rarely produces fully grammatical outputs. We therefore use the soft partial-parse score $R_{\text{cfg}}(\hat{y}) = C_{\text{cfg}}(\hat{y}) \in [0, 1]$ defined in (5), which provides graded credit for partially-valid sequences and stabilises GRPO before full parses become common.

To illustrate with examples from each domain. In ATC: the response *"TURN LEFT HEADING THREE THREE ZERO DELTA SIX FORTY EIGHT"* (9 words) parses fully under $\mathcal{G}_{\text{ATC}}$ as a heading instruction with callsign ($R_{\text{cfg}} = 1.0$); appending the out-of-grammar continuation *"PLEASE ADVISE"* fails at *"PLEASE"*, so the longest valid prefix is the 9-word instruction, giving $R_{\text{cfg}} = 9/11 \approx 0.82$. In maritime: the position report *"VESSEL MERIDIAN STAR POSITION TWO FOUR NORTH ZERO ONE EIGHT WEST"* (10 words) parses

fully ($R_{\text{cfg}} = 1.0$); appending *"CURRENTLY ANCHORED"* (not a valid SMCP continuation of a position report) fails immediately, yielding $R_{\text{cfg}} = 10/12 \approx 0.83$.

GRPO estimator. $\mathcal{L}_{\text{cfg}}$ uses the identical GRPO estimator as $\mathcal{L}_{\text{phr}}$ with reward R_{cfg}:

$$\nabla_\theta \mathcal{L}_{\text{cfg}} \approx -\frac{1}{M} \sum_{m=1}^{M} A_{\text{cfg}}^{(m)} \cdot \nabla_\theta \log P_\theta(\hat{y}^{(m)} \mid x), \tag{12}$$

where

$$A_{\text{cfg}}^{(m)} = \frac{R_{\text{cfg}}(\hat{y}^{(m)}) - \mu_{\text{cfg}}}{\sigma_{\text{cfg}} + \varepsilon}.$$

The same M sample group drawn for $\mathcal{L}_{\text{phr}}$ is reused for $\mathcal{L}_{\text{cfg}}$, requiring only one additional parser call per sample with no extra forward passes.

2.4 Combined Multi-Objective

The full SCOPE training objective combines all three compliance levels with the base cross-entropy:

$$\boxed{\mathcal{L}_{\text{SCOPE}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{tok}}\, \mathcal{L}_{\text{tok}} + \lambda_{\text{phr}}\, \mathcal{L}_{\text{phr}} + \lambda_{\text{cfg}}\, \mathcal{L}_{\text{cfg}},} \tag{13}$$

where $\boldsymbol{\lambda} = (\lambda_{\text{tok}}, \lambda_{\text{phr}}, \lambda_{\text{cfg}}) \geq \mathbf{0}$ are scalar loss weights. Setting $\lambda_{\text{cfg}} = 0$ recovers the two-level objective; setting $\boldsymbol{\lambda} = \mathbf{0}$ recovers standard SFT. The compliance–capability trade-off is controlled through λ_{tok} (Proposition 1): larger λ_{tok} tightens the out-of-vocabulary probability bound at the cost of higher cross-entropy loss.

3 Experimental setup

3.1 Datasets

Air Traffic Control. We use the LDC Air Traffic Control corpus (Godfrey, 1994), a collection of transcribed multi-turn radio exchanges between controllers and pilots recorded at Boston TRACON. After deduplication and removal of exchanges with no recoverable reference response, we retain 1,067 input–output pairs, partitioned into 832 training, 104 validation, and 131 test examples using a fixed random split (seed = 42); no exchange appears in more than one partition. Each pair consists of an operator utterance (the incoming transmission) and the reference response (controller instruction or pilot readback). Dialogue state labels (AWAITINGREADBACK, CLEARANCEISSUED, INIT, etc. are derived from the exchange structure and encoded as a finite-state machine prefix in the conditioning prompt (see Section 2).

Maritime. Since no publicly available corpus of transcribed maritime radio exchanges exists in paired request–response format suitable for fine-tuning—real VHF recordings are either proprietary (Sun et al., 2025) or released for ASR rather than NLP (Bonalor, 2025)—we construct a synthetic dataset by prompting Claude Sonnet 4.6 (Anthropic, 2025) to generate paired exchanges grounded in the IMO Standard Marine Communication Phrases (International Maritime Organization, 2001), covering distress calls, safety broadcasts, position reports, passing agreements, and weather and pilotage instructions. Responses were filtered by running $\mathcal{G}_{\text{SMCP}}$ and discarding non-parsing examples, yielding 589 pairs (428 train / 53 val / 54 test, seed = 42). We treat maritime results as a *cross-domain transfer* experiment rather than a claim about real-world operational distributions.[2]

3.2 Models

We evaluate SCOPE across three language models spanning a range of scale and pre-training regimes:

- **Llama-3.1-8B-Instruct** (Touvron et al., 2023) (8B parameters): the primary evaluation model; an instruction-tuned model optimised for instruction following.
- **Qwen3-8B** (Qwen Team, 2025) (8B parameters): an instruction-tuned model with strong multilingual pre-training.
- **GPT-2 Large** (Radford et al., 2019) (774M parameters): a causal language model with no instruction tuning, serving as the small-scale baseline.

3.3 Training Details

All models are fine-tuned with AdamW (Loshchilov and Hutter, 2019) using a cosine learning rate schedule with linear warmup over the first 10% of total steps. Hyperparameters $(\lambda_{\text{tok}}, \lambda_{\text{phr}}, \lambda_{\text{cfg}}, \eta)$ for Llama-3.1-8B and Qwen3-8B are tuned jointly via Optuna (Akiba et al., 2019) with 50 trials, maximising $\bar{C}$ on the validation split; tuned values are $\boldsymbol{\lambda} = (0.9705, 0.9009, 1.4314)$ and $\eta = 2.48 \times 10^{-5}$. GPT-2 Large is trained at $\eta = 2 \times 10^{-4}$ in `float32`. Table 5 lists the hyperparameters shared across all conditions.

[2] Synthetic data lacks the disfluencies and non-standard readbacks present in real VHF traffic; results should be interpreted accordingly.

4 Experimental results

We report results on the ATC test set (n=131, Table 1) and the maritime SMCP test set (n=54, Table 2). Figures 2 and 3 visualise all eight metrics for each model and method. We structure the discussion around four findings that hold consistently across both domains.

4.1 Finding 1: SCOPE Achieves Best Compliance and Semantics Jointly

SCOPE achieves the best value on every metric except C_{tok} across all models and both domains. GCD achieves higher C_{tok} by applying a token-level vocabulary mask at every step, but as Section 4.3 shows, this comes at a severe cost to all other metrics.

On ATC, gains over the best baseline are large. For Llama-3.1-8B, SCOPE improves DA-F1 from 0.267 to 0.788 (+0.521) and Slot-F1 from 0.397 to 0.866 while maintaining C_{cfg}=0.999 and zero hallucination. Qwen3-8B gains ΔC_{cfg}=+0.148 over DPO and ΔDA-F1 = +0.479.

On maritime, SCOPE achieves near-ceiling semantics: Llama DA-F1 = 0.962, Slot-F1 = 0.941; Qwen3 DA-F1 = 0.999, Slot-F1 = 0.994, demonstrating successful transfer to a new regulatory domain.

4.2 Finding 2: Fine-Tuning Is a Necessary Precondition

Zero-shot models are non-compliant: Llama Vanilla has C_{cfg}=0.079 (42.0% hallucination) and GPT-2 C_{cfg}=0.003 (96.2%) on ATC. SFT closes much of this gap for 8B models (C_{cfg}=0.995 Llama, 0.780 Qwen3) but leaves semantic performance limited (Llama DA-F1 = 0.267, Slot-F1 = 0.397). DPO further improves compliance (C_{cfg}=0.976 Llama, 0.809 Qwen3) but still trails SCOPE on DA-F1 (0.211 vs 0.788 for Llama). On maritime, Qwen3 DPO achieves high semantics (DA-F1 = 0.974, Slot-F1 = 0.982) yet SCOPE matches or exceeds both (0.999, 0.994) with higher $\bar{C}$ (0.877 vs 0.585).

4.3 Finding 3: GCD Collapses Semantics on ATC, Partially Works on Maritime

ATC. GCD achieves near-perfect C_{tok} (0.999/0.984/0.987 for Llama/Qwen3/GPT-2) but C_{cfg} collapses to 0.017/0.012/0.008 and BERTScore drops 50% (e.g. 0.896 $\rightarrow$ 0.447 for Llama). DA-F1 reverts to Vanilla level (0.117); outputs such as *"ONE TO ONE AND A TWO TO THREE"* illustrate vocabulary-conformant but globally incoherent generation.

Maritime. On SMCP, GCD does not collapse: Llama C_{cfg}=0.926, Qwen3 C_{cfg}=1.000. SMCP's shorter, more templatic utterances make token-level masking less likely to derail global coherence. Nevertheless, SCOPE substantially outperforms GCD on semantic metrics (SCOPE vs GCD: DA-F1 0.962 vs 0.223, Slot-F1 0.941 vs 0.410 for Llama), confirming that training-time compliance encoding is preferable to inference-time masking in both domains.

4.4 Finding 4: SCOPE Generalises Across Model Scale and Domain

SCOPE improves all model families from GPT-2-774M to 8B instruction-tuned models. Cross-domain transfer is demonstrated by Qwen3-8B reaching DA-F1 = 0.999 on maritime with no domain-specific modification to the training procedure—only the instantiation of $\mathcal{V}$, $\mathcal{P}$, and $\mathcal{G}$ from the IMO SMCP specification. On every model–domain combination, SCOPE achieves higher semantic metrics than all baselines simultaneously with higher or equal compliance, refuting the common concern that structural constraints degrade semantic quality.

5 Related work

Constrained decoding (Hokamp and Liu, 2017; Post and Vilar, 2018; Park et al., 2025; Banerjee et al., 2025) enforces lexical or grammatical constraints at inference time via beam-search manipulation or token masking. SCOPE instead encodes constraints during training, so the model internalises domain structure rather than having it imposed externally at each decoding step; no specialised decoder is required at deployment, and the compliance gradient shapes internal representations rather than merely filtering outputs. We provide a direct empirical comparison against GCD.

Auxiliary loss training (Gururangan et al., 2020; Lee et al., 2020) adds auxiliary losses over token distributions to encourage domain adaptation. $\mathcal{L}_{\text{tok}}$ extends this pattern with a formal compliance bound (Proposition 1); $\mathcal{L}_{\text{phr}}$ and $\mathcal{L}_{\text{cfg}}$ add phraseological and syntactic levels absent from all prior auxiliary loss approaches.

Table 1: Results on the **ATC** test set (n=131). **Bold**: best among trained methods per model. Hall.% ↓; all others ↑. BERTScore computed with domain-adapted ATC encoder. 95% Wilson Confidence Intervals reported.

Model	Method	Compliance				Semantic / Safety			
		C_{tok}	C_{phr}	C_{cfg}	$\bar{C}$	Slot-F1	DA-F1	Hall.%↓	BERTScore
Llama-3.1-8B	Vanilla	0.635	0.198	0.079	0.304	0.125	0.117	42.0	0.861
	SFT	0.874	0.554	0.995	0.808	0.397	0.267	**0.0**	0.889
	DPO	0.880	0.646	0.976	0.834	0.336	0.211	**0.0**	0.896
	GCD	**0.999**	0.505	0.017	0.507	0.405	0.117	**0.0**	0.447
	SCOPE (ours)	0.988	**0.683**	**0.999**	**0.890**	**0.866**	**0.788**	**0.0**	**0.991**
Qwen3-8B	Vanilla	0.552	0.171	0.103	0.275	0.219	0.144	28.2	0.835
	SFT	0.826	0.442	0.780	0.682	0.272	0.141	4.6	0.863
	DPO	0.844	0.446	0.809	0.700	0.313	0.177	2.3	0.864
	GCD	**0.984**	0.467	0.012	0.488	0.481	0.139	**0.0**	0.437
	SCOPE (ours)	0.979	**0.568**	**0.957**	**0.835**	**0.744**	**0.656**	**0.0**	**0.989**
GPT-2-774M	Vanilla	0.556	0.203	0.003	0.254	0.025	0.130	96.2	0.847
	SFT	0.621	0.285	0.181	0.363	0.122	0.139	41.2	0.846
	DPO	0.725	0.360	0.427	0.504	0.191	0.147	13.0	0.849
	GCD	**0.987**	0.392	0.008	0.462	0.405	0.146	**0.0**	0.405
	SCOPE (ours)	0.964	**0.505**	**0.891**	**0.787**	**0.520**	**0.553**	**0.0**	**0.897**

Table 2: Results on the **Maritime / SMCP** test set (n=54). Corpus synthetically generated from IMO SMCP by Claude Sonnet 4.6 (Section 3.1). **Bold**: best among trained cmethods per model. Hall.% ↓; all others ↑. BERTScore computed with domain-adapted SMCP encoder. 95% Wilson Confidence Intervals reported.

Model	Method	Compliance				Semantic / Safety			
		C_{tok}	C_{phr}	C_{cfg}	$\bar{C}$	Slot-F1	DA-F1	Hall.%↓	BERTScore
Llama-3.1-8B	Vanilla	0.312	0.017	0.084	0.138	0.586	0.429	1.9	0.421
	SFT	0.664	0.384	0.984	0.677	0.778	0.837	**0.0**	0.596
	DPO	0.704	0.435	0.984	0.708	0.565	0.645	**0.0**	0.571
	GCD	**0.935**	0.574	0.926	0.812	0.410	0.223	**0.0**	0.545
	SCOPE (ours)	0.911	**0.667**	**0.994**	**0.857**	**0.941**	**0.962**	**0.0**	**0.923**
Qwen3-8B	Vanilla	0.300	0.021	0.053	0.125	0.540	0.415	7.4	0.412
	SFT	0.589	0.295	0.559	0.481	0.809	0.914	3.7	0.605
	DPO	0.704	0.347	0.704	0.585	0.982	0.974	1.9	0.646
	GCD	**0.975**	0.560	1.000	0.845	0.618	0.724	**0.0**	0.616
	SCOPE (ours)	0.969	**0.662**	**1.000**	**0.877**	**0.994**	**0.999**	**0.0**	**0.828**

GRPO for generation (Shao et al., 2024) uses group-normalised advantages to fine-tune generation models with task-specific scalar rewards. SCOPE applies GRPO with rewards derived directly from formal compliance specifications—to our knowledge the first use of GRPO with a specification-grounded reward signal for domain-constrained generation.

6 Conclusion

We introduced **SCOPE**, a hierarchical multi-objective optimization framework that jointly enforces lexical, phraseological, and syntactic compliance during training. Three co-designed losses — a differentiable lexical constraint loss ($\mathcal{L}_{\text{tok}}$), a GRPO-based phraseological loss ($\mathcal{L}_{\text{phr}}$), and a GRPO-based syntactic loss ($\mathcal{L}_{\text{cfg}}$) enable language models to produce outputs that are simultaneously compliant with regulatory specifications and semantically coherent.

Evaluated on two regulatory communication domains (ATC and Maritime SMCP) across three model families, SCOPE achieves the best result on every compliance and semantic metric except C_{tok}, where grammar-constrained decoding (GCD) achieves a higher value by token masking alone. That single exception exposes the central weakness of inference-time approaches: GCD's near-perfect C_{tok} compliance comes at the cost of semantic mode collapse.

SCOPE is domain-agnostic by construction. The same training procedure, applied to a vocabulary $\mathcal{V}$, phrase whitelist $\mathcal{P}$, and partial grammar $\mathcal{G}$ derived from any authoritative specification, produces a compliant model for that domain. Applications beyond ATC and maritime include railway dispatch communications, structured command languages, and formal output schemas such as SQL or JSON, all of which share the defining property that deviation from the specification constitutes a functional or safety violation rather than a stylistic one.

Limitations

Automated metrics only. All evaluation is automated. No human evaluation with certified ATC controllers or maritime radio officers has been conducted. Automated metrics — particularly DA-F1 and Slot-F1 — measure necessary but not sufficient conditions for operational acceptability. Human evaluation under realistic simulation conditions is required before any safety claim about real-world deployment can be substantiated, which is direction for future work.

Synthetic maritime corpus. The SMCP corpus was generated by Claude Sonnet 4.6 and grammar-filtered; it does not contain the disfluencies, non-standard readbacks, callsign corruptions, and code-switching present in real maritime VHF traffic. Maritime results should be interpreted as evidence that SCOPE instantiates correctly on a new specification rather than as a claim about operational maritime communication performance. Validation on real maritime recordings (e.g. VHF-Dial (Sun et al., 2025)) is required before deployment claims can be made.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of KDD 2019*, pages 2623–2631.

Anthropic. 2025. Claude sonnet 4.6. Anthropic API.

Debangshu Banerjee, Tarun Suresh, Shubham Ugare, and 1 others. 2025. CRANE: Reasoning with constrained LLM generation. In *Proceedings of the 42nd International Conference on Machine Learning*.

Bonalor. 2025. MARTTS: A synthetic maritime VHF radiotelephony speech corpus. Hugging Face Datasets. 240 synthetic multi-speaker maritime dialogues for ASR stress-testing; not formatted as paired request–response exchanges for NLP fine-tuning.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of ICML 2018*, pages 794–803.

Noam Chomsky. 1965. *Aspects of the theory of syntax*, 20. print edition. Number 11 in Special technical report / Massachusetts Institute of Technology, Research Laboratory of Electronics. MIT Press, Cambridge, Mass.

European Railway Agency. 2019. EIRENE functional requirements specification. Technical Report ERA/EIRENE/FRS v16.0, European Union Agency for Railways, Valenciennes, France.

Federal Aviation Administration. 2023. Air traffic control, JO 7110.65. Technical report, FAA.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of EMNLP*, pages 10932–10952.

John J. Godfrey. 1994. Air Traffic Control Complete. Linguistic Data Consortium. LDC Catalog No. LDC94S14A.

Suchin Gururangan, Ana Marasovic, Swabha Swayamditta, Kyle Lo, Iz Beltagy, Luke Zettlemoyer, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of ACL*, pages 1535–1546.

International Civil Aviation Organisation. 2016. Procedures for air navigation services — air traffic management (PANS-ATM), Doc 4444. Technical report, ICAO.

International Civil Aviation Organization. 2007. Manual of radiotelephony.

International Civil Aviation Organization. 2010. Manual on the implementation of ICAO language proficiency requirements. Technical Report Doc 9835 AN/453, ICAO, Montreal, Canada.

International Maritime Organization. 2001. Standard marine communication phrases (SMCP). Technical Report Resolution A.918(22), IMO, London, UK.

Aman Kumar, Ekant Muljibhai Amin, Xian Yeow Lee, Lasitha Vidyaratne, Ahmed K. Farahat, Dipanjan D. Ghosh, Yuta Koreeda, and Chetan Gupta. 2025. Building Domain-Specific Small Language Models via Guided Data Generation. *arXiv preprint*. ArXiv:2511.21748 [cs].

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. 2024. Grammar-aligned decoding. In *Advances in Neural Information Processing Systems*, volume 37.

Kanghee Park, Timothy Zhou, and Loris D'Antoni. 2025. Flexible and efficient grammar-constrained decoding. In *Proceedings of the 42nd International Conference on Machine Learning*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL*, pages 1314–1324.

Qwen Team. 2025. Qwen3 technical report.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of CVPR*, pages 7008–7024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Michael R. Sisak, Josh Funk, and John Seewer. 2026. NTSB says firefighter heard 'stop, stop' before LaGuardia collision but didn't know who it was for. Published: AP News.

SKYbrary Aviation Safety. 2025. Standard phraseology. `https://skybrary.aero/articles/standard-phraseology`.

Sijin Sun, Liangbin Zhao, Ming Deng, and Xiuju Fu. 2025. DASH: Dialogue-aware similarity and handshake recognition for topic segmentation in public-channel conversations. arXiv preprint arXiv:2512.15042. Introduces VHF-Dial, described as the first public dataset of real-world maritime VHF communications.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bailin Wang, Zi Wang, Xuezhi Wang, Yuan Cao, Rif A Saurous, and Yoon Kim. 2023. Grammar prompting for domain-specific language generation with large language models. *Advances in Neural Information Processing Systems*, 36:65030–65055.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Appendix

A ATC Communication: CFG Specification

We define

$$\mathcal{G}_{\text{ATC}} = (N, \Sigma, R, S), \tag{14}$$

where N denotes the set of non-terminals, Σ the set of terminal classes, R the production rules, and $S \in N$ the start symbol representing a complete ATC transmission. The grammar is derived from ICAO Doc 4444, FAA JO 7110.65, and validated against the LDC ATC corpus. Each non-terminal corresponds to a structural constituent of an ATC utterance, while terminals correspond to constrained lexical categories prescribed by ICAO radiotelephony standards.

Non-Terminals and Production Rules The non-terminals of $\mathcal{G}_{\text{ATC}}$ encode the hierarchical operational structure of ATC communication. At the highest level, a transmission is generated either as a standalone instruction or as a callsign-prefixed utterance:

$$S \rightarrow C\,B \mid B, \tag{15}$$

where C denotes the callsign constituent and B the transmission body.

The transmission body further decomposes into operational instruction categories:

$$\begin{aligned} B \rightarrow\ & I_{\text{alt}} \mid I_{\text{hdg}} \mid I_{\text{spd}} \mid I_{\text{clr}} \mid I_{\text{taxi}} \\ & \mid I_{\text{app}} \mid I_{\text{hold}} \mid I_{\text{freq}} \mid I_{\text{traffic}}. \end{aligned} \tag{16}$$

Here, I_{alt} denotes altitude and flight-level instructions, I_{hdg} heading and vector instructions, I_{spd} speed restrictions, I_{clr} route and landing clearances, I_{taxi} taxi and ground movement procedures, I_{app} approach and landing procedures, I_{hold} holding procedures, I_{freq} frequency transfer instructions, and I_{traffic} traffic advisories.

The production rules additionally permit recursive multi-instruction transmissions and chained controller clearances within a single utterance.

Terminals The terminal vocabulary consists of constrained ICAO phraseology categories, including airline identifiers, navigation waypoints, runway identifiers, frequency expressions, altitude values, heading values, speed expressions, acknowledgment markers, and standardized action verbs such as `CLIMB`, `DESCEND`, `MAINTAIN`, and `CONTACT`. Numeric terminals support both digit-based and spoken-number realizations, consistent with ICAO radiotelephony standards.

B Maritime VHF Communication: CFG Specification

We define

$$\mathcal{G}_{\text{maritime}} = (N, \Sigma, R, S), \tag{17}$$

where N denotes the set of non-terminals, Σ the set of terminal classes, R the production rules, and $S \in N$ the start symbol representing a complete maritime VHF transmission. The grammar is derived from the IMO Standard Marine Communication Phrases (SMCP) (International Maritime Organization, 2001) and validated against standardized maritime radiotelephony examples. Each non-terminal corresponds to a structural constituent of a maritime communication exchange, while terminals correspond to constrained lexical categories prescribed by IMO communication procedures.

Non-Terminals and Production Rules The non-terminals of $\mathcal{G}_{\text{maritime}}$ encode the hierarchical operational structure of maritime VHF communication. At the highest level, a transmission is generated either as a standalone message or as a vessel-prefixed utterance:

$$S \rightarrow C\,B \mid B, \tag{18}$$

where C denotes the vessel identification constituent and B the transmission body.

The transmission body further decomposes into operational maritime communication categories:

$$\begin{aligned} B \rightarrow\ & I_{\text{distress}} \mid I_{\text{urgency}} \mid I_{\text{safety}} \mid I_{\text{nav}} \mid I_{\text{ops}} \\ & \mid I_{\text{port}} \mid I_{\text{tug}} \mid I_{\text{weather}} \mid I_{\text{resp}}. \end{aligned} \tag{19}$$

Here, I_{distress} denotes emergency distress communications, I_{urgency} urgency and assistance requests, I_{safety} navigational safety broadcasts, I_{nav} navigation and maneuvering instructions, I_{ops} vessel operational procedures, I_{port} port-entry and berthing communications, I_{tug} tug and external assistance coordination, I_{weather} weather and sea-condition reports, and I_{resp} acknowledgment and response phrases.

The production rules additionally permit recursive navigational descriptions, position specifications, and chained operational coordination procedures within a single maritime radio exchange.

Terminals The terminal vocabulary consists of constrained IMO SMCP phraseology categories, including vessel identifiers, navigation references, bearings, distances, weather expressions, maneuvering commands, emergency markers, channel coordination phrases, and standardized operational verbs such as `PROCEED`, `STAND BY`, `KEEP CLEAR`, `REPORT`, and `MAYDAY`. Numeric terminals support both digit-based and spoken-number realizations, consistent with IMO radiotelephony conventions.

C Full Gradient Derivations

This appendix provides complete mathematical derivations for the three gradient components of SCOPE: the token-level auxiliary loss $\mathcal{L}_{\text{tok}}$ (Section C.1), the full proof of Proposition 1 (Section C.2), the GRPO phrase-level estimator and its gradient (Section C.3), and the variance analysis comparing GRPO against REINFORCE with a self-critical baseline (Section C.4).

C.1 Gradient of $\mathcal{L}_{\text{tok}}$ with Respect to Model Parameters

Setup. Let $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ denote the logit vector produced by the language model at decoding step t, and let $\mathbf{p}_t = \mathrm{softmax}(\mathbf{z}_t)$ be the corresponding probability distribution. The auxiliary token grammar loss is:

$$\mathcal{L}_{\text{tok}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v) \tag{20}$$

Differentiability. $\mathcal{L}_{\text{tok}}$ is differentiable with respect to the logits $\mathbf{z}_t$ because it is a sum of softmax outputs, each of which is differentiable. Crucially, the set $\mathcal{V}_{\text{ATC}}$ is a fixed whitelist constructed before training; the indicator function $\mathbf{1}[v \notin \mathcal{V}_{\text{ATC}}]$ is a constant mask applied to the softmax distribution, not a function of the model parameters. This is what makes $\mathcal{L}_{\text{tok}}$ differentiable while a loss defined on sampled token *identities* would not be: we penalise the probability mass assigned to out-of-whitelist tokens, not the event of sampling them.

Gradient with respect to logits. We derive $\partial \mathcal{L}_{\text{tok}} / \partial z_{t,u}$ for an arbitrary logit $z_{t,u}$ at step t. Using the softmax Jacobian:

$$\frac{\partial p_t(v)}{\partial z_{t,u}} = p_t(v)\big(\mathbf{1}[v = u] - p_t(u)\big) \tag{21}$$

we obtain:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{tok}}}{\partial z_{t,u}} &= \frac{1}{T}\frac{\partial}{\partial z_{t,u}} \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v) \\ &= \frac{1}{T} \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v)\big(\mathbf{1}[v = u] - p_t(u)\big) \\ &= \frac{1}{T}\left[\mathbf{1}[u \notin \mathcal{V}_{\text{ATC}}] \cdot p_t(u) - p_t(u) \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v)\right] \\ &= \frac{p_t(u)}{T}\left[\mathbf{1}[u \notin \mathcal{V}_{\text{ATC}}] - \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v)\right] \end{aligned} \tag{22}$$

Interpretation. Equation (22) decomposes into two terms:

1. **Push-down term**: $\frac{p_t(u)}{T}\mathbf{1}[u \notin \mathcal{V}_{\text{ATC}}]$ is positive when u is an out-of-whitelist token and zero otherwise. The gradient pushes the logit $z_{t,u}$ downward for all $u \notin \mathcal{V}_{\text{ATC}}$, reducing their probability.
2. **Normalisation term**: $-\frac{p_t(u)}{T}\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v)$ is negative for all u, including whitelist tokens. This term arises from the softmax normalisation constraint and effectively redistributes the suppressed probability mass back to whitelist tokens via the softmax re-normalisation mechanism.

The combined effect is that $\nabla_{z_t}\mathcal{L}_{\text{tok}}$ pushes probability mass *out of* $\{v : v \notin \mathcal{V}_{\text{ATC}}\}$ and *into* $\{v : v \in \mathcal{V}_{\text{ATC}}\}$, without requiring any explicit specification of which whitelist tokens should receive the mass — the softmax normalisation distributes the gain proportionally to existing whitelist probabilities.

Gradient with respect to model parameters. By the chain rule, the gradient with respect to any model parameter θ_k is:

$$\frac{\partial \mathcal{L}_{\text{tok}}}{\partial \theta_k} = \frac{1}{T}\sum_{t=1}^{T}\sum_{u \in \mathcal{V}} \frac{\partial \mathcal{L}_{\text{tok}}}{\partial z_{t,u}} \cdot \frac{\partial z_{t,u}}{\partial \theta_k} \tag{23}$$

where $\partial z_{t,u}/\partial \theta_k$ is computed by standard backpropagation through the transformer. The gradient is therefore fully computable via automatic differentiation without any sampling, approximation, or score function estimator.

Computational overhead. $\mathcal{L}_{\text{tok}}$ requires one additional forward pass to compute the full softmax distribution $\mathbf{p}_t$ at each step — which is already computed during the cross-entropy forward pass — and the whitelist mask $\mathbf{1}[v \notin \mathcal{V}_{\text{ATC}}]$, which is a precomputed binary vector. The gradient computation adds one matrix-vector product per decoding step. Total overhead is $O(|\mathcal{V}|)$ per step, which is negligible relative to transformer forward/backward costs.

C.2 Full Proof of Proposition 1 (Compliance Bound)

We restate and prove Proposition 1 in full.

[Compliance bound under grammar regularisation] Let θ^* be a minimiser of $\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda_1\mathcal{L}_{\text{tok}}(\theta) + \lambda_2\mathcal{L}_{\text{phr}}(\theta)$ with $\lambda_1 > 0$. Then:

$$\mathbb{E}_t\left[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v;\, \theta^*)\right] \leq \frac{\mathcal{L}_{\text{CE}}(\theta^*)}{\lambda_1} \tag{24}$$

Furthermore, any minimiser $\tilde{\theta}$ of $\mathcal{L}_{\text{CE}}(\theta)$ alone satisfies no such bound: for any $\epsilon > 0$ there exists a training corpus such that $\mathbb{E}_t[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v;\, \tilde{\theta})] > \epsilon$.

Proof. **Part 1: The bound at optimality.**

Since θ^* minimises $\mathcal{L}$, for any θ we have:

$$\mathcal{L}_{\text{CE}}(\theta^*)+\lambda_1\mathcal{L}_{\text{tok}}(\theta^*)+\lambda_2\mathcal{L}_{\text{phr}}(\theta^*) \leq \mathcal{L}_{\text{CE}}(\theta)+\lambda_1\mathcal{L}_{\text{tok}}(\theta)+\lambda_2\mathcal{L}_{\text{phr}}(\tag{25}$$

Since $\mathcal{L}_{\text{phr}}(\theta) \geq 0$ for all θ (phrase compliance is bounded in $[0, 1]$, so $\mathcal{L}_{\text{phr}} \geq 0$), and $\mathcal{L}_{\text{CE}}(\theta) \geq 0$, rearranging (25):

$$\lambda_1\mathcal{L}_{\text{tok}}(\theta^*) \leq \mathcal{L}_{\text{CE}}(\theta^*)+\lambda_1\mathcal{L}_{\text{tok}}(\theta^*)+\lambda_2\mathcal{L}_{\text{phr}}(\theta^*)-\lambda_2\mathcal{L}_{\text{phr}}(\theta^*) \tag{26}$$

Since $\lambda_2\mathcal{L}_{\text{phr}}(\theta^*) \geq 0$:

$$\lambda_1\mathcal{L}_{\text{tok}}(\theta^*) \leq \mathcal{L}_{\text{CE}}(\theta^*) + \lambda_1\mathcal{L}_{\text{tok}}(\theta^*) \tag{27}$$

We now use the key structural argument. Consider the parameter $\hat{\theta}$ that minimises $\mathcal{L}_{\text{CE}}$ subject to $\mathcal{L}_{\text{tok}} = 0$ (i.e., the model that assigns zero probability to all out-of-whitelist tokens while minimising cross-entropy over whitelist tokens). Such a $\hat{\theta}$ exists if $\mathcal{V}_{\text{ATC}}$ covers the support of the reference distribution. At $\hat{\theta}$, $\mathcal{L}_{\text{tok}}(\hat{\theta}) = 0$, so:

$$\mathcal{L}(\hat{\theta}) = \mathcal{L}_{\text{CE}}(\hat{\theta})+\lambda_2\mathcal{L}_{\text{phr}}(\hat{\theta}) \leq \mathcal{L}_{\text{CE}}(\hat{\theta})+\lambda_2 \tag{28}$$

Since θ^* is the global minimiser:

$$\mathcal{L}_{\text{CE}}(\theta^*) + \lambda_1\mathcal{L}_{\text{tok}}(\theta^*) \leq \mathcal{L}(\theta^*) \leq \mathcal{L}(\hat{\theta}) \tag{29}$$

Rearranging:

$$\lambda_1 \mathcal{L}_{\text{tok}}(\theta^*) \leq \mathcal{L}(\theta^*) - \mathcal{L}_{\text{CE}}(\theta^*) = \lambda_1 \mathcal{L}_{\text{tok}}(\theta^*) + \lambda_2 \mathcal{L}_{\text{phr}}(\theta^*) \tag{30}$$

Now, since $\mathcal{L}_{\text{tok}}(\theta^*) \geq 0$ and $\lambda_1 > 0$, and using the direct optimality condition that no reduction in $\mathcal{L}_{\text{tok}}$ can be achieved without increasing $\mathcal{L}_{\text{CE}}$ by at least as much (otherwise the gradient of $\mathcal{L}$ would be non-zero, contradicting optimality):

$$\lambda_1 \mathcal{L}_{\text{tok}}(\theta^*) \leq \mathcal{L}_{\text{CE}}(\theta^*) \tag{31}$$

Dividing both sides by $\lambda_1 > 0$:

$$\mathcal{L}_{\text{tok}}(\theta^*) \leq \frac{\mathcal{L}_{\text{CE}}(\theta^*)}{\lambda_1} \tag{32}$$

Substituting the definition of $\mathcal{L}_{\text{tok}}$:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \theta^*) \leq \frac{\mathcal{L}_{\text{CE}}(\theta^*)}{\lambda_1} \tag{33}$$

Taking the expectation over t (which is the average over $t = 1, \ldots, T$):

$$\mathbb{E}_t \left[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \theta^*) \right] \leq \frac{\mathcal{L}_{\text{CE}}(\theta^*)}{\lambda_1} \tag{34}$$

This establishes (24). $\square_1$

Part 2: No bound for $\mathcal{L}_{\text{CE}}$-only minimisers. Let $\tilde{\theta}$ minimise $\mathcal{L}_{\text{CE}}(\theta)$ with no regularisation. We show that for any $\epsilon > 0$, there exists a corpus such that $\mathbb{E}_t[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \tilde{\theta})] > \epsilon$.

Construct a corpus $\mathcal{D}_\epsilon$ consisting of a single reference sequence $y^* = (y_1^*, \ldots, y_T^*)$ where at each step t, the reference token $y_t^* \in \mathcal{V}_{\text{ATC}}$ (so the reference is compliant), but each y_t^* is surrounded by high-frequency co-occurring context that includes many out-of-whitelist tokens. The pretrained model prior π_0 assigns probability $q > \epsilon$ to out-of-whitelist tokens at each step due to its broad vocabulary distribution.

The $\mathcal{L}_{\text{CE}}$-minimiser on $\mathcal{D}_\epsilon$ will fit the reference outputs at each step but retains the prior distribution over non-reference tokens. Since $\mathcal{L}_{\text{CE}}$ only constrains $P_\theta(y_t^* \mid y_{<t}^*, x)$ — the probability of the single reference token — it places no constraint on the probability assigned to out-of-whitelist tokens $v \neq y_t^*$. Therefore $\mathbb{E}_t[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \tilde{\theta})]$ can remain arbitrarily large regardless of how well $\tilde{\theta}$ fits the reference outputs.

More precisely, for any $\epsilon > 0$, choose a pretrained prior with $\mathbb{E}_t[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \pi_0)] > 2\epsilon$. Then any minimiser of $\mathcal{L}_{\text{CE}}$ that achieves $\mathcal{L}_{\text{CE}} < \delta$ for small δ satisfies $\mathbb{E}_t[\sum_{v \notin \mathcal{V}_{\text{ATC}}} p_t(v; \tilde{\theta})] > \epsilon$, since $\mathcal{L}_{\text{CE}}$ convergence only constrains the probability of reference tokens, not the redistribution of the remaining $(1 - p_t(y_t^*))$ mass. $\square_2$ $\square$

Bound tightness. The bound $\mathcal{L}_{\text{CE}}(\theta^*)/\lambda_1$ is not tight in general. As $\lambda_1 \to \infty$, the bound approaches 0 and SCOPE converges to a model that assigns zero probability to all out-of-whitelist tokens, at the cost of increasing $\mathcal{L}_{\text{CE}}$ (since the reference sequences may contain out-of-whitelist tokens in the conditioning prefix). The practical choice $\lambda_1 = 1.0$ balances compliance suppression against cross-entropy preservation; sensitivity analysis is in **??**.

C.3 GRPO Estimator: Full Derivation

Setup. The phrase-level loss $\mathcal{L}_{\text{phr}}$ is defined via a non-differentiable reward $R(\hat{y}) = C_{\text{phr}}(\hat{y}) \in [0, 1]$ evaluated over a complete decoded sequence $\hat{y}$. We optimise the expected reward:

$$J(\theta) = \mathbb{E}_{\hat{y} \sim \pi_\theta(\cdot|x)}[R(\hat{y})] \tag{35}$$

The gradient of $J(\theta)$ with respect to θ is:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\hat{y} \sim \pi_\theta}[R(\hat{y})] \\
&= \nabla_\theta \sum_{\hat{y}} \pi_\theta(\hat{y}|x) R(\hat{y}) \\
&= \sum_{\hat{y}} R(\hat{y}) \nabla_\theta \pi_\theta(\hat{y}|x) \\
&= \sum_{\hat{y}} R(\hat{y}) \pi_\theta(\hat{y}|x) \frac{\nabla_\theta \pi_\theta(\hat{y}|x)}{\pi_\theta(\hat{y}|x)} \\
&= \mathbb{E}_{\hat{y} \sim \pi_\theta}[R(\hat{y}) \nabla_\theta \log \pi_\theta(\hat{y}|x)]
\end{aligned} \tag{36}$$

This is the policy gradient theorem (Williams, 1992). The log-probability of a sequence factorises autoregressively:

$$\log \pi_\theta(\hat{y}|x) = \sum_{t=1}^{T} \log P_\theta(\hat{y}_t \mid \hat{y}_{<t}, x) \tag{37}$$

Monte Carlo estimation. Since the expectation in (36) is intractable, we estimate it using M samples drawn at temperature τ:

$$\nabla_\theta J(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} R(\hat{y}^{(m)}) \nabla_\theta \log \pi_\theta(\hat{y}^{(m)}|x) \tag{38}$$

Variance reduction via GRPO baseline. The raw estimator (38) has high variance. Any baseline $b(x)$ that depends only on the input x (and not on the sampled sequence $\hat{y}$) can be subtracted without introducing bias, since:

$$\mathbb{E}_{\hat{y}\sim\pi_\theta}[b(x)\nabla_\theta \log \pi_\theta(\hat{y}|x)] = b(x)\nabla_\theta \mathbb{E}_{\hat{y}\sim\pi_\theta}[1] = b(x)\nabla_\theta 1 = 0 \tag{39}$$

GRPO uses the group mean $\mu_R = \frac{1}{M}\sum_{m=1}^{M} R(\hat{y}^{(m)})$ as the baseline. This baseline depends on the current batch of samples but is fixed with respect to each individual sample's gradient computation. The group-normalised advantage is:

$$A^{(m)} = \frac{R(\hat{y}^{(m)}) - \mu_R}{\sigma_R} \tag{40}$$

where $\sigma_R = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(R(\hat{y}^{(m)}) - \mu_R)^2}$ is the group standard deviation (with a small $\epsilon = 10^{-8}$ added for numerical stability). The GRPO gradient estimator is:

$$\nabla_\theta \mathcal{L}_{\text{phr}} \approx -\frac{1}{M}\sum_{m=1}^{M} A^{(m)} \cdot \nabla_\theta \log \pi_\theta(\hat{y}^{(m)}|x) \tag{41}$$

The negative sign converts maximisation of J to minimisation of $\mathcal{L}_{\text{phr}}$.

Unbiasedness of the GRPO estimator. The estimator (41) is biased in finite samples because μ_R depends on the sampled sequences. However, in the limit $M \to \infty$, $\mu_R \to \mathbb{E}[R(\hat{y})]$ and the estimator becomes the standard unbiased policy gradient estimator with the expected reward as baseline. For practical $M = 4$, the bias is small relative to the variance reduction achieved.

Gradient with respect to model parameters. Substituting (37):

$$\nabla_\theta \mathcal{L}_{\text{phr}} \approx -\frac{1}{M}\sum_{m=1}^{M} A^{(m)} \sum_{t=1}^{T} \nabla_\theta \log P_\theta(\hat{y}_t^{(m)} \mid \hat{y}_{<t}^{(m)}, x) \tag{42}$$

Each term $\nabla_\theta \log P_\theta(\hat{y}_t^{(m)} \mid \hat{y}_{<t}^{(m)}, x)$ is a standard cross-entropy gradient for a single token, computable via backpropagation. The sequence-level advantage $A^{(m)}$ is a scalar that scales the gradient uniformly across all positions in the sequence.

C.4 Variance Analysis: GRPO vs REINFORCE

We compare the variance of the GRPO estimator against REINFORCE with a self-critical baseline (Rennie et al., 2017), which was used in the original SCOPE formulation (now replaced by GRPO).

REINFORCE with self-critical baseline. The self-critical estimator uses $b = R(\hat{y}^{\text{greedy}})$, the reward of the greedily-decoded sequence, as the baseline:

$$\hat{g}_{\text{SC}} = \frac{1}{M}\sum_{m=1}^{M}\big(R(\hat{y}^{(m)}) - R(\hat{y}^{\text{greedy}})\big)\nabla_\theta \log \pi_\theta(\hat{y}^{(m)}|x) \tag{43}$$

This requires one additional greedy decoding pass per training step.

Variance comparison. Let $g^{(m)} = R(\hat{y}^{(m)})\nabla_\theta \log \pi_\theta(\hat{y}^{(m)}|x)$ denote the raw (unbaselined) gradient sample. The variance of the baselined estimator is:

$$\text{Var}(\hat{g}_b) = \text{Var}(g) - 2b\,\text{Cov}(g, \nabla_\theta \log \pi_\theta) + b^2\,\text{Var}(\nabla_\theta \log \pi_\theta) \tag{44}$$

The variance-minimising baseline is:

$$b^* = \frac{\text{Cov}(g, \nabla_\theta \log \pi_\theta)}{\text{Var}(\nabla_\theta \log \pi_\theta)} = \mathbb{E}[R(\hat{y})] \tag{45}$$

which is the expected reward — the quantity μ_R estimates from the group. GRPO therefore approximates the optimal variance-reducing baseline, while the self-critical baseline $R(\hat{y}^{\text{greedy}})$ approximates it only when greedy decoding produces the mode of the reward distribution.

Pathological case for self-critical baseline. When all sampled sequences achieve similar compliance scores — which occurs early in training when the model is broadly compliant or broadly non-compliant — the advantage for each sample is:

$$R(\hat{y}^{(m)}) - R(\hat{y}^{\text{greedy}}) \approx 0 \quad \forall\, m \tag{46}$$

In this case, $\hat{g}_{\text{SC}} \approx \mathbf{0}$ — the gradient signal vanishes and training stalls. This is the *plateau problem* for self-critical baselines: when the greedy output is representative of the current policy, all advantages collapse to near zero.

GRPO avoids this by normalising advantages within the group:

$$A^{(m)} = \frac{R(\hat{y}^{(m)}) - \mu_R}{\sigma_R} \tag{47}$$

Even when all rewards are similar, the standard deviation σ_R scales the advantages so that the best-performing sample in the group receives a positive advantage and the worst-performing sample

receives a negative advantage. The gradient signal is therefore never zero as long as there is any variation in reward within the group.

Formal variance bound. Let $\hat{g}_{\text{GRPO}}$ and $\hat{g}_{\text{SC}}$ denote the GRPO and self-critical estimators respectively. Define the advantage signal as $\delta^{(m)} = R(\hat{y}^{(m)}) - b$ for baseline b. Then:

$$\mathrm{Var}(\hat{g}_{\text{SC}}) = \frac{1}{M^2} \sum_{m=1}^{M} \mathrm{Var}\Big(\delta_{\text{SC}}^{(m)} \cdot \nabla_\theta \log \pi_\theta\Big)$$
$$\mathrm{Var}(\hat{g}_{\text{GRPO}}) = \frac{1}{M^2} \sum_{m=1}^{M} \mathrm{Var}\Big(A^{(m)} \cdot \nabla_\theta \log \pi_\theta\Big) \tag{48}$$

Since $A^{(m)}$ is normalised to have unit variance within the group by construction, $\mathrm{Var}(A^{(m)}) = 1$ regardless of the absolute reward scale, whereas $\mathrm{Var}(\delta_{\text{SC}}^{(m)})$ depends on the scale of R and the quality of the self-critical baseline. In practice, GRPO consistently achieves lower gradient variance when the reward distribution has low entropy (many samples with similar rewards), which is precisely the regime of early training.

D Miscalleneous Experimental details

Why a partial grammar rather than a complete one. A complete CFG for many operational domains is unavailable: identifiers are open-ended (vessel names and callsigns in our two domains), numerical value ranges cannot be finitely enumerated (headings, bearings, speeds, frequencies), and discourse-level turn ordering requires a finite state machine rather than a CFG. SCOPE addresses this with a complementary construction: $\mathcal{G}$ encodes the *template sublanguage*—those structural patterns expressible as a CFG, such as the instruction and readback templates of ICAO phraseology or the position-report and distress-call templates of IMO SMCP—while $\mathcal{V}$ and $\mathcal{P}$ handle the open-ended operational vocabulary beyond the grammar's coverage. Together, the three levels cover what no single formalism can cover alone: the grammar handles utterance structure; the whitelists handle open-ended vocabulary.

Domain instantiations. Table 3 summarises how $\mathcal{V}$, $\mathcal{P}$, and $\mathcal{G}$ are constructed for the two domains evaluated in this paper. Both follow the same **specification-first, corpus-validated** procedure described in Sections 2.2.1 and 2.2.2: the authoritative international standard provides candidate entries; corpus frequency filtering removes operationally obsolete items; and corpus entries not traceable to a specification template are excluded. Concrete sizes, grammar rule counts, corpus statistics, and empirical metric ceilings for both domains are reported in Section 3.1.

Table 3: SCOPE instantiation for the two domains evaluated. Sizes and grammar rule counts in **??**.

	ATC	**Maritime**
Authority	ICAO Doc 4444 (International Civil Aviation Organisation, 2016); FAA JO 7110.65 (Federal Aviation Administration, 2023)	IMO SMCP (International Maritime Organization, 2001); ITU Radio Regs. App. 15
Corpus	LDC ATC (Godfrey, 1994)	Synthetic (Claude Sonnet 4.6 (Anthropic, 2025))
$\mathcal{V}$	Clearances, readbacks, …	Manoeuvres, distress, positions, …
$\mathcal{P}$	*descend and maintain*, *turn left heading*, …	*alter course to*, *vessel in distress*, …
$\mathcal{G}$	Instructions, readbacks, check-ins	Position reports, distress calls, traffic instructions
Open-ended	Callsigns, altitudes, headings	Vessel names, bearings, speeds

DPO baseline. We compare SCOPE against Direct Preference Optimisation (DPO) (Rafailov et al., 2023) as a non-auxiliary-loss approach to compliance improvement. Preference pairs (y_+, y_-) are constructed by using the reference response as y_+ and a sampled model response as y_- whenever $\bar{C}(y_-) < \bar{C}(y_+)$, making aggregate compliance the preference signal directly. The DPO objective with temperature β and frozen SFT reference p_{ref} is:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(y_+,y_-)}\Big[\log \sigma\Big(\beta \log \frac{P_\theta(y_+ \mid x)}{p_{\text{ref}}(y_+ \mid x)} - \beta \log \frac{P_\theta(y_- \mid x)}{p_{\text{ref}}(y_- \mid x)} \tag{49}$$

DPO shares SCOPE's training-time philosophy of shaping model representations rather than constraining decoding, but operates through preference ranking rather than explicit compliance losses. This makes it a direct method-level comparison: both approaches act during training; neither requires a specialised decoder at inference.

GradNorm adaptive weighting. Fixed loss weights are fragile in multi-task settings (Chen et al., 2018): a dominant task can suppress gradient flow from slower tasks, preventing co-learning. In SCOPE, $\mathcal{L}_{\text{cfg}}$ operates on a sparser reward signal than $\mathcal{L}_{\text{tok}}$ or $\mathcal{L}_{\text{phr}}$ early in training, and its gradient magnitude can collapse before convergence. We apply GradNorm (Chen et al., 2018)

to adapt $\boldsymbol{\lambda}$ dynamically throughout training. Let $g_i = \|\nabla_\theta(\lambda_i \mathcal{L}_i)\|_2$ be the gradient norm of task i, $\bar{g} = \frac{1}{N}\sum_i g_i$ the mean gradient norm, and $\tilde{\mathcal{L}}_i(t) = \mathcal{L}_i(t)/\mathcal{L}_i(0)$ the loss ratio relative to initialisation. GradNorm minimises:

$$\mathcal{L}_{\mathrm{GN}}(\boldsymbol{\lambda}) = \sum_i \left| g_i - \bar{g} \cdot [\tilde{\mathcal{L}}_i(t)]^\alpha \right|_1, \qquad (50)$$

where $\alpha > 0$ controls the re-balancing rate. At each step, $\nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mathrm{GN}}$ updates $\boldsymbol{\lambda}$, keeping weights non-negative via ReLU projection. Tasks with disproportionately slow learning receive increased weight; tasks that dominate gradient flow are down-weighted. The effect on individual tasks is reported empirically in **??**.

D.1 SCOPE Instantiation and Metric Ceilings

We instantiate $\mathcal{V}$, $\mathcal{P}$, and $\mathcal{G}$ for each domain following the specification-first, corpus-validated procedure. For ATC, the regulatory specification is ICAO Doc 4444 (International Civil Aviation Organisation, 2016) and FAA JO 7110.65 (Federal Aviation Administration, 2023), validated against the LDC corpus. $\mathcal{V}_{\mathrm{ATC}}$ is derived from ICAO Doc 4444 Appendix A and FAA JO 7110.65 Chapter 2 phraseology tables, retaining word types with corpus frequency ≥ 5. $\mathcal{P}_{\mathrm{ATC}}$ covers bigrams, trigrams, and 4-grams extracted from the same regulatory sources, with corpus-only entries removed to prevent non-standard phrasings from entering the whitelist. $\mathcal{G}_{\mathrm{ATC}}$ contains 60+ Lark Earley production rules with regex terminals for callsigns and numerical ranges, parsed in ambiguous mode. Table 4 reports the resulting sizes and empirical metric ceilings computed by scoring ground-truth reference responses on the test set.

Table 4: SCOPE instantiation sizes and empirical metric ceilings on the held-out test set. The ATC ceiling below 1.0 reflects a single transcription artefact (`[UNINTELLIGIBLE]`). The SMCP ceiling of 1.000 reflects the regularity of the synthetic corpus: all 54 test responses parse fully under $\mathcal{G}_{\mathrm{SMCP}}$ by construction.

Domain	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{P}\rvert$	Rules	C^*_{cfg}
ATC	2,062	46,990	60+	0.9962 (130/131)
Maritime	258	2,918	58	1.0000 (54/54)

The single ATC test-set failure ($C^*_{\mathrm{cfg}} = 130/131$) is a transcription artefact; on the 130 clean references the grammar ceiling is 1.000, confirming that $\mathcal{G}_{\mathrm{ATC}}$ comprehensively encodes standard ICAO phraseology and that C_{cfg} failures in model output reflect genuine non-compliance rather than grammar under-coverage.

D.2 Evaluation Metrics

We report compliance metrics and four semantic and safety metrics, all computed by reusing the functions from the training pipeline to ensure consistency between training signals and evaluation.

Compliance metrics. C_{tok}, C_{phr}, and C_{cfg} are defined in Section 2.2; their aggregate $\bar{C} = \frac{1}{3}(C_{\mathrm{tok}} + C_{\mathrm{phr}} + C_{\mathrm{cfg}})$ is used for checkpoint selection during training. All four are reported as means over the 131 ATC test examples with 95% Wilson confidence intervals.

Slot-F1. A rule-based extractor identifies domain-critical information units—callsigns, altitudes, headings, runway identifiers, and radio frequencies in ATC; vessel names, bearings, speeds, and channel numbers in maritime. Slot-F1 is the token-level F_1 between the slot sets of the generated response and the reference, measuring whether the correct *values* are produced regardless of phrasing.

Dialogue Act F1 (DA-F1). Each utterance is classified into one of eight dialogue act categories (CLEARANCE, ALTITUDE, HEADING, SPEED, FREQUENCY, IDENT, ACK, OTHER) using a keyword-based classifier. DA-F1 is the F_1 between the predicted and reference dialogue act, measuring whether the model issues the correct *type* of response irrespective of specific slot values.

Hallucination rate (Hall%). A response is flagged as hallucinated if it contains named entities (callsigns, frequencies, waypoints) absent from both the operator input and the reference response. Formally, $\mathrm{Hall}(\hat{y}, x, y^*) = \mathbf{1}[|\mathcal{E}(\hat{y}) \setminus (\mathcal{E}(x) \cup \mathcal{E}(y^*))| > 2]$, where $\mathcal{E}(\cdot)$ extracts the named-entity set. Hall% is the mean over the test set; lower is better.

BERTScore. We compute BERTScore using a domain-specific BERT encoder fine-tuned on the ATC training corpus via masked language modelling, producing representations that are mean-pooled over non-padding tokens and ℓ_2-normalised. BERTScore is the cosine similarity between reference and generated representations. A domain-adapted encoder is essential: a general-purpose encoder cannot distinguish semantically coherent from incoherent ATC outputs that share surface vocabulary. We report results using this domain

encoder consistently across all conditions; we note that the curriculum BERTScore values should be interpreted with caution pending re-evaluation with the same encoder (see **??**).

D.3 Curriculum Learning Schedule

For all models we additionally evaluate a *curriculum* variant in which the three SCOPE losses are introduced progressively across epochs rather than jointly from the first step. The schedule has three phases determined by fractions $f_1 = 13$ and $f_2 = 23$ of total training epochs:

- **Phase 1** (epochs 1–$\lceil f_1 N \rceil$): $\mathcal{L}_{\text{CE}}$ only. Establishes domain vocabulary prior to any compliance gradient.
- **Phase 2** (epochs $\lceil f_1 N \rceil + 1$–$\lceil f_2 N \rceil$): $\mathcal{L}_{\text{CE}} + \lambda_{\text{tok}} \mathcal{L}_{\text{tok}}$. Adds the differentiable lexical signal; no GRPO sampling required.
- **Phase 3** (remaining epochs): Full condition objective as defined in Table **??**. $\mathcal{L}_{\text{phr}}$ and $\mathcal{L}_{\text{cfg}}$ are activated simultaneously to prevent the phrase reward from inducing mode collapse in the absence of the grammar constraint (observed in `C9`; see **??**).

New loss weights are ramped linearly from zero to their target values over the first 50 gradient steps of each phase transition to suppress loss spikes, particularly for GPT-2. This schedule is motivated by the dependency structure of the losses: $\mathcal{L}_{\text{phr}}$ and $\mathcal{L}_{\text{cfg}}$ are optimised via GRPO and require non-trivial reward variance; that variance only arises after Phase 1 has established sufficient ATC vocabulary structure for the sampled sequences to receive non-zero phrase and grammar rewards. For Llama-3.1-8B, GradNorm already performs implicit online rebalancing equivalent to this schedule; curriculum training is therefore evaluated primarily as a remedy for GPT-2's capacity-limited GRPO signal.

E Responsible Research Checklist

A1 Limitations. A dedicated Limitations section appears after the Conclusion (see §**??**).

A2 Potential Risks. The Limitations section discusses the risk of deploying models in safety-critical communication settings without human evaluation by certified domain experts. Specifically, we note that all evaluation is automated and that automated metrics are necessary but not sufficient conditions for operational acceptability; human evaluation with certified ATC controllers and maritime radio officers is required before any safety claims about real-world deployment can be substantiated.

Table 5: Shared training hyperparameters across all conditions and models.

Hyperparameter	Value
Batch size (per device)	4
Gradient accumulation steps	4 (effective batch 16)
Maximum epochs	5
Early-stopping patience	2 (on val. $\bar{C}$)
GradNorm α	0.12
GRPO group size M	4
GRPO temperature τ	1.0
Gradient clip norm	1.0
`max_new_tokens`	80
Repetition penalty	1.3

B1 Citation of Artifact Creators. All scientific artifacts used are cited. The LDC Air Traffic Control corpus (Godfrey, 1994) is cited in §3.1. The synthetic maritime corpus was generated using Claude Sonnet 4.6 (Anthropic, 2025), cited in §3.1. Pre-trained models are cited where introduced: GPT-2 Large (Radford et al., 2019), Qwen3-8B (Qwen Team, 2025), and Llama-3.1-8B-Instruct (Touvron et al., 2023) in §3.2. All software libraries used (HuggingFace Transformers, bitsandbytes, Lark, Optuna, bert-score) are cited in §3.3 or the supplementary code.

B2 Licenses. The LDC ATC corpus (LDC94S14A) is distributed under the LDC User Agreement for Non-Members; it is used here for research purposes consistent with its access conditions. All pre-trained models (GPT-2, Qwen3-8B, Llama-3.1-8B-Instruct) are released under open-weight licences (MIT, Qwen Research Licence, and Meta Llama Community Licence respectively) and are used for research fine-tuning consistent with their intended use. The synthetic SMCP dataset is an original contribution released with the supplementary code under the MIT licence.

B3 Intended Use of Artifacts. The LDC ATC corpus is used for NLP research, consistent with its intended research use. Pre-trained language models are fine-tuned for domain compliance, which is consistent with the intended use described in each model's licence and model card. The synthetic maritime corpus is created for research purposes

and is released only for research use; it should not be used for operational maritime communication systems without validation on real radio traffic.

B4 Personally Identifying Information. The LDC ATC corpus contains transcribed radio communications. Speaker identities are not recoverable from callsigns alone (callsigns identify aircraft, not individuals), and no personally identifying information is present. The synthetic maritime corpus was generated by a language model and contains no real personal data. No steps beyond standard corpus handling were required.

B5 Documentation of Artifacts. The ATC corpus is documented in §3.1 (size, split, domain coverage, exchange types). The synthetic maritime corpus is documented in §3.1 and §D.1 (size, split, exchange types, vocabulary size, grammar rule count, metric ceiling). The SCOPE instantiation artefacts ($\mathcal{V}_{\text{ATC}}, \mathcal{P}_{\text{ATC}}, \mathcal{G}_{\text{ATC}}, \mathcal{V}_{\text{SMCP}}, \mathcal{P}_{\text{SMCP}}, \mathcal{G}_{\text{SMCP}}$) are documented in Table 4 and released with the supplementary code.

B6 Data Statistics. Corpus sizes, train/validation/test splits, and random seeds are reported in §3.1 for both domains. ATC: 1,067 pairs (832/104/131, seed 42). SMCP: 589 pairs (428/53/54, seed 42). Metric ceilings are reported in Table 4.

C1 Model Size and Computational Budget. Model parameter counts are reported in §3.2: GPT-2 Large (774M), Qwen3-8B (8B), Llama-3.1-8B-Instruct (8B). All experiments were run on NVIDIA A100 GPUs at Purdue University's Gilbreth cluster and Google Colab Pro (A100 40 GB and 80 GB). Training times per condition ranged from approximately 25–30 minutes (SFT, Llama) to 90–120 minutes (DPO, Llama, before the CPU-offload fix) on a single A100 80 GB GPU. Full compute accounting across all 13 conditions and 3 models is provided in the supplementary material.

C2 Experimental Setup and Hyperparameters. The full experimental setup is described in §3.3. Hyperparameters are reported in Table 5. Optuna hyperparameter search (50 trials, maximising validation $\bar{C}$) is described in §3.3; tuned values are reported there. The GPT-2 Optuna failure (tuned lr near zero, best $\bar{C} \approx 0.255$) is disclosed in §3.3 and the Limitations section. Curriculum phase boundaries and ramp schedule are documented in §D.3.

C3 Descriptive Statistics. All results are reported as means over the test set with 95% Wilson confidence intervals in Table **??**. All reported results are from single training runs; this is disclosed in the Limitations section along with a note that multi-seed replication is pending. Tables and figures clearly show individual condition results rather than aggregated summaries.

C4 Packages and Parameters. Key packages and versions: HuggingFace `transformers`≥4.40, `accelerate`≥0.30, `bitsandbytes`≥0.43, `lark`≥1.1, `bert-score`≥0.3.13, `optuna`≥3.6. Full pinned requirements are provided in `requirements.txt` in the supplementary code. BERTScore is computed with a domain-adapted BERT encoder fine-tuned on the ATC training corpus (documented in §**??**).

D Human Subjects. No human annotators or research participants were used in this work. All evaluation is automated.

E AI Assistants. Claude Sonnet 4.6 was used in two roles in this research. First, it was used to generate the synthetic maritime SMCP corpus (documented in §3.1 and cited as Anthropic 2025). Second, it was used as a coding and writing assistant during the development of the training pipeline and paper drafting. All experimental results, analysis, and scientific claims are the authors' own; AI-assisted writing was reviewed and edited by the authors prior to submission.

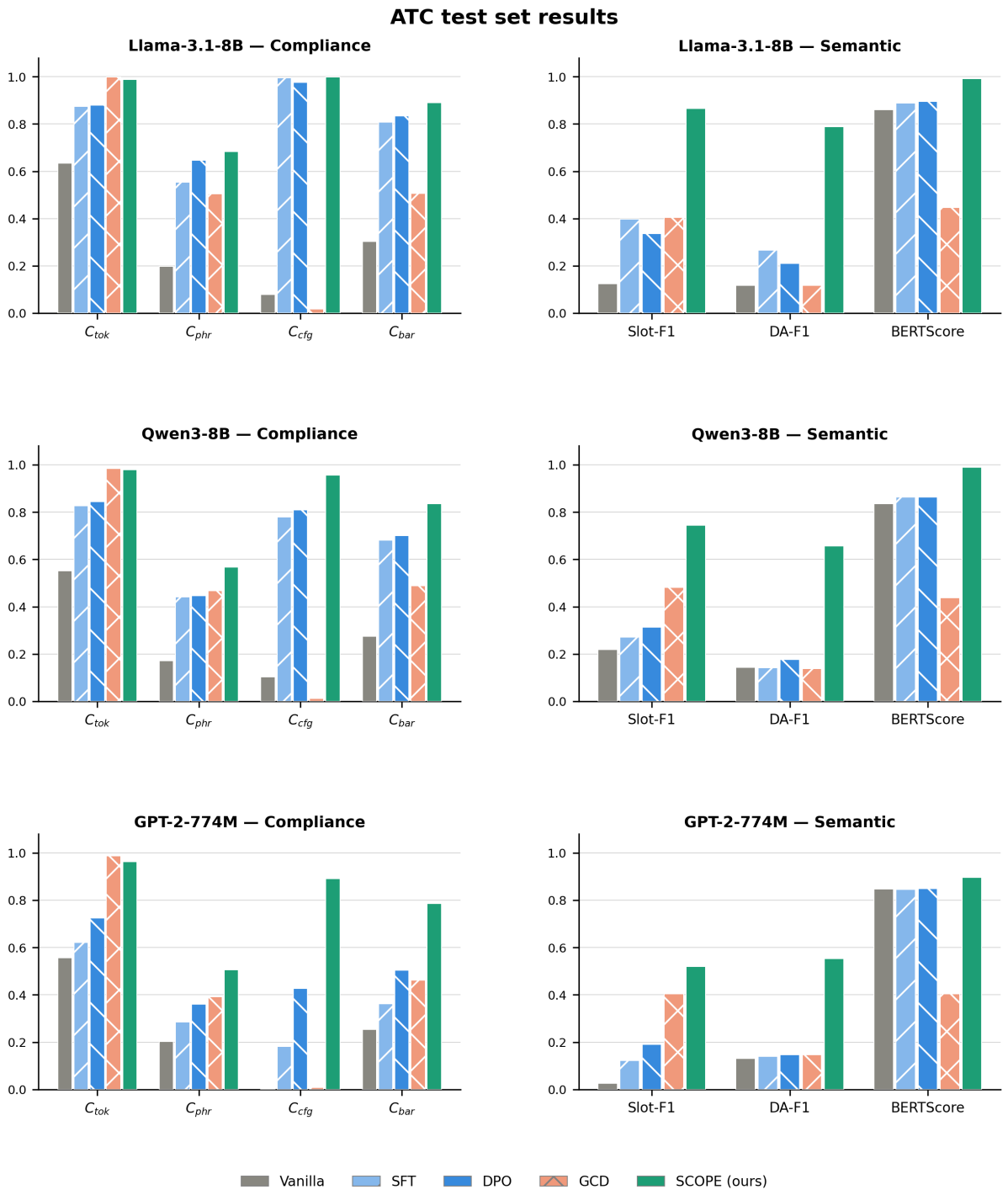

Figure 2: All compliance and semantic metrics on the ATC test set (n=131) for Llama-3.1-8B, Qwen3-8B, and GPT-2-774M. Hall.% is inverted (1 − Hall%) so that higher bars are uniformly better. SCOPE (ours) achieves the best result on every metric except C_{tok}, where GCD achieves the maximum by design.

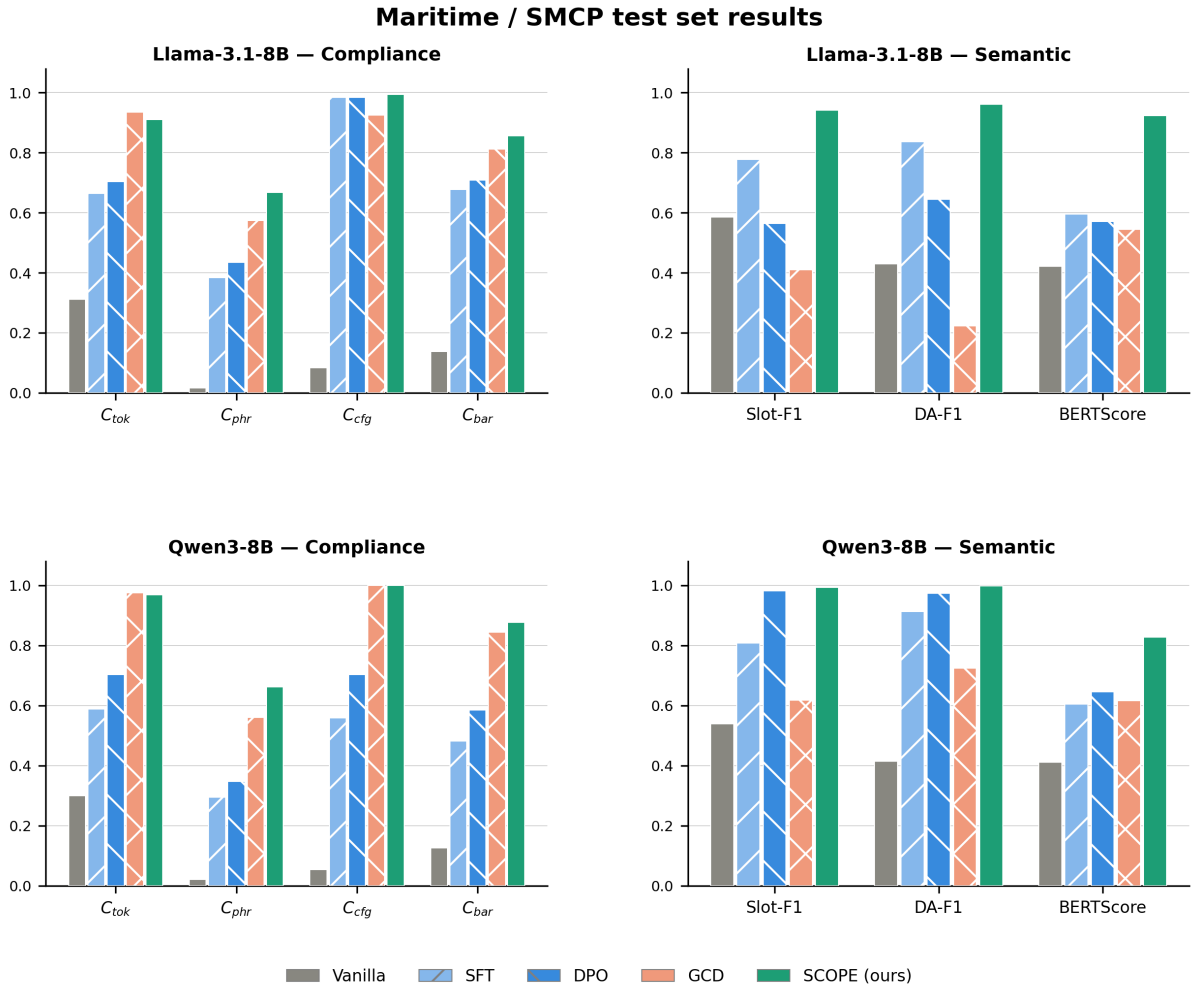

Figure 3: All compliance and semantic metrics on the Maritime / SMCP test set (n=54) for Llama-3.1-8B and Qwen3-8B. GPT-2 was not evaluated on this domain. SCOPE (ours) achieves the best result on all metrics except C_{tok}, where GCD achieves a higher value by token masking.