

# Towards Automated Air Traffic Safety Assessment Around Non-Towered Airports Using Large Language Models

Torsten Darrell\*, Mahyar Ghazanfari<sup>†</sup>  
*George Washington University, Washington, DC, 20052*

Jordan Kam<sup>‡</sup>, Alexandre Bayen<sup>§</sup>  
*University of California, Berkeley, Berkeley, CA, 94720*

Amin Tabrizian<sup>¶</sup> and Peng Wei<sup>||</sup>  
*George Washington University, Washington, DC, 20052*

We investigate frameworks for post-flight safety analysis at non-towered airports using large language models (LLMs). Non-towered airports rely on the Common Traffic Advisory Frequency (CTAF) for air traffic coordination and experience frequent near mid-air collisions due to the pilot self-announcement communication protocol. We propose a general vision-language model (VLM) approach to analyze the transcribed CTAF radio communications in natural language, METeorological Aerodrome Report (METAR) weather data, Automatic Dependent Surveillance-Broadcast (ADS-B) flight trajectories, and Visual Flight Rules sectional charts of the airfield. We provide a preliminary study at Half Moon Bay Airport, with a qualitative real world case study and a quantitative evaluation using a new synthetic dataset of communications and weather modalities. We qualitatively evaluate our framework on real flight data using Gemini 2.5 Pro, demonstrating accurate identification of a right-of-way violation. The synthetic dataset is derived from real examples and includes a 12-category hazard taxonomy, and is used to benchmark three open-source (Qwen 2.5-7B, Mistral-7B, Gemma-2-9B) and three closed-source (GPT-4o, GPT-5.4, Claude Sonnet 4.6) LLM models on the subset of inputs related to CTAF and METAR. Even limited to CTAF and METAR inputs and open source LLMs, instances of our framework typically achieve a macro  $F_1$  score above 0.85 on a binary nominal/danger classification task. Future work includes a quantitative evaluation across all modalities and a larger number of real world examples. Taken together, our results suggest that VLM analysis of safety at non-towered airports may be a valuable future capability.

## I. Introduction

Traditional air traffic control (ATC) employs complex decision-making to ensure mid-air collision-avoidance in the United States national airspace system (NAS) [1]. Airspace operations at non-towered airports follow procedures outlined in the Federal Aviation Administration’s (FAA) advisory circular (AC) 90-66C document [2]. This communication is administered between pilots over the common traffic advisory frequency (CTAF) to offer a level of safety in the absence of a human controller. While this collision-avoidance method has been in place for nearly fifty years, its heavy reliance on pilot’s voluntary decisions to announce their aircraft’s position and intent to others has led to many near mid-air collisions at these non-towered airports [3]. CTAF’s lack of layered safety highlights the shortcomings of purely pilot-to-pilot deconfliction. As the amount of general aviation (GA) in low-altitude airspace increases, the development of automated ATC as a safety layer at these non-towered airports could close this safety critical technology gap.

ATC’s deconfliction architecture relies on rigid rule-based or data-driven systems that function well only when inputs are complete and precisely formatted (e.g., sensor data or communication) [4]. Furthermore, these architectures are primarily human-in-the-loop. However, non-towered airport operations are inherently unstructured. CTAF exchanges are spontaneous, abbreviated, and often incomplete radio calls whose meaning depends heavily on contextual cues such as traffic position, weather, and visual conditions [2]. Non-towered airports also typically lack the infrastructure required to log flight data essential for monitoring airspace operations.

---

\*Research Intern, Department of Mechanical and Aerospace Engineering, George Washington University, AIAA Student Member.

<sup>†</sup>Ph.D. Student, Department of Mechanical and Aerospace Engineering, George Washington University, AIAA Student Member.

<sup>‡</sup>Undergraduate Student, Aerospace Program, University of California, Berkeley, AIAA Student Member.

<sup>§</sup>Full Professor, Department of Electrical Engineering and Computer Science, University of California, Berkeley.

<sup>¶</sup>Ph.D. Student, Department of Computer Science, George Washington University, AIAA Student Member.

<sup>||</sup>Associate Professor, Department of Mechanical and Aerospace Engineering, George Washington University, AIAA Associate Fellow.

Recent advances in artificial intelligence (AI) have produced remarkable progress across audio, vision, and language domains, enabling large language models (LLMs) and vision-language models (VLMs) to perform a broad range of translation, comprehension, and reasoning tasks [5–9], opening the possibility they could handle the unstructured, heterogenous, and semantically-rich nature of non-towered ATC data. LLMs have demonstrated strong reasoning capabilities over natural language, structured data, and contextual information [10]. Pre-trained on large-scale text corpora, these models learn relationships among language, numerical data, and higher-level concepts, enabling them to infer deeper contextual information such as intent. VLMs extend these capabilities to multimodal settings by jointly reasoning over visual and textual inputs, making them well suited for aviation applications involving heterogeneous data sources such as flight trajectories, airport diagrams, weather information, and pilot communication transcripts.

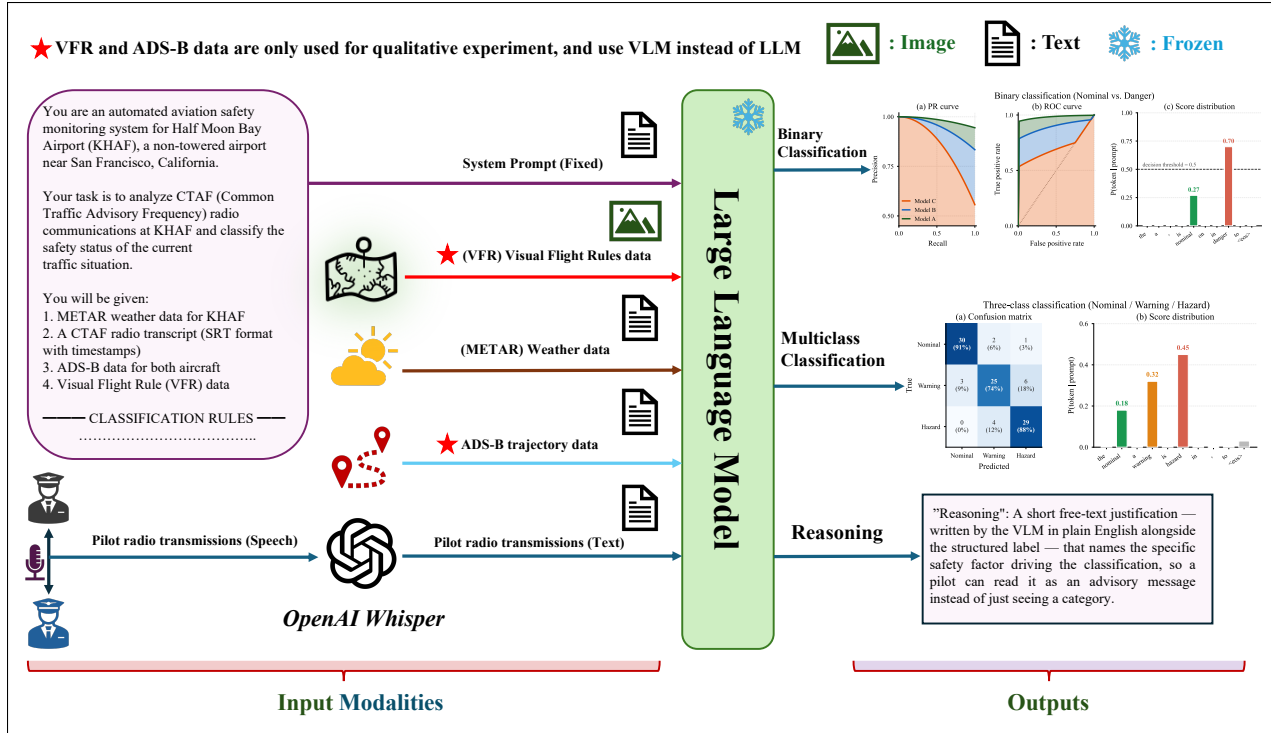
We propose LLM-based models for post-flight safety assessments of airspace operations surrounding non-towered airports. We consider a general VLM approach to analyze the relevant air traffic control (ATC) signals including transcribed CTAF radio communications in natural language, METeorological Aerodrome Report (METAR) weather data, Automatic Dependent Surveillance-Broadcast (ADS-B) flight trajectories, and Visual Flight Rules (VFR) sectional charts of the airfield. A qualitative experiment using Half Moon Bay Airport (KHAF) as a case study suggests that a contemporary VLM can discriminate safe vs unsafe conditions when prompted with these inputs. Initial quantitative evaluations of model performance when prompted solely with CTAF and METAR data confirm performance across a variety of models and parameter settings. Future work includes a quantitative evaluation across all modalities and a larger number of real world examples. We believe that LLM and VLM based models can serve as an important element to support safe ATC operations around non-towered airports and/or in general scenarios where human air traffic controllers are unavailable.

## II. Related Work

While there is a long history of automated collision avoidance, including terrain avoidance [11] and head-to-head collision resolution systems [12] that have been deployed in modern aviation based on explicit predictive models and control theory [13], analysis of natural language CTAF communications for aircraft safety or intent prediction has been relatively underexplored. Prior work has considered natural language processing for pilot communications, including extensive investigation of Automatic Speech Recognition (ASR) for automatic transcription, and the development of aviation safety databases and incident evaluations [14]. Other work using LLMs, such as ChatGPT, have been used for summarization of historical ground delay program (GDP) data to support optimal air traffic flow management in strategic decision making [15]. Lastly, prior research has studied non-towered airports and modeled ATC as a Markov Decision Process (MDP) towards autonomous ATC of GA operations [16].

Prior works have used LLMs for safety report summarization of aviation report data. LLMs have been used to summarize incident narratives from the Aviation Safety Reporting System (ASRS) [17] and extract causal patterns related to human factors, while [18] investigated domain adaptation of transformer-based architectures for documentation such as Letters of Agreements. Both works demonstrating improved accuracy in domain-specific classification. The AviationGPT framework [19] further introduced a domain-tuned LLaMA-based [20] LLM that achieved substantial performance gains on aviation text corpora. In parallel, [21] employed an LLM to detect pilot workload levels from multimodal physiological and behavioral data. Beyond these textual and design applications, [22] proposed a chain-of-thought-based LLM flight planner for eVTOL routing under wind hazards, achieving high plan validity while incorporating human preferences via natural-language input. [23] on the other hand, introduced an embodied ATC agent using LLMs with function-calling and an experience library to autonomously resolve multi-aircraft conflicts in simulated airspace, achieving near-human reasoning performance.

More recently, [15] developed CHATATC, a fine-tuned conversational agent trained on over 80,000 GDP issuances to support strategic air-traffic-flow management. The study demonstrated that LLM-based summarization and dialogue systems can assist Traffic Managers by retrieving, synthesizing, and contextualizing historical NAS data for decision support. In addition, [24] introduced a multimodal model that integrates ASR and LLM-based intent extraction with spatial reasoning for goal prediction in untowered airspace, showing that language-conditioned multimodal fusion significantly reduces goal-prediction error. Collectively, these studies highlight the expanding role of LLMs as reasoning and decision-support systems across aviation, covering domains such as safety analysis, workload estimation, multimodal intent prediction, route generation, and traffic-flow management. Yet, existing efforts primarily focus on prediction and decision-support tasks rather than understanding pilot communications.



**Fig. 1 Framework overview.** Pilot radio communications are transcribed by Whisper Large-v3 and passed to a frozen LLM alongside METAR weather, ADS-B trajectory, and a system prompt text. The model emits both a structured safety classification (binary, with a three-class breakdown in Appendix VIII.B) and a free-text reasoning string that can be surfaced to the pilot as a CTAF-style advisory.

### III. Framework

#### A. Overview

Our framework (Fig. 1) consists of a LLM prompted with a heterogeneous set of data modalities based on ATC inputs which generates both a structured text safety classification and a free-text reasoning output. Five inputs are concatenated into a single prompt: a fixed task prompt (text), METAR weather data (text), ADS-B flight trajectories (text), transcribed CTAF communications between pilots (text), and a VFR sectional chart of the airfield (image). The model’s structured output is a binary nominal-versus-danger classifier; we also evaluate using a finer-grained three-class safety classifier, reported in Appendix VIII.B.

The LLM is frozen throughout this work, no weights are updated, and no fine-tuning is performed using the synthetic or real flight data. The model’s behavior is shaped entirely through prompt engineering, in-context learning, and the choice of reasoning. This design choice preserves modularity, and the LLM used can be swapped or upgraded independently of the prompt template or the ASR component. We evaluate using a variety of LLM/VLM models as described below. The model implements a mapping:

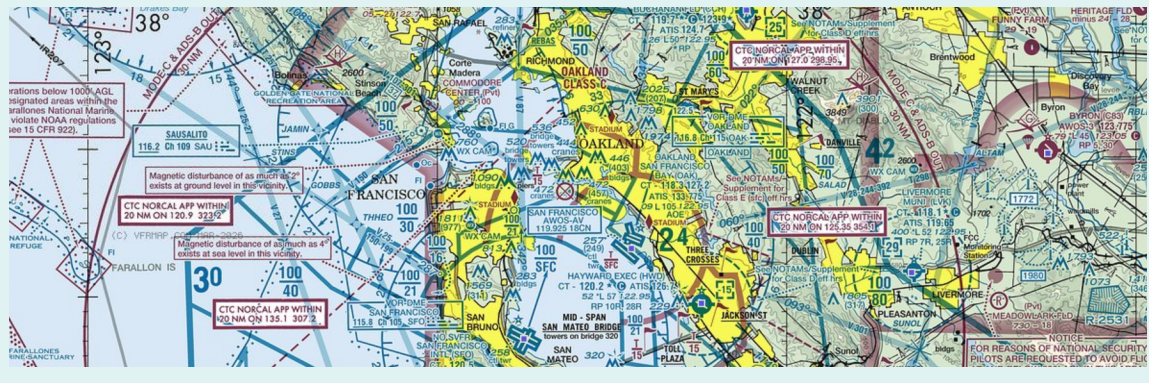
$$f_{\theta} : (x_{\text{prompt}}, x_{\text{CTAF}}, x_{\text{METAR}}, x_{\text{ADS-B}}, x_{\text{VFR}}) \rightarrow (y, \alpha)$$

where  $y$  is the predicted safety category and  $\alpha$  is the free-text reasoning.

Whisper performs a deterministic mapping  $g_{\phi} : x_{\text{audio}} \rightarrow x_{\text{CTAF}}$ , transcribing raw radio audio (with timestamps preserved in SubRip Text format) into the textual representation later provided to the model.

## Input Modalities

### VFR Sectional Chart — KHAF (Half Moon Bay Airport)



### METAR — KHAF

KHAF 041655Z AUTO 27005KT 9SM BKN060  
14/13 A2999 RMK A02

Wind 270°/5 kt · Visibility 9 SM · BKN ceiling 6,000 ft  
· Temp 14°C / Dewpoint 13°C · Altimeter 29.99 inHg

### ADS-B State Data

	N100AB	N200CD
Altitude (ft)	1,800↓	900↑
Speed (kt)	75	60
Heading (°)	301	210
Vert. rate (fpm)	-600	+200

📶 N100AB: straight-in RNAV 📶 N200CD: right-base turn

### CTAF Transcript (Whisper Large-v3 ASR · KHAF 122.8 MHz)

Standard call frame “Half Moon Bay traffic, . . . , Half Moon Bay” omitted for brevity. 📶 N100AB = straight-in RNAV 📶 N200CD = VFR pattern • *italics* = other traffic

- 01:13 📶 N200CD “turning right crosswind 30.”
- 02:52 📶 N200CD “turning right base, 30.”
- 03:44 📶 N100AB “12 miles to the southeast, 3,400 ft, descending on the instrument approach for runway 30, full stop.”
- 04:10 📶 N100AB “coming up on 6-mile final for 30, 1,500 ft.”
- 04:43 📶 N100AB “10 miles to the southeast, 3,300 ft, straight in on the instrument approach runway 30, full stop.”
- 05:49 📶 N100AB “4-mile final.”
- 05:55 📶 N100AB “7 miles to the southeast, 2,600 ft, straight in on the instrument for runway 30, full stop.”
- 06:11 📶 N200CD “turning crosswind.”
- 06:46 📶 N100AB “5 miles southeast, 1,900 ft, straight in instrument approach, runway 30, full stop.”
- 07:40 📶 N200CD “on downwind — will turn base after the traffic on short final.”
- 07:49 📶 N100AB “3 miles to the southeast, straight in on the instrument approach, full stop, runway 30.”
- 08:20 📶 N200CD “turning right base, 30.”
- 08:31 📶 N100AB “2-mile final for runway 30 — traffic turning final in front of us; we’re gonna have to go around if you’re not off that runway.”
- 08:49 📶 N100AB “we see you deviating off to the right now, 1½-mile final, full stop for 30.”
- 10:29 📶 N200CD “turning final, about a 2-mile final.”

Fig. 2 Qualitative results of the VLM safety analysis pipeline at KHAF, using a VFR sectional, METAR, ADS-B state data, and a Whisper-transcribed CTAF feed. Gemini identifies a right-of-way violation on final approach, resolved by a go-around warning. Call signs shown are random placeholders, not the real call signs; all other data is unmodified.

### CTAF Transcript (continued)

- 12:49 *Cherokee 1AA* “4 miles to the southeast at 2,000 ft descending, will fly over the ocean, cross midfield into the right downwind.”
- 13:21 *Skyhawk 2BB* “10 miles to the southeast on RNAV 30 inbound.”
- 14:47 📍 **N200CD** “turning right downwind, right traffic 30.”
- 15:21 *Skyhawk 2BB* “5 miles to the southeast on RNAV 30, full stop.”
- 15:50 *Archer 3CC* “6 miles to the southeast, maneuvering over water, then crossing midfield to enter right downwind 30.”
- 16:09 *Cherokee 1AA* “overhead midfield entering right downwind 30 — continuing climb, departing area to the south, 1,500 ft.”
- 16:23 📍 **N200CD** “right to base, 30.”
- 16:43 *Skyhawk 2BB* “on the RNAV, is that you at 2,300?” *Skyhawk 2BB*: “Affirm, on the RNAV 30.”
- 17:00 *Skyhawk 2BB* “we’ll be going under you.”
- 17:35 *Skyhawk 2BB* “5 miles to the southeast on RNAV 30, full stop.”
- 18:45 *Skyhawk 2BB* “3 miles final on 30.”
- 19:10 *Archer 3CC* “offshore, turning to cross midfield and enter the right downwind 30.”
- 19:27 *Skyhawk 2BB* “on the RNAV 30.”
- 20:10 *Archer 3CC* “crossing midfield to enter right downwind 30.”
- 20:23 *Skyhawk 2BB* “short final 30.”
- 20:43 *Archer 3CC* “turning onto the downwind 30.”
- 20:54 📍 **N200CD** “turning downwind — I have the traffic on the downwind.”
- 21:39 *Archer 3CC* “turning right base 30.”
- 22:28 *Archer 3CC* “on final 30.”
- 23:01 📍 **N200CD** “turning right base 30.”
- 23:17 *Unknown* “2 miles to the north, transitioning the area at 2,500 ft.”
- 23:44 📍 **N200CD** “turning final 30.”
- 26:55 📍 **N200CD** “turning crosswind to right traffic 30.”
- 28:32 📍 **N200CD** “turning right base 30.”
- 29:29 *Mania 4DD* “right downwind 30, departing the pattern to the southeast.”
- 30:58 *Desmond 5EE* “about 2 miles to the airfield, making right traffic, runway 30.”

### VLM Safety Analysis

#### Primary Conflict (07:40–08:51) — Right-of-Way Violation on Final Approach

At 07:40, **N200CD** reports downwind with an ambiguous sequencing call (“will turn base after the traffic on short final”), suggesting awareness of the inbound but unclear intent. At 08:20, **N200CD** turns base, misjudging the closure rate of the established straight-in **N100AB**. At 08:31, **N100AB** on a 2-mile final issues a go-around threat, directly resolving the conflict: **N200CD** immediately deviates right and **N100AB** confirms visual separation and continues to a full stop.

*Instructional point.* An aircraft established on final approach has right-of-way over aircraft in the traffic pattern (FAA AC 90-66C). **N200CD** failed to yield; **N100AB** demonstrated correct airmanship by clearly announcing the conflict, which prompted immediate corrective action and prevented a potential near mid-air collision.

**Label:** HAZARD | **Confidence:** 0.91 | **Hazard type:** *simultaneous\_final / right-of-way violation* | **Model:** Gemini-2.5-Pro

## B. Qualitative Example

We probed the framework’s ability to perform safety assessment using an example at KHAF using four aviation data sources across two modalities (text and image) from real flight data. The data sources include ADS-B trajectory data, CTAF communications, METAR weather data, and a VFR sectional chart. Figure 2 presents a live CTAF recording from KHAF transcribed by Whisper Large-v3 and analyzed by Gemini 2.5 Pro. By inputting KHAF’s VFR sectional chart as an image to the model, this qualitative study serves as a preliminary investigation into the effectiveness of using VLMs for ATC safety analysis. The same figure also demonstrates the complete process of testing the qualitative real-world scenario at KHAF. Gemini 2.5 Pro was capable of accurately identifying a right-of-way violation that led to a near mid-air collision.

# IV. Synthetic Dataset and METAR+CTAF Benchmark Scenarios

## A. Dataset

For our quantitative study, we restrict the inputs to METAR weather data and CTAF communication transcripts, which are processed using a LLM. We develop and evaluate on CTAF-KHAF-Synthetic, a synthetic benchmark of 100 flight-operations scenarios at KHAF non-towered airport derived from real data examples. The full generator prompt is reproduced in Appendix VIII.A. The 12 hazard categories are drawn from incident patterns documented in FAA AC 90-66C [2] and the NASA ASRS non-tower report set [3], so each scenario type corresponds to a class of unsafe (or nominal) situation that actually occurs at non-towered airports. Pilot phraseology follows AC 90-66C self-announcement conventions, and METAR and per-aircraft ADS-B values use the same fields, units, and ranges observed in real KHAF operations.

Each scenario is first assigned a binary ground-truth safety label of either `nominal` or `danger`, reflecting whether the traffic situation contains any operational safety concern. Each scenario additionally includes a hazard-type label drawn from a 12-category taxonomy spanning communication gaps, pattern conflicts, runway incursions, instrument flight rules or VFR misalignment, and nominal operations, along with raw and decoded METAR text, a CTAF radio transcript in SRT format, and multi-voice audio synthesized from the transcript using OpenAI’s TTS-1-HD. For the finer-grained three-class formulation used in Appendix VIII.B, the `danger` class is further subdivided into `warning` and `hazard` based on collision imminence. A `warning` corresponds to a potentially unsafe situation that pilots can still resolve through standard advisory actions, whereas a `hazard` represents an imminent collision risk or serious operational conflict. Under this formulation, the labels are nearly balanced (33 nominal, 34 warning, 33 hazard). The scenarios were constructed from the 12-category hazard taxonomy by composing CTAF radio call sequences that exhibit each target safety condition, paired with per-aircraft ADS-B state vectors and a representative METAR. Each transcript was then synthesized into multi-voice audio using OpenAI’s TTS-1-HD, with a distinct voice assigned to each aircraft on the frequency so that speaker attribution is preserved in the audio. The complete dataset and code are available [here](#).

Every scenario including its ground-truth safety label, hazard-type label, and FAA-conforming advisory text was reviewed by human experts with GA pilot flight experience to verify that the assigned label is unambiguous and that the radio call sequence would be plausible in real CTAF operations. Six held-out scenarios, two per class, are reserved as in-context learning (ICL) examples and are excluded from evaluation. The remaining 94 scenarios constitute the test set used in every experiment reported in this article. A representative sample is shown in the S003: Simultaneous Final on Runway 30 box below, illustrating the multi-aircraft conversational structure typical of CTAF traffic, the natural co-occurrence of incomplete or conflicting position calls, and the FAA-conforming advisory text that defines the ground-truth label. METAR codes provide weather information to pilots operating near an airfield, assisting with flight planning and instrument adjustments. The corresponding synthetic METAR text data was sourced from NOAA’s Aviation Weather Center [25].

## B. LLM Evaluation

We benchmark six LLMs spanning open-source and closed-source models quantitatively using the synthetic dataset. The three open-source LLMs include: Qwen 2.5-7B-Instruct [26], Mistral-7B-Instruct-v0.3 [27], and Gemma-2-9B-IT [28]. The open-source models are run locally in fp16 with no quantization. The three closed-source models: GPT-4o, GPT-5.4, and Claude Sonnet 4.6 are accessed through their respective HTTP APIs. All six LLMs are frozen at evaluation time. The qualitative experiment is ran using the Gemini 2.5 Pro model [29]. Each scenario is evaluated under three prompting strategies that vary the number of examples. Zero-shot: The system prompt and target scenario only, no

exemplars. One-shot: One held-out exemplar per class (3 total). Few-shot: Two held-out exemplars per class (6 total). Exemplars are drawn from the fixed 6-scenario ICL pool and never appear in the test set. The same prompt template is run under two reasoning protocols. In direct prompting the model emits the structured JSON in a single turn. In chain-of-thought (CoT) prompting [30], a first turn elicits step-by-step reasoning (“Before classifying, reason step by step. . .”) and a second turn extracts the JSON from that reasoning.

Each combination of model, strategy, and reasoning protocol constitutes one evaluation condition, yielding a  $6 \times 3 \times 2 = 36$ -condition design per task framing. A fixed system prompt anchors the LLM in the safety-classification task. It establishes the airport, runway, and traffic-pattern conventions following FAA AC 90-66C [2]; defines each safety class with a list of distinguishing criteria. The prompt then prescribes a strict JSON output schema. The complete prompt for the binary task is shown in the **System Prompt** box below. A three-class prompt used in Appendix VIII.B follows the same structure but splits the **danger** class into **warning** and **hazard**. Every prediction is parsed from a JSON object with three fields: **label** (the categorical safety class), **confidence** (a self-reported scalar in [0, 1]), and **reasoning** (a one-sentence free-text advisory in FAA-conforming phraseology). The structured fields drive quantitative evaluation; the **reasoning** string is a human-readable advisory that can be displayed to pilots directly.

### Sample Scenario — S003: Simultaneous Final on Runway 30

**Setting.** Half Moon Bay Airport (KHAF), runway 30, right-traffic pattern.

**METAR.** KHAF 142135Z AUTO 18005KT 5SM -BR FEW010 BKN020 18/16 A2999 RMK A02

*Decoded:* Marginal VFR — 5 SM visibility in mist, broken ceiling at 2,000 ft, wind 180° at 5 kt, 18°C / dewpoint 16°C.

#### Aircraft on frequency.

- ▶ **N910YZ** — Cessna 172, two-mile straight-in RNAV final runway 30, *full stop*.  
ADS-B ( $t=0$ ): 37.4967°N, 122.4644°W; 800 ft MSL; heading 300°; 85 kt.
- ♦ **N602SK** — Piper Seneca, right base runway 30, *touch-and-go*.  
ADS-B ( $t=0$ ): 37.5147°N, 122.4756°W; 900 ft MSL; heading 210°; 85 kt.

**CTAF transcript.** *The standard CTAF call frame “Half Moon Bay traffic, . . . , Half Moon Bay” is omitted for brevity.*

00:00.0 📡 **N602SK** “right base runway three zero, touch and go.”  
 00:04.8 📡 **N910YZ** “two-mile straight-in RNAV final runway three zero, full stop.”  
 00:10.1 📡 **N602SK** “turning right final runway three zero, touch and go.”  
 00:15.3 📡 **N910YZ** “one-and-a-half-mile final runway three zero, full stop.”  
 00:21.0 📡 **N602SK** “short final runway three zero, touch and go.”  
 00:25.5 📡 **N910YZ** “one-mile final runway three zero, full stop.”  
 00:31.0 📡 **N910YZ** “traffic on short final, say position.”  
 00:36.1 📡 **N602SK** “on short final three zero, negative contact.”  
 00:41.0 📡 **N910YZ** “half-mile final, I have traffic now, you’re directly below me.”  
 00:46.2 📡 **N602SK** “traffic in sight now, you’re overtaking us on final.”  
 00:51.3 📡 **N910YZ** “going around runway three zero, traffic conflict on final.”  
 00:57.1 📡 **N602SK** “continuing runway three zero, near midair on short final.”

**Ground-truth advisory** (FAA-conforming).

*“Traffic alert: N910YZ, Cessna one-mile straight-in final Runway Three Zero, and N602SK, Seneca short final Runway Three Zero, converging on the same runway. N910YZ, go around immediately; N602SK, continue landing or clear the runway without delay; both aircraft maintain visual separation.”*

**Ground-truth label:** **HAZARD** | **Hazard type:** *simultaneous\_final* | **Source:** CTAF-KHAF benchmark, scenario S003

## V. Quantitative Results

We evaluate all six models on the 94-scenario test split of CTAF-KHAF-Synthetic across 36 evaluation conditions (six models  $\times$  three prompting strategies  $\times$  two reasoning protocols). The primary task is binary safety classification (nominal/danger), while a finer-grained three-class formulation (nominal/warning/hazard) is reported in Appendix VIII.B. Additional robustness ablations targeting ASR quality, audio noise, and transcript masking are reported in Appendix VIII.C. Table 2 reports macro- $F_1$ , accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC) across all evaluation conditions, while Table 1 reports per-class  $F_1$  scores. Every LLM exceeds macro- $F_1 = 0.85$  in its best configuration; the strongest configuration is Qwen-2.5-7B at zero-shot direct ( $F_1 = 0.964$ , AUROC = 0.995), with Claude Sonnet 4.6 close behind at few-shot + CoT ( $F_1 = 0.952$ ). GPT-4o and Gemma-2-9B both reach  $F_1 \approx 0.93$ . Most models approach near-ceiling performance on the binary task, although notable failures

### System Prompt — Binary Safety Classification

You are an automated aviation safety monitoring system for Half Moon Bay Airport (KHAF), a non-towered airport near San Francisco, California. Your task is to analyze CTAF (Common Traffic Advisory Frequency) radio communications at KHAF and classify the safety status of the current traffic situation.

#### Inputs

- METAR weather data for KHAF (raw + decoded)
- CTAF radio transcript (SRT format with timestamps)

**Task.** Classify the situation as exactly one of **nominal** or **danger**.

**NOMINAL** — all is well.

- All required position calls are present (crosswind, downwind, base, final)
- Traffic is sequenced and separated with no conflicts
- Weather is VMC and appropriate for operations
- Single aircraft announcing each leg, no other traffic

**DANGER** — any potential or imminent safety issue. Use **danger** whenever there is *any* conflict, communication gap, or unsafe condition:

- Communication gaps: missing position calls, NORDO traffic, delayed announcements
- Pattern conflicts: converging traffic, wrong-runway calls, improper entries
- Active conflicts: simultaneous final, runway incursions, mid-air risk
- Weather mismatches: VFR pilot inadvertently in IMC
- Late or omitted go-around announcements
- Any situation a CTAF advisory would flag as caution, alert, or emergency

**Key question.** “Would a CTAF advisory flag this for any reason (caution, alert, or emergency)?” If yes  $\Rightarrow$  **danger**.

**CTAF rules** (FAA AC 90-66C [2]).

- Pilots must self-announce: crosswind, downwind, base, final, runway clear
- Straight-in: announce at 10, 5, and 3 NM
- Go-around must be announced immediately
- No ATC — pilots are solely responsible for separation

**Output format.** Respond with *only* the following JSON, no other text:

```
{
  "label": "<nominal | danger>",
  "confidence": <0.0-1.0>,
  "reasoning": "<one sentence stating the key safety factor>"
}
```

remain for Mistral-7B under one-shot and few-shot CoT prompting, where macro- $F_1$  drops to approximately 0.30.

### A. Effect of ICL and CoT

Figure 4 shows macro- $F_1$  vs. strategy on the binary task. The within-model trends show that most LLMs benefit modestly from additional ICL exemplars, and CoT yields mixed effects (Fig. 3). Mistral’s CoT collapse is even more dramatic in the binary framing, dropping the danger-class  $F_1$  from 0.923 at zero-shot + CoT to 0.091 at few-shot + CoT (Table 1). Conversely, Qwen at one-shot + CoT matches its zero-shot direct performance ( $F_1 = 0.964$ ), suggesting the open-source CoT picture is highly model-specific rather than uniformly beneficial or harmful.

### B. Confusion Structure

Table 3 summarizes the  $2 \times 2$  confusion structure for the best-performing configuration of each LLM, broken out into true-negative (TN), false-positive (FP), false-negative (FN), and true-positive (TP) counts together with the safety-relevant  $FN$  rate (missed-danger) and  $FP$  rate (false-alarm). A clear difference emerges between the larger closed-source models and the smaller open-source models. GPT-5.4 exhibits a conservative over-alerting behavior, achieving zero missed-danger cases (FN rate 0.0%) at the cost of a high false-alarm rate (32.3% of nominal scenarios misclassified as danger). Claude Sonnet 4.6 and GPT-4o maintain a more balanced trade-off, with low FN rates (3.2% and 9.5% respectively) while keeping false alarms relatively limited (6.5% for Claude and 0.0% for GPT-4o). In contrast, the smaller open-source models tend to under-alert rather than over-alert. Mistral-7B misses 14.3% of true-danger scenarios while issuing no false alarms, and Gemma-2-9B misses 9.5% with the same zero-false-alarm profile. Qwen-2.5-7B is the strongest open-source model, reducing the FN rate to only 1.6% while preserving a low false-positive rate (6.5%). Overall, the table shows that larger models are generally more willing to issue cautionary danger predictions, whereas smaller open-source models (with the exception of Qwen) are more likely to miss hazardous situations.

### C. Threshold-Independent Ranking Quality

Figure 5 reports PR and ROC curves for the best run of each LLM. All six LLMs hug the top-left corner of the ROC and the top-right corner of the PR plot, with  $AP \geq 0.97$  and  $AUROC \geq 0.95$  for every model. Gemma-2-9B and Claude Sonnet 4.6 attain the highest AP (0.996 each); Qwen and GPT-4o are also at  $\geq 0.99$ . GPT-5.4 trails at  $AP = 0.969 / AUROC = 0.952$ , in part because its score is necessarily derived from self-reported confidence rather than token logprobs (the GPT-5 series rejects the logprobs parameter at the time of writing); Claude Sonnet 4.6 is similarly

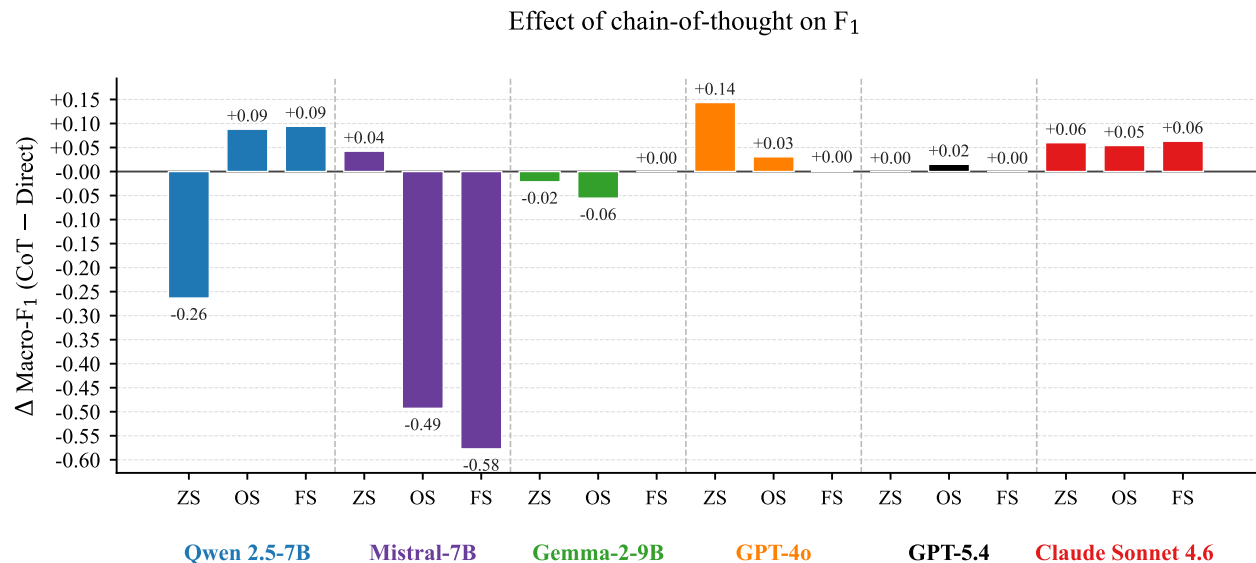
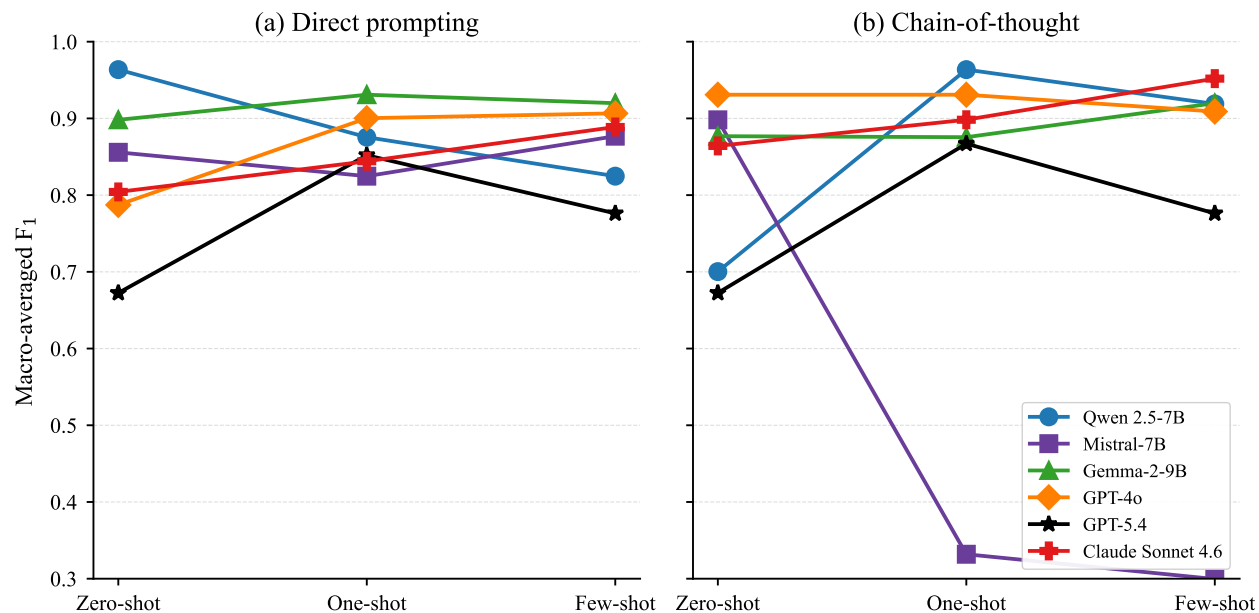


Fig. 3 Per-condition macro- $F_1$  change from adding chain-of-thought on the binary task. Mistral’s collapse is now visible at every non-zero-shot setting; GPT-4o and Claude show modest gains.

**Table 1 Per-class  $F_1$  scores on the binary CTAF-KHAF benchmark (Nominal vs. Danger) across all models, prompting strategies, and reasoning methods. Best per (model, class) row in bold.**

Model	Class	Direct prompting			Chain-of-thought		
		ZS	OS	FS	ZS	OS	FS
<i>Open-source</i>							
Qwen 2.5-7B	Nominal	<b>0.951</b>	0.845	0.795	0.560	<b>0.951</b>	0.896
	Danger	<b>0.976</b>	0.906	0.855	0.841	<b>0.976</b>	0.942
Mistral-7B	Nominal	0.827	0.795	0.849	<b>0.873</b>	0.517	0.508
	Danger	0.885	0.855	0.904	<b>0.923</b>	0.147	0.091
Gemma-2-9B	Nominal	0.873	<b>0.912</b>	0.899	0.849	0.845	0.899
	Danger	0.923	<b>0.950</b>	0.941	0.904	0.906	0.941
<i>Closed-source</i>							
GPT-4o	Nominal	0.681	0.862	0.879	<b>0.912</b>	<b>0.912</b>	0.886
	Danger	0.894	0.938	0.934	<b>0.950</b>	<b>0.950</b>	0.932
GPT-5.4	Nominal	0.488	0.784	0.667	0.488	<b>0.808</b>	0.667
	Danger	0.857	0.920	0.886	0.857	<b>0.926</b>	0.886
Claude Sonnet 4.6	Nominal	0.733	0.778	0.847	0.814	0.857	<b>0.935</b>
	Danger	0.875	0.910	0.930	0.915	0.939	<b>0.968</b>



**Fig. 4 Macro- $F_1$  vs. prompting strategy on the binary task. Direct (left) and CoT (right). Most LLMs exceed  $F_1 = 0.85$  in their best configuration; Mistral + CoT at higher ICL counts is the conspicuous outlier.**

flagged. Importantly, the high AUROC values indicate that the model’s score-space rank ordering of dangerous vs. nominal scenarios is reliable across operating points.

**Table 2 Binary classification results (Nominal vs. Danger) on the CTAF-KHAF benchmark. Best per (model, metric) row in bold. AUROC is computed against the danger class using token logprobs where available.**

Model	Metric	Direct prompting			Chain-of-thought		
		ZS	OS	FS	ZS	OS	FS
<i>Open-source</i>							
Qwen 2.5-7B	Macro-F <sub>1</sub>	<b>0.964</b>	0.876	0.825	0.700	<b>0.964</b>	0.919
	Accuracy	<b>0.968</b>	0.883	0.830	0.766	<b>0.968</b>	0.926
	AUROC	<b>0.995</b>	0.973	0.983	0.994	0.979	0.984
Mistral-7B	Macro-F <sub>1</sub>	0.856	0.825	0.877	<b>0.898</b>	0.332	0.300
	Accuracy	0.862	0.830	0.883	<b>0.904</b>	0.383	0.362
	AUROC	0.897	0.968	<b>0.984</b>	0.942	0.548	0.425
Gemma-2-9B	Macro-F <sub>1</sub>	0.898	<b>0.931</b>	0.920	0.877	0.876	0.920
	Accuracy	0.904	<b>0.936</b>	0.926	0.883	0.883	0.926
	AUROC	0.951	0.979	<b>0.982</b>	0.978	0.969	0.968
<i>Closed-source</i>							
GPT-4o	Macro-F <sub>1</sub>	0.787	0.900	0.907	<b>0.931</b>	<b>0.931</b>	0.909
	Accuracy	0.840	0.915	0.915	<b>0.936</b>	<b>0.936</b>	0.915
	AUROC	<b>0.991</b>	0.983	0.976	0.977	0.988	0.978
GPT-5.4	Macro-F <sub>1</sub>	0.672	0.852	0.776	0.672	<b>0.867</b>	0.776
	Accuracy	0.777	0.883	0.830	0.777	<b>0.894</b>	0.830
	AUROC	0.902	0.935	0.924	0.935	<b>0.952</b>	0.915
Claude Sonnet 4.6	Macro-F <sub>1</sub>	0.804	0.844	0.889	0.864	0.898	<b>0.952</b>
	Accuracy	0.830	0.872	0.904	0.883	0.915	<b>0.957</b>
	AUROC	0.937	0.974	0.964	0.961	0.982	<b>0.994</b>

## VI. Limitations and Future Work

A key limitation of the proposed approach lies in the computational cost and system-level efficiency of LLM/VLMs. Although we employ a frozen model without fine-tuning, inference over multimodal inputs remains resource-intensive, posing challenges for real-time or edge deployment in general aviation settings. Recent work has shown that efficient deployment of LLMs requires careful system-level optimization across memory, compute, and scheduling layers, rather than model design alone [31]. Our current architecture does not incorporate such optimizations, and therefore may not yet meet the latency and power constraints required for continuous onboard operation.

A second limitation concerns the data on which the system is evaluated. Real-world CTAF recordings paired with ground-truth safety labels are scarce in the public domain, and curating them at scale would require dedicated audio collection at multiple non-towered airports together with expert annotation. As a stand-in we constructed CTAF-KHAF-Synthetic, which captures a structured 12-category hazard taxonomy at a single airport but cannot fully reproduce the acoustic noise, accent variation, and timing irregularities of live radio traffic. The quantitative results in this article should therefore be read as a measurement on a controlled, single-airport benchmark rather than a deployment-ready estimate. A third limitation follows directly from the design of this synthetic dataset: it does not include per-scenario VFR sectional imagery or weather radar imagery, so the VLM side of the architecture cannot be exercised quantitatively on the benchmark. We were therefore restricted to evaluating LLMs on the textual subset of the inputs (METAR text and CTAF transcript), and could only demonstrate the full framework qualitatively, on a single real recording at KHAF (Sec III.B). A larger benchmark with paired visual modalities would be required for a like-for-like quantitative comparison between LLMs and VLMs in this domain.

Future work includes extending this study into a larger, controlled benchmark in which every scenario is paired with all four data sources (METAR, ADS-B, VFR sectional, and weather radar imagery). At first starting with synthetic data, and then moving into real flight data from the same non-towered airport. This larger more multimodal study would enable a like-for-like quantitative evaluation to our qualitative study. Our use of frozen models was also a limitation, and

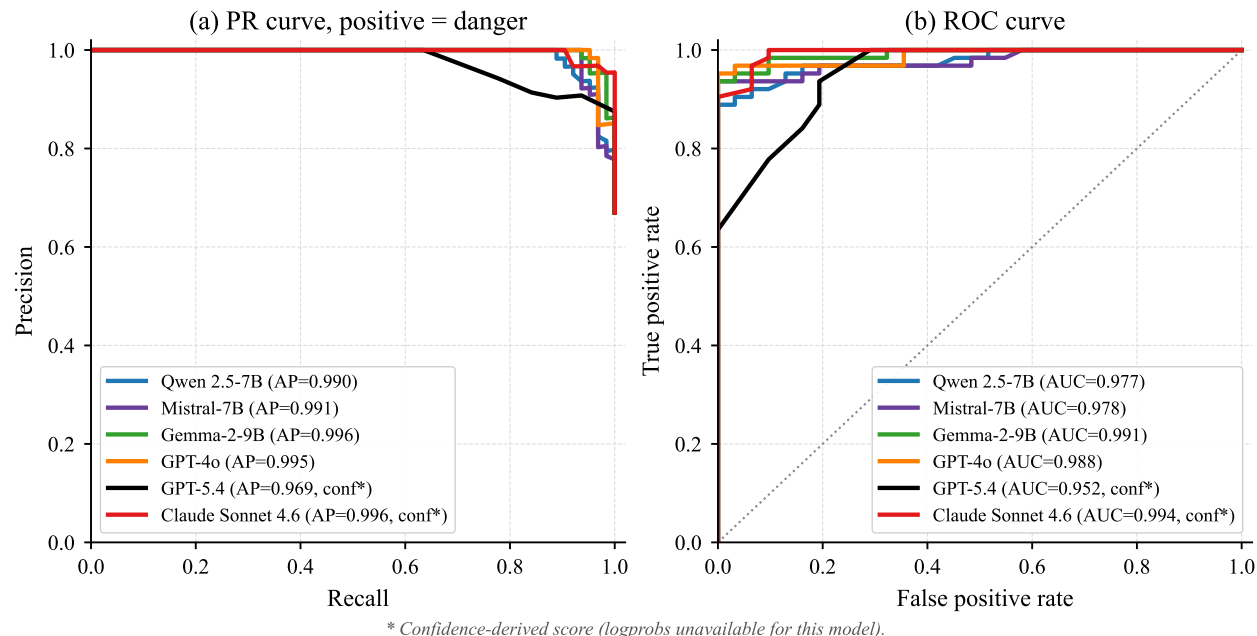
**Table 3 Confusion structure for the best run of each LLM on the binary task (positive class = *danger*). Rates are computed out of  $N = 31$  true-nominal and  $D = 63$  true-danger scenarios in the 94-scenario test split: TN rate (specificity) =  $TN/N$ , FP rate (false-alarm) =  $FP/N$ , TP rate (recall) =  $TP/D$ , FN rate (missed-danger) =  $FN/D$ . Arrows indicate the desired direction for a safety advisor.**

Source	Model	Best run	TN rate $\uparrow$	FP rate $\downarrow$	TP rate $\uparrow$	FN rate $\downarrow$
<i>Open-source</i>	Qwen 2.5-7B	OS+CoT	93.5%	6.5%	98.4%	1.6%
	Mistral-7B	ZS+CoT	100.0%	0.0%	85.7%	14.3%
	Gemma-2-9B	OS	100.0%	0.0%	90.5%	9.5%
<i>Closed-source</i>	GPT-4o	OS+CoT	100.0%	0.0%	90.5%	9.5%
	GPT-5.4	OS+CoT	67.7%	32.3%	100.0%	0.0%
	Claude Sonnet 4.6	FS+CoT	93.5%	6.5%	96.8%	3.2%

fine-tuning could improve performance; this remains future work. Future work will also include evaluating the affect of the VFR sectional chart on the models performance, justifying the usage of VLMs over LLMs on the same task.

## VII. Conclusion

This article proposed an VLM-based safety assessment framework for non-towered airports that inputs CTAF radio transcripts, and METAR weather data, ASD-B trajectory data, and VFR sectional images, and produce both a structured safety label and a free-text CTAF-style advisory. We demonstrated an initial proof of concept with an example at KHAF run with Gemini 2.5 Pro (VLM) which was able to correctly identify a right-of-way violation that produced a near mid-air collision at KHAF. We developed a synthetic dataset to quantitatively evaluate performance using the METAR and CTAF data, and benchmarked six frozen LLMs on a binary safety-classification task, where every model exceeded macro  $F_1$  of 0.85 in its best configuration and the strongest open-source result tied the strongest closed-source result.



**Fig. 5 Precision-recall and receiver-operating-characteristic curves on the binary task, one curve per LLM (best run by macro- $F_1$ ). conf\* marks score-distribution-based fallback for models whose APIs do not expose token logprobs.**

## References

- [1] Mogford, R. H., Guttman, J. A., Morrow, S., Kopardekar, P., et al., “The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature,” 1995.
- [2] Federal Aviation Administration, “Advisory Circular 90-66C: Non-Towered Airport Flight Operations,” Tech. rep., U.S. Department of Transportation, Federal Aviation Administration, June 2023. URL [https://www.faa.gov/documentlibrary/media/advisory\\_circular/ac\\_90-66c.pdf](https://www.faa.gov/documentlibrary/media/advisory_circular/ac_90-66c.pdf).
- [3] National Aeronautics and Space Administration (NASA), Ames Research Center, “ASRS Database Report Set: Non-Tower Airport Incidents,” Technical Memorandum, ASRS Report Set39 TH: 262-7, NASA Ames Research Center, Jun. 2025. URL [https://asrs.arc.nasa.gov/docs/rpsts/non\\_twr.pdf](https://asrs.arc.nasa.gov/docs/rpsts/non_twr.pdf), update No. 39, June 12 2025.
- [4] National Academies of Sciences, E., and Medicine, “In-Time Aviation Safety Management: Challenges and Research for an Evolving Aviation System,” Tech. rep., The National Academies Press, 2022. URL <https://www.faa.gov/sites/faa.gov/files/2022-05/508.NationalAcademyReport.pdf>, section “System Monitoring – Data Fusion, Completeness, and Quality” points out that “the range of IASMS capabilities . . . will be limited by the completeness of the data available, by its quality and consistency, by the ability to fuse it. . .”.
- [5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I., “Attention is All you Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
- [6] Cui, W., Yu, D., Jiao, X., Meng, Z., Zhang, G., Wang, Q., Guo, Y., and King, I., “Recent Advances in Speech Language Models: A Survey,” *Proceedings of ACL (Long Paper)*, 2025. URL <https://aclanthology.org/2025.acl-long.682.pdf>, preprint available at arXiv.
- [7] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., and Chen, E., “A Survey on Multimodal Large Language Models,” *National Science Review*, Vol. 11, No. 12, 2024, p. nwae403. <https://doi.org/10.1093/nsr/nwae403>.
- [8] Stahlberg, F., “Neural Machine Translation: A Review,” *Journal of Artificial Intelligence Research (JAIR)*, 2020. URL <https://jair.org/index.php/jair/article/download/12007/26611/24616>, cited > 600times.
- [9] Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., and Bäck, T., “Reasoning with Large Language Models: A Survey,” 2024. URL <https://arxiv.org/html/2407.11511v1>, arXiv preprint.
- [10] Cheng, S., Zhuang, Z., Xu, Y., Yang, F., Zhang, C., Qin, X., Huang, X., Chen, L., Lin, Q., Zhang, D., Rajmohan, S., and Zhang, Q., “Call Me When Necessary: LLMs can Efficiently and Faithfully Reason over Structured Environments,” *Findings of the Association for Computational Linguistics: ACL 2024*, edited by L.-W. Ku, A. Martins, and V. Srikumar, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 4275–4295. <https://doi.org/10.18653/v1/2024.findings-acl.254>, URL <https://aclanthology.org/2024.findings-acl.254/>.
- [11] Kuchar, J., and Wan, L. M., “The Traffic Alert and Collision Avoidance System (TCAS): Past, Present and Future,” *Lincoln Laboratory Journal*, Vol. 16, No. 2, 2007, p. 14–32. Survey of airborne automated collision avoidance systems.
- [12] Al-Basman, M., and Hu, J., “Probability of conflict analysis of 3D aircraft flight based on two-level Markov chain approximation approach,” *2010 International Conference on Networking, Sensing and Control (ICNSC)*, IEEE, 2010, pp. 608–613.
- [13] Kochenderfer, M. J., Holland, J. E., and Chryssanthacopoulos, J. P., “Next-generation airborne collision avoidance system,” 2012.
- [14] Paradis, C. V., Hong, C., Matthews, B., Davies, M. D., and Hooey, B., “Kaona: Deep Searching and Curating Aviation Safety Reporting Systems,” *AIAA SCITECH 2025 Forum*, 2025. <https://doi.org/10.2514/6.2025-2152>, URL <https://arc.aiaa.org/doi/abs/10.2514/6.2025-2152>.
- [15] Abdulhak, S., Hubbard, W., Gopalakrishnan, K., and Li, M. Z., “CHATATC: Large Language Model-Driven Conversational Agents for Supporting Strategic Air Traffic Flow Management,” *arXiv preprint*, Vol. arXiv:2402.14850, 2024. URL <https://arxiv.org/abs/2402.14850>.
- [16] Mahboubi, Z., and Kochenderfer, M. J., “Autonomous Air Traffic Control for Non-Towered Airports,” *Proc. USA/Eur. Air Traffic Manage. Res. Develop. Seminar*, 2015, pp. 1–6.

- [17] Tikayat Ray, A., Bhat, A. P., White, R. T., Nguyen, V. M., Pinon Fischer, O. J., and Mavris, D. N., “Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS),” *Aerospace*, Vol. 10, No. 9, 2023, p. 770.
- [18] Nielsen, D., Clarke, S. S., and Kalyanam, K. M., “Towards an Aviation Large Language Model by Fine-tuning and Evaluating Transformers,” *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, IEEE, 2024, pp. 1–5.
- [19] Wang, L., Chou, J., Tien, A., Zhou, X., and Baumgartner, D., “AviationGPT: A Large Language Model for the Aviation Domain,” *AIAA AVIATION FORUM AND ASCEND 2024*, 2024, p. 4250.
- [20] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., “LLama: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [21] Gao, Y., Yue, L., Sun, J., Shan, X., Liu, Y., and Wu, X., “WorkloadGPT: A Large Language Model Approach to Real-Time Detection of Pilot Workload,” *Applied Sciences*, Vol. 14, No. 18, 2024, p. 8274.
- [22] Tabrizian, A., Ghazanfari, M., and Wei, P., “Chain-of-Thought Flight Planner: End-to-End LLM Routing Under Wind Hazards,” *AIAA AVIATION FORUM AND ASCEND 2025*, 2025, p. 3711.
- [23] Andriūškevičius, J., and Sun, J., “Automatic Control With Human-Like Reasoning: Exploring Language Model Embodied Air Traffic Agents,” *arXiv preprint arXiv:2409.09717*, 2024.
- [24] Sangeetha, S. V., Chiu, C.-Y., Li, S. H., and Kousik, S., “Language Conditioning Improves Accuracy of Aircraft Goal Prediction in Untowered Airspace,” *arXiv preprint arXiv:2509.14063*, 2025.
- [25] Aviation Weather Center, “AviationWeather.gov,” <https://aviationweather.gov/>, 2025. Accessed: 2025-11-07.
- [26] Team, Q., et al., “Qwen2 Technical Report,” *arXiv preprint arXiv:2407.10671*, Vol. 2, No. 3, 2024.
- [27] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W., “Mistral 7B,” *arXiv preprint arXiv:2310.06825*, 2023.
- [28] Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al., “Gemma 2: Improving open language models at a practical size,” *arXiv preprint arXiv:2408.00118*, 2024.
- [29] Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al., “Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities,” *arXiv preprint arXiv:2507.06261*, 2025.
- [30] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al., “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, 2022, pp. 24824–24837.
- [31] Gogineni, K., Suvizi, A., and Venkataramani, G., “LLMs on a budget: System-level approaches to power-efficient and scalable fine-tuning,” *IEEE Open Journal of the Computer Society*, 2025.

## VIII. Appendix

### Dataset Generator Prompts — CTAF-KHAF-SYNTHETIC

**User message (per scenario).** For every scenario, the procedurally sampled aircraft, position events, and METAR are formatted into a single user message:

```
SCENARIO: {hazard_type} ({label})
METAR: {raw METAR text}
DURATION: ~{duration}s
AIRCRAFT:
{callsign A} ({type}) - radio|NORDO
{callsign B} ({type}) - radio|NORDO
POSITION EVENTS:
t={t}s {callsign} {phase} {dist}NM {alt}ft radio|NORDO
...
Write the SRT transcript.
```

This user message is sent to GPT-4o twice, each time paired with one of the two system prompts shown below.

#### System Prompt 1 — CTAF Transcript Generation.

You generate realistic CTAF radio transcripts for Half Moon Bay Airport (KHAF), runway 30, right-traffic pattern. Given exact aircraft positions, write an SRT-format transcript of pilot radio calls.

```
FORMAT (strict):
{index}
{HH:MM:SS,mmm} --> {HH:MM:SS,mmm}
{text}
RULES:
- Use NATO phonetic alphabet for letters (Alpha, Bravo...) and "niner" for 9.
- Each self-announced call: "Half Moon Bay traffic, [callsign], [position], [runway 30], [intention], Half Moon Bay."
- NORDO aircraft: only mentioned by other pilots.
- Timing: each utterance 3-6 s; gap between calls 3-8 s. Timestamps start at 00:00:00,000.
- CRITICAL: total scenario duration MUST be under 90 s; if the last timestamp would exceed 90 s, stop writing calls early.
- CRITICAL: write at most 10 lines total; stop at 10 even if not all events are covered.
- Return ONLY raw SRT - no markdown fences, no triple-backtick blocks.
- Cover the KEY position events only - not every single distance update.
- Pilots call position at major phase changes: entering downwind, turning base, turning final, short final, going around, clear of runway.
- Do NOT have pilots repeat the same position multiple times unless there is a conflict.
- For IMC / disoriented pilots: write hesitant, confused speech ("uh", "I got...", corrections).
- Return ONLY the SRT content.
```

#### System Prompt 2 — Ground-Truth Safety Advisory.

You are an AI aviation safety advisor monitoring CTAF at KHAF (Half Moon Bay Airport). Write a concise ground-truth safety advisory (2-4 sentences, ~100-200 words) based on the scenario. Identify aircraft by callsign and type, state their positions precisely, assess the safety situation, and give recommended actions if needed. Return ONLY the advisory text.

### A. Synthetic Dataset Generation Prompt

The CTAF-KHAF-SYNTHETIC benchmark is constructed in two stages. First, for each of the twelve hazard-type categories in our taxonomy (e.g., *simultaneous\_final*, *silent\_traffic*, *nominal\_single\_aircraft*), a deterministic Python procedure samples aircraft callsigns and types, decides which aircraft are NORDO, generates a sequence of position events at specific timestamps (entering downwind, turning base, turning final, short final, going around, clearing the runway, and so on), and draws a representative METAR for Half Moon Bay Airport (KHAF). Second, these per-scenario events are packed into a structured user message and passed to GPT-4o twice, with two different system prompts: once with the *Transcript* system prompt to produce a strict SRT-format pilot-radio transcript that follows FAA AC 90-66C phraseology, and once with the *Advisory* system prompt to produce a two-to-four-sentence ground-truth safety advisory in FAA-conforming language. Both LLM calls use temperature 0.7 and a 1,200-token output cap, wrapped in retry logic with exponential backoff. The complete user-message template and the two system prompts are reproduced in the box

below. After generation, every scenario was reviewed by human experts with general-aviation flight experience to verify that the radio call sequence is plausible in real CTAF operations and that the assigned safety label is unambiguous (see Sec. IV of the main paper). The verified transcripts are then synthesized into multi-voice MP3 audio using OpenAI’s TTS-1-HD, with a distinct voice assigned to each aircraft on the frequency so that speaker attribution is preserved in the audio.

### B. Three-Class Classification

The system prompt used in the three-class classification is shown in the box below. Aggregated macro- $F_1$  and accuracy values for all six LLMs, three prompting strategies, and two reasoning protocols are reported in Table 4; per-class  $F_1$  for the same 36 conditions is in Table 5. The three-class task is increased in difficulty: the strongest configuration (GPT-4o, Few-shot + CoT) reaches macro- $F_1 = 0.828$  and accuracy 0.830, leaving a clear gap to perfect classification. Closed-source LLMs dominate the leaderboard—GPT-4o (best configuration  $F_1 = 0.828$ ), Claude Sonnet 4.6, and GPT-5.4 ( $F_1 = 0.820$ ) all sit in the 0.82–0.83 band, well above the best open-source results (Qwen  $F_1 = 0.743$ , Gemma-2-9B  $F_1 = 0.726$ , Mistral-7B  $F_1 = 0.640$ ). The closed-source LLMs also remain consistent across prompting strategies, whereas the open-source LLMs exhibit substantial within-model variance.

**Table 4 Three-class classification results (Nominal / Warning / Hazard) on the CTAF-KHAF benchmark. Best per (model, metric) row in bold.**

Model	Metric	Direct prompting			Chain-of-thought		
		ZS	OS	FS	ZS	OS	FS
<i>Open-source</i>							
Qwen 2.5-7B	Macro- $F_1$	0.687	0.615	0.515	0.416	0.692	<b>0.743</b>
	Accuracy	0.681	0.628	0.553	0.479	0.681	<b>0.755</b>
Mistral-7B	Macro- $F_1$	0.560	0.570	<b>0.640</b>	0.390	0.504	0.520
	Accuracy	0.585	0.638	<b>0.681</b>	0.468	0.617	0.628
Gemma-2-9B	Macro- $F_1$	0.663	0.666	<b>0.726</b>	0.456	0.616	0.687
	Accuracy	0.670	0.670	<b>0.755</b>	0.500	0.606	0.681
<i>Closed-source</i>							
GPT-4o	Macro- $F_1$	0.791	0.766	0.789	0.764	0.781	<b>0.828</b>
	Accuracy	0.798	0.777	0.798	0.766	0.777	<b>0.830</b>
GPT-5.4	Macro- $F_1$	0.782	<b>0.820</b>	0.818	0.747	0.808	0.759
	Accuracy	0.777	<b>0.819</b>	<b>0.819</b>	0.745	0.809	0.766
Claude Sonnet 4.6	Macro- $F_1$	0.792	0.715	0.675	0.770	0.761	<b>0.827</b>
	Accuracy	0.787	0.723	0.713	0.777	0.766	<b>0.830</b>

Figure 6 plots macro- $F_1$  as a function of prompting strategy, separating direct prompting from chain-of-thought. Adding ICL exemplars improves most LLMs under direct prompting (e.g., Gemma-2-9B improves from 0.663 at zero-shot to 0.726 at few-shot), but degrades performance for others (e.g., Claude Sonnet 4.6 drops from 0.792 at zero-shot to 0.675 at few-shot). Under CoT, the trend reverses for some models: Mistral-7B’s hazard-class  $F_1$  drops from 0.453 at zero-shot + CoT to 0.000 at one-shot + CoT and few-shot + CoT (Table 5), indicating that the additional exemplars compounded a failure mode rather than mitigating it.

Figure 7 reports the per-condition  $F_1$  delta between CoT and direct prompting. The pattern is striking: CoT helps Qwen at the few-shot setting ( $\Delta F_1 = +0.23$ ) and offers small gains for closed-source LLMs, but it consistently hurts Mistral (deltas of  $-0.17$  to  $-0.12$ ) and Gemma ( $-0.21$  to  $-0.04$ ). On inspection of individual records, Mistral’s CoT failure is not a parsing or reasoning error: turn-1 reasoning frequently identifies the correct hazard, but the JSON-extraction turn discards it and emits a generic “nominal” or “warning” template. This is a known weakness of small instruction-tuned LLMs on multi-turn structured-output tasks. Figure 8 shows per-hazard accuracy for the best run of each LLM. Two patterns emerge: (i) unambiguous hazards (*simultaneous final, nominal single-aircraft, nominal instrument approach*) are solved by every LLM with accuracy near 1.0; (ii) scenarios that hinge on *imminence* rather

### System Prompt — Three-Class Safety Classification

You are an automated aviation safety monitoring system for Half Moon Bay Airport (KHAF), a non-towered airport near San Francisco, California. Your task is to analyze multimodal flight-operations data and classify the safety status of the current traffic situation.

#### Inputs

- METAR weather data for KHAF (raw + decoded)
- CTAF radio transcript (SRT format with timestamps)

**Task.** Classify the situation as exactly one of **nominal**, **warning**, or **hazard**.

**NOMINAL** — all is well.

- All required position calls are present (crosswind, downwind, base, final)
- Traffic is sequenced and separated, no conflicts
- Weather is VMC and appropriate for operations

**WARNING** — a potential problem exists but no collision is imminent yet.

- An aircraft flying the wrong pattern direction without conflict
- Two aircraft converging on final with separation > 0.5 NM
- Missing position calls from one aircraft, no immediate conflict

*Key question.* “Can the pilots resolve this themselves with standard advisory actions?” If yes ⇒ **warning**.

**HAZARD** — a collision or serious incident is imminent or already occurring.

- Two aircraft simultaneously on final for the same runway (< 0.5 NM)
- An aircraft on the runway while another is on short final
- Wrong-runway announcement during an active approach
- Same altitude and converging — mid-air collision risk

*Key question.* “Would a CTAF advisory say IMMEDIATELY OR SAFETY ALERT?” If yes ⇒ **hazard**.

**Critical distinction.** The difference between **warning** and **hazard** is **imminence**, not severity.

**Output format.** Respond with *only* the following JSON, no other text:

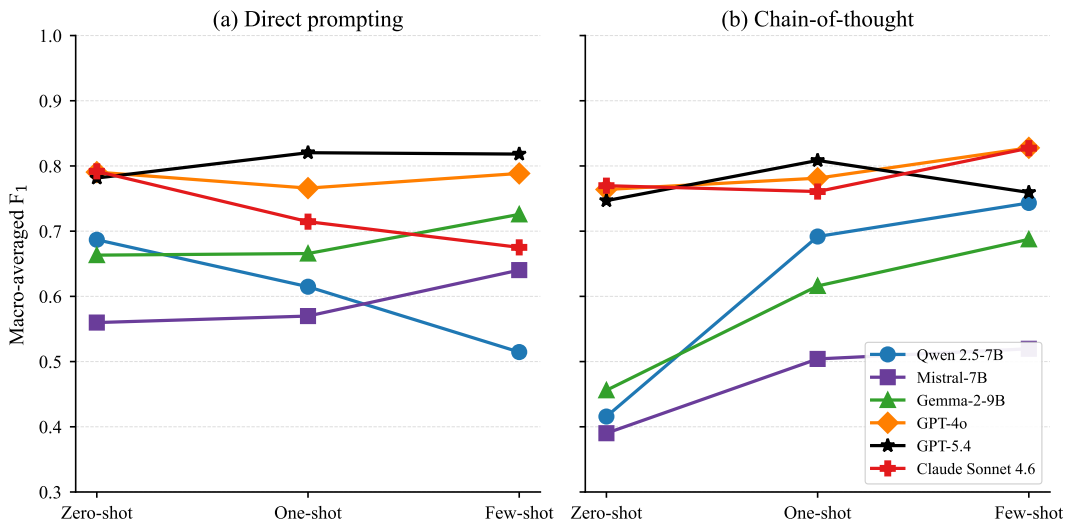
```
{
  "label": "<nominal | warning | hazard>",
  "confidence": <0.0-1.0>,
  "reasoning": "<one sentence stating the key safety factor>"
}
```

than overt symptoms (*runway incursion risk*, *go-around conflict*) yield highly model-dependent accuracy, ranging from ~ 0.0 to 1.0 across the six LLMs. This is consistent with our system-prompt definition where the warning-versus-hazard distinction is explicitly imminence-based and therefore most exposed to model-side disagreement on subtle scenarios.

The confusion-matrix grid in Fig. 9 shows how each LLM’s errors are distributed. Across the closed-source LLMs (GPT-4o, GPT-5.4, Claude), the dominant error mode is *warning-as-hazard* (over-conservative classification), which is operationally desirable in a safety advisor. Open-source LLMs exhibit more scattered confusion, with Mistral in particular under-predicting hazard. Figure 10 reports average inference latency per scenario. Latency is measured as the wall-clock time of one full inference call per scenario, averaged across the 94 test scenarios; for chain-of-thought runs it sums both the reasoning turn and the JSON-extraction turn. The within-model spread across prompting strategies in Fig. 10(b) reflects prompt-length scaling: few-shot prompts contain six exemplars, which lengthens the prefill stage and reduces KV-cache reuse across scenarios. GPT-4o is the fastest model overall (~ 2.07 s/scenario), and Qwen and Mistral are the fastest open-source LLMs (~ 2.4 s/scenario each). Among the open-source models, Gemma-2-9B is the slowest (~ 4.5 s/scenario), reflecting its larger parameter count and the lack of cache reuse imposed by the model’s

**Table 5 Per-class  $F_1$  scores on the three-class CTAF-KHAF benchmark across all models, prompting strategies, and reasoning methods. Best per (model, class) row in bold.**

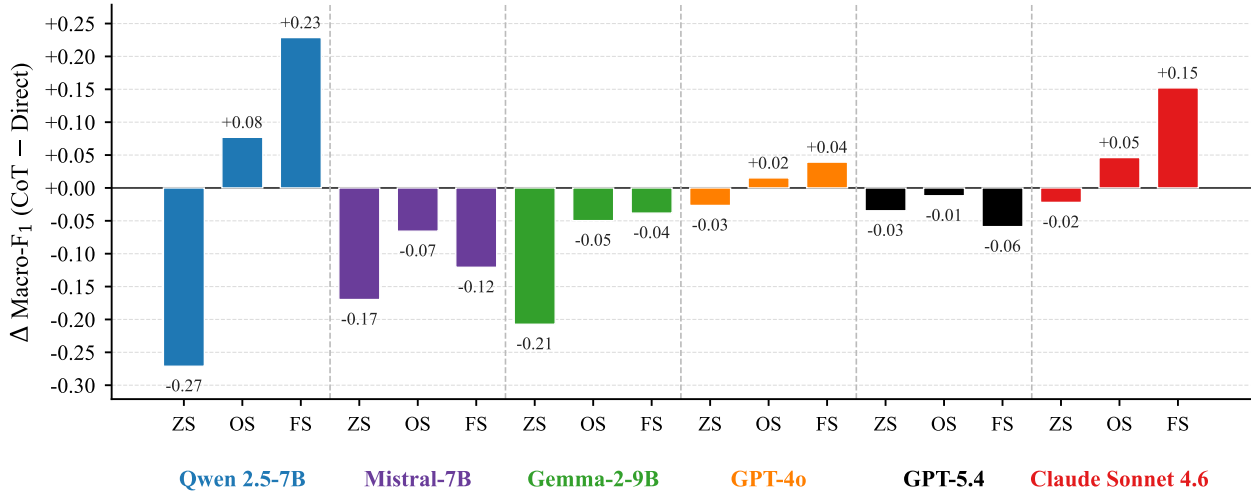
Model	Class	Direct prompting			Chain-of-thought		
		ZS	OS	FS	ZS	OS	FS
<i>Open-source</i>							
Qwen 2.5-7B	Nominal	0.828	0.857	0.756	0.512	0.852	<b>0.968</b>
	Warning	0.545	0.469	0.269	0.569	0.613	<b>0.709</b>
	Hazard	<b>0.688</b>	0.519	0.519	0.167	0.610	0.553
Mistral-7B	Nominal	0.827	0.805	0.861	0.488	0.899	<b>0.921</b>
	Warning	0.400	0.200	0.356	0.561	0.614	<b>0.638</b>
	Hazard	0.453	<b>0.704</b>	<b>0.704</b>	0.121	0.000	0.000
Gemma-2-9B	Nominal	<b>0.939</b>	0.769	0.909	0.591	0.769	0.852
	Warning	0.508	0.475	0.489	<b>0.561</b>	0.431	0.516
	Hazard	0.542	0.753	<b>0.779</b>	0.216	0.648	0.694
<i>Closed-source</i>							
GPT-4o	Nominal	0.899	0.899	0.899	0.968	0.915	<b>0.969</b>
	Warning	0.667	0.618	0.655	0.694	0.687	<b>0.733</b>
	Hazard	0.806	0.781	<b>0.812</b>	0.630	0.742	0.781
GPT-5.4	Nominal	0.836	<b>0.900</b>	0.842	0.792	0.897	0.871
	Warning	0.696	<b>0.730</b>	<b>0.730</b>	0.636	0.712	0.607
	Hazard	0.812	0.831	<b>0.882</b>	0.812	0.817	0.800
Claude Sonnet 4.6	Nominal	0.833	0.825	0.886	0.899	0.862	<b>0.923</b>
	Warning	0.714	0.536	0.400	0.644	0.621	<b>0.733</b>
	Hazard	<b>0.828</b>	0.783	0.740	0.767	0.800	0.825



**Fig. 6 Macro- $F_1$  vs. prompting strategy for the three-class task. Left panel: direct prompting; right panel: chain-of-thought. Closed-source LLMs (GPT-4o, GPT-5.4, Claude) are clustered in the top band and respond modestly to ICL; open-source LLMs show higher variance and CoT-induced regressions.**

chat template. The slowest model overall is Claude Sonnet 4.6 ( $\approx 7.5$  s/scenario), well above every other LLM in our benchmark. The latency vs. macro- $F_1$  scatter in the right panel shows that GPT-4o offers the most favorable cost-quality

Effect of chain-of-thought on  $F_1$



**Fig. 7** Per-condition macro- $F_1$  change from adding chain-of-thought on the three-class task. Qwen benefits at higher ICL counts; closed-source LLMs see small mixed effects; Mistral and Gemma both regress, with Mistral collapsing entirely on the hazard class at the higher-shot CoT settings.

trade-off in our deployment regime, since the higher-quality alternatives (Claude and the few-shot+CoT runs of GPT-5.4) sit further to the right on the latency axis.

### C. Ablation Studies

We run three ablations targeting different stages of the pipeline: ASR quality (Section VIII.D), additive audio noise (Section VIII.E), and direct masking of the transcript text (Section VIII.F). All ablations use the three open-source LLMs (Qwen 2.5-7B, Mistral-7B, Gemma-2-9B) on the same 94-scenario test set, with all main-experiment hyperparameters held fixed. Closed-source LLMs are excluded from ablations to bound API cost. As in Sec. V, only textual inputs (METAR plus CTAF transcript) are passed to the models, so the LLM terminology is appropriate throughout.

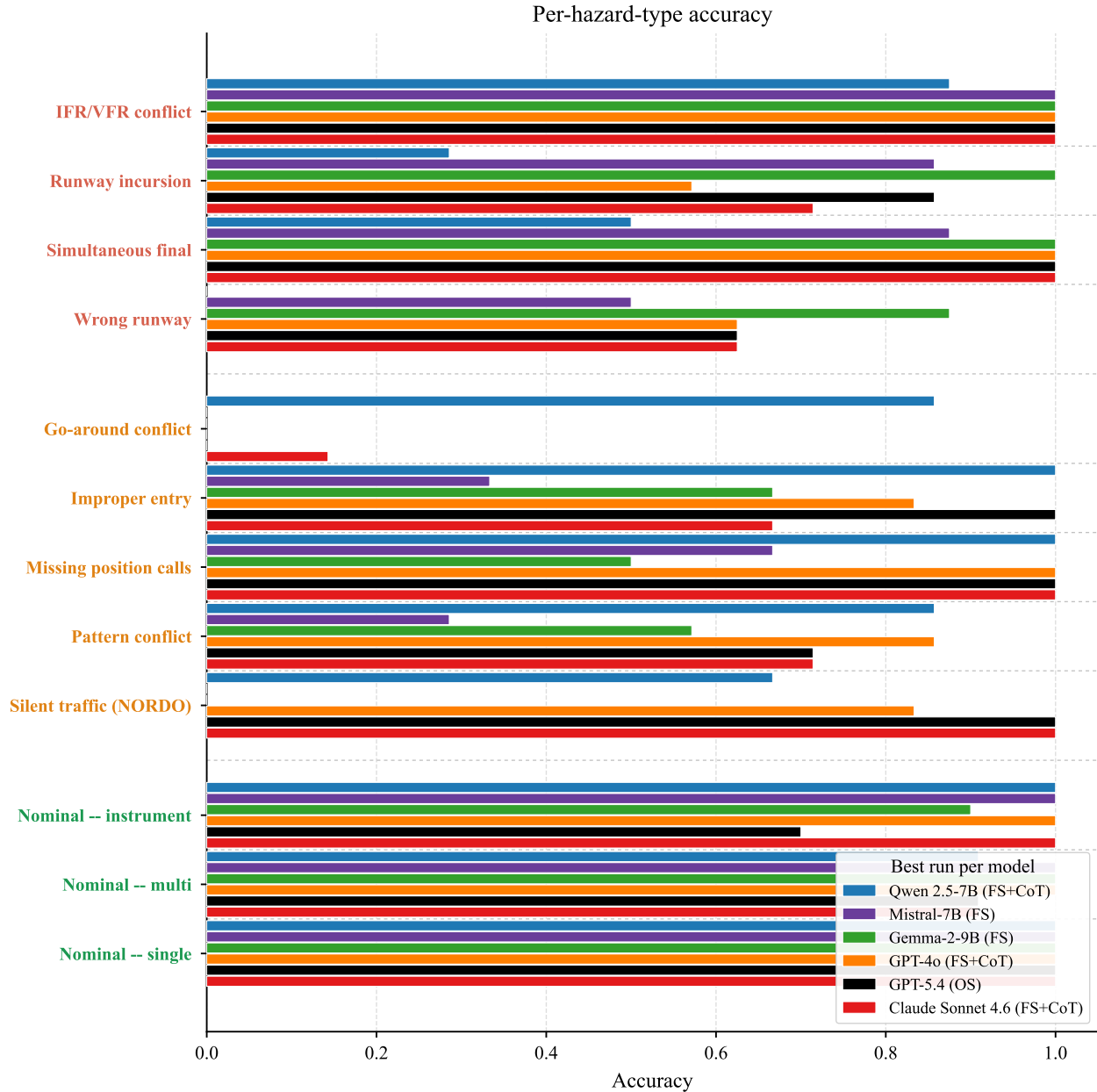
### D. ASR Quality

We re-transcribe each scenario’s clean audio with three Whisper sizes—base (74M parameters), medium (769M), and large-v3 (1.55B; the default in the main experiments)—and re-evaluate each open-source LLM on each transcript under all three prompting strategies. Figure 11 reports macro- $F_1$  as a function of Whisper size for each open-source LLM under zero-shot, one-shot, and few-shot prompting. Under zero-shot prompting, all three LLMs improve as the Whisper model scales from base to medium and large-v3, indicating that better transcription quality consistently benefits downstream hazard classification in the absence of in-context examples. Qwen-2.5-7B and Gemma-2-9B show the clearest positive trends, while Mistral-7B exhibits a smaller but still consistent improvement with increasing ASR size.

Under one-shot prompting, the trends become more model-dependent. Gemma-2-9B and Mistral-7B continue to improve as the Whisper model size increases, suggesting that both models are still able to benefit from incremental ASR gains in the presence of limited prompting examples. In contrast, Qwen-2.5-7B degrades as the ASR model becomes larger, indicating that improved transcription quality does not always translate into improved downstream reasoning once one-shot prompting is introduced.

Under few-shot prompting, Qwen-2.5-7B and Gemma-2-9B again improve with increasing Whisper size, although the gains remain relatively modest. Mistral-7B, however, becomes largely insensitive to ASR scale, showing only minor fluctuations without a consistent trend: performance improves slightly at the medium Whisper size before returning close to its original level with large-v3. This suggests that the model’s downstream behavior is dominated more by prompting dynamics than by transcription fidelity.

Figure 12 further shows that ASR scaling does not affect all safety classes equally. Qwen primarily benefits in



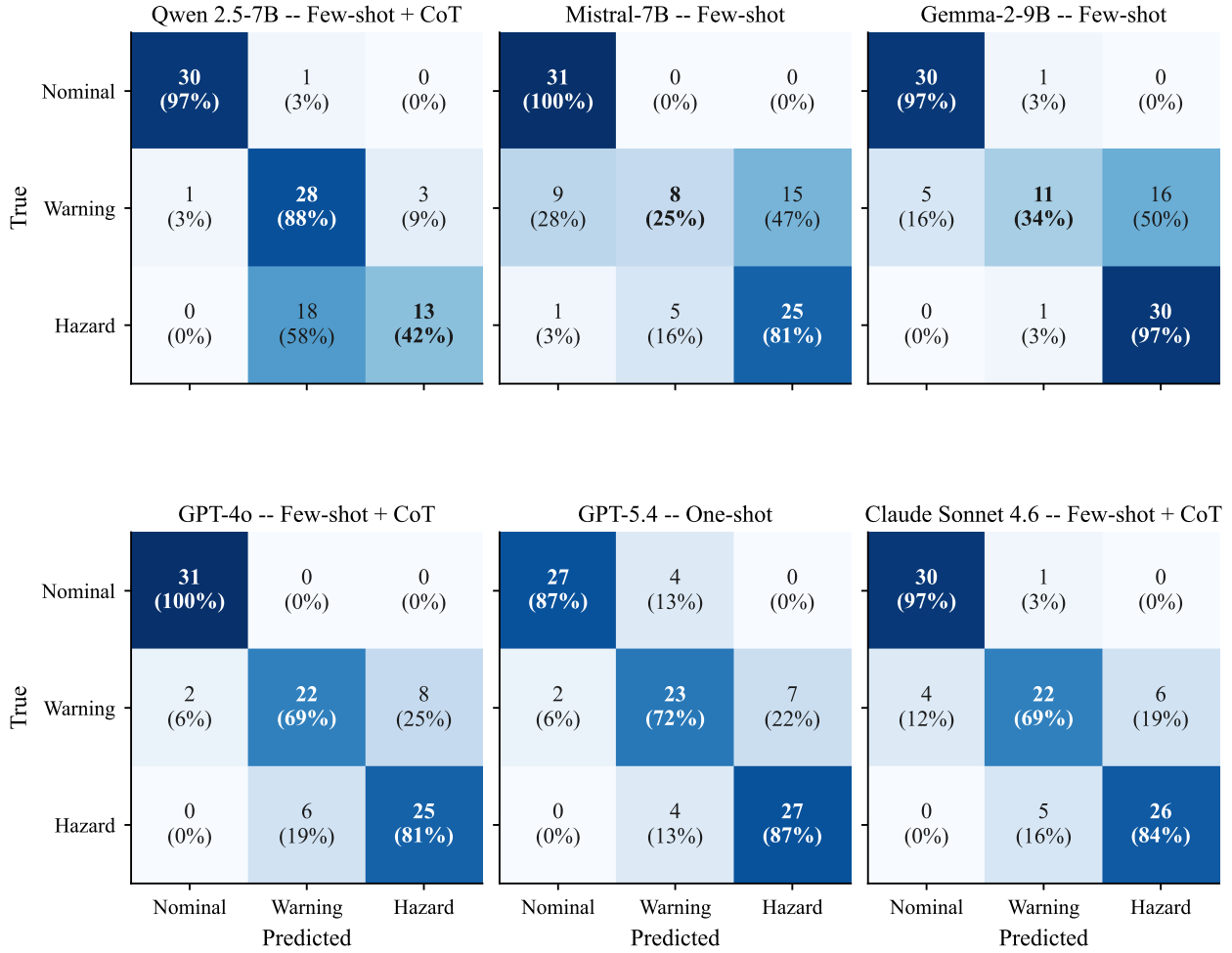
**Fig. 8** Per-hazard-type accuracy on the three-class task, showing the best run per LLM. Hazard categories on the y-axis are colored by their ground-truth safety class.

the nominal class as Whisper size increases, while the warning and hazard classes remain comparatively stable. Mistral exhibits little class-wise variation across all ASR configurations, reinforcing its overall insensitivity to ASR scale. Gemma shows the strongest dependence on ASR size, particularly for the nominal and hazard classes, whereas warning detection remains consistently more difficult.

Overall, the results suggest that larger ASR models provide only limited downstream benefits for CTAF hazard classification. Even the smallest Whisper variant preserves sufficient semantic information for most traffic situations, and the primary performance bottleneck appears to lie in the LLM’s reasoning and class-boundary calibration rather than in speech recognition quality itself.

Figure 12 disaggregates per safety class. Per-class  $F_1$  is also nearly invariant under ASR-size choice, with the warning class showing the most variability—consistent with the warning class being the bottleneck class in the main

Confusion matrices (best run per model)



**Fig. 9 Confusion matrices for the best run of each LLM on the three-class task. Diagonal cells are bolded; row-normalized percentages indicate per-class recall.**

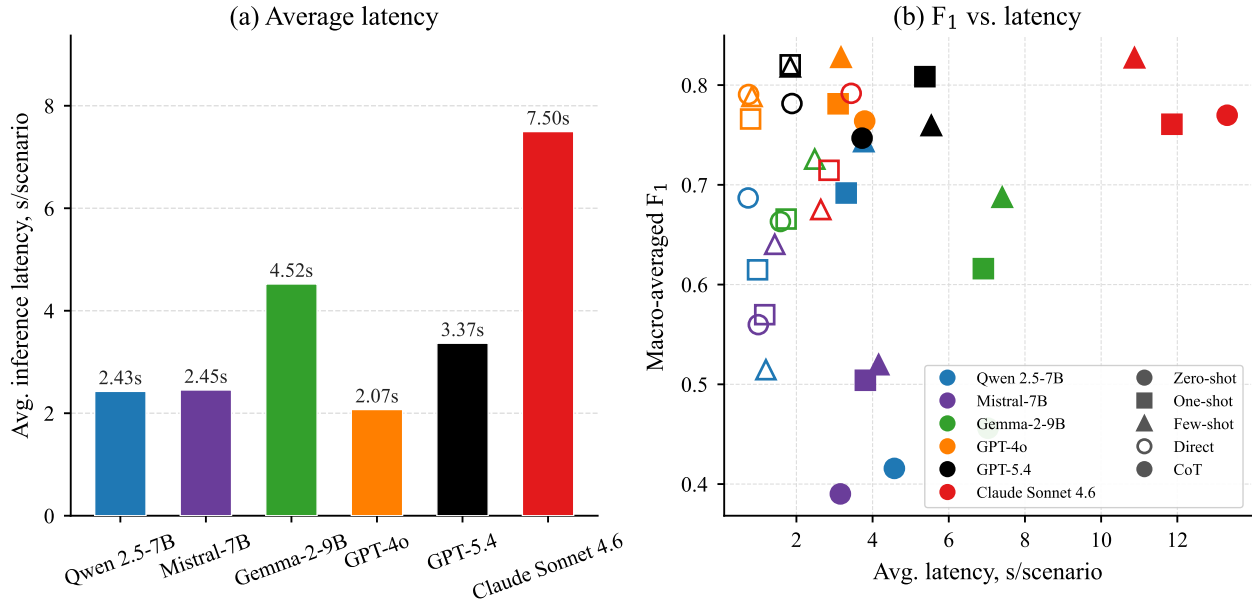
experiments above.

**E. Audio Noise Robustness**

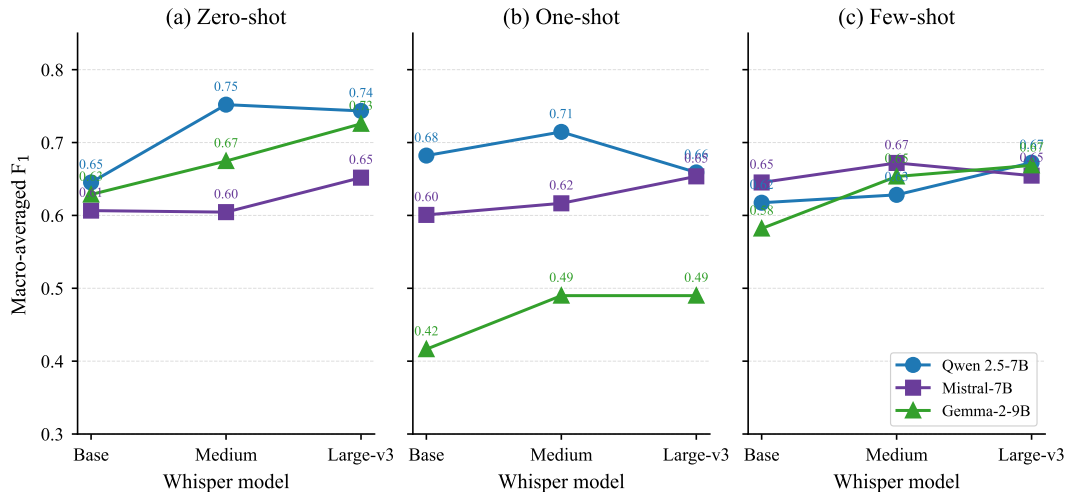
We inject additive white Gaussian noise into each scenario’s clean audio at five noise-to-signal ratios (NSR): 5%, 10%, 25%, 50%, and 75%. The corrupted audio is transcribed with Whisper-Large-v3 (held fixed) and classified under zero-shot prompting only. Figure 13 reports macro- $F_1$  as a function of noise-to-signal ratio (NSR) for each open-source LLM under zero-shot prompting with Whisper-large-v3 transcription. Overall, all three LLMs remain relatively stable as audio noise increases, indicating that the CTAF hazard-classification pipeline is reasonably robust to moderate transcription degradation caused by noisy network communication.

Qwen-2.5-7B exhibits the strongest overall robustness to noise. Its performance remains largely stable across the full NSR range, with only minor fluctuations as noise increases. This suggests that Qwen is able to maintain reliable downstream reasoning even when the ASR input becomes progressively noisier.

Mistral-7B and Gemma-2-9B show greater sensitivity to increasing noise. Both models experience an initial



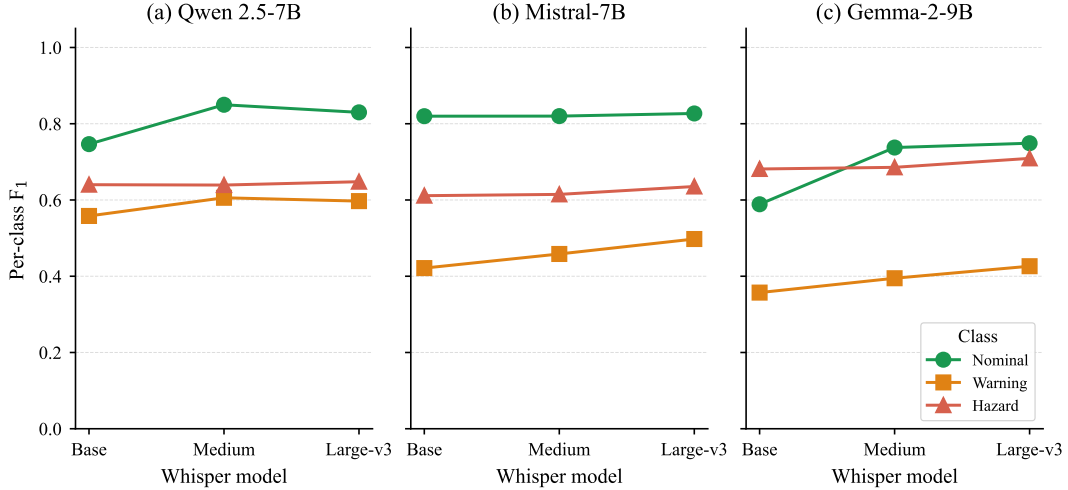
**Fig. 10** Inference latency on the three-class task. Left: average latency per scenario, by LLM. Right: macro- $F_1$  vs. latency scatter, with marker shape indicating prompting strategy and fill indicating CoT vs. direct.



**Fig. 11** Macro- $F_1$  vs. Whisper size for each open-source LLM, one panel per prompting strategy.

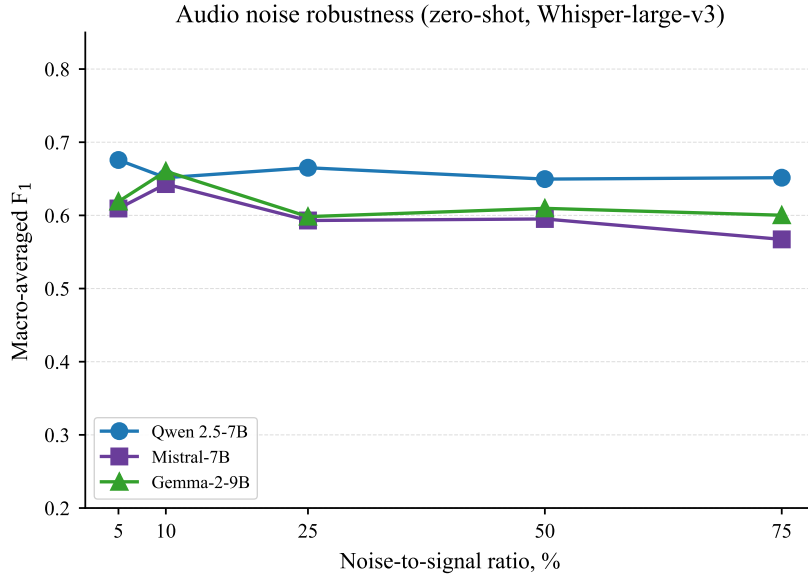
improvement at low NSR levels before degrading as noise increases further. After this early peak, their performance gradually stabilizes at a lower level across the higher-noise settings. This behavior suggests that small amounts of perturbation do not significantly disrupt the semantic structure of the CTAF transcripts, but heavier degradation eventually reduces the LLM’s ability to consistently separate safety classes. Figure 14 further shows that noise affects the three safety classes differently. Across all LLMs, the nominal class remains the most stable under increasing noise, indicating that routine traffic situations are comparatively easy to preserve even when transcription quality deteriorates. In contrast, the warning and hazard classes exhibit greater variation, particularly for Gemma-2-9B and Mistral-7B. Hazard-related performance tends to fluctuate more strongly as NSR increases, suggesting that subtle linguistic cues associated with conflict imminence are more vulnerable to ASR degradation than standard traffic phraseology.

Qwen-2.5-7B again demonstrates the most stable class-wise behavior, with relatively minor changes across all three safety categories. Mistral-7B exhibits moderate instability in the warning and hazard classes as noise increases, while Gemma-2-9B shows the clearest degradation trend under higher NSR conditions, particularly for hazard detection.



**Fig. 12** Per-class  $F_1$  vs. Whisper size, averaged over prompting strategy. Color encodes safety class (green = nominal, orange = warning, red = hazard).

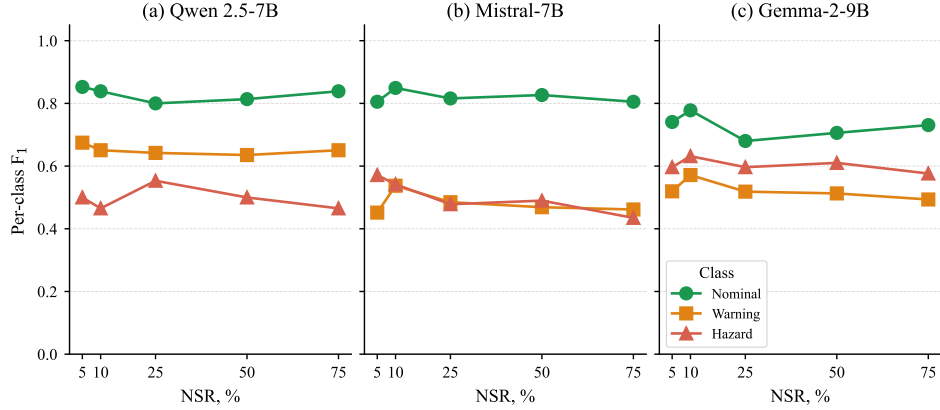
Overall, the results indicate that the open-source LLM pipeline is fairly resilient to noisy CTAF audio, especially for nominal traffic situations. While increased noise can reduce the reliability of warning and hazard detection, the degradation remains gradual rather than catastrophic, suggesting that downstream LLM reasoning retains substantial robustness even under imperfect ASR conditions.



**Fig. 13** Macro- $F_1$  vs. noise-to-signal ratio for each open-source LLM under zero-shot prompting and Whisper-Large-v3 transcription.

### F. Transcript Text Masking

To probe tolerance for partial transcript loss directly—without running the audio chain through Whisper—we mask the ground-truth transcript at five rates ( $r \in \{10, 20, 40, 60, 80\}\%$ ) under two schemes. Random fraction of words replaced with a fixed mask token. LLMs still see the conversational structure but with gappy content. And utterance masking, a random fraction of complete radio calls replaced with a placeholder. LLMs see fewer turns overall. Figure 15 shows macro- $F_1$  versus mask rate for both masking schemes, while Fig. 17 highlights the corresponding model-by-rate

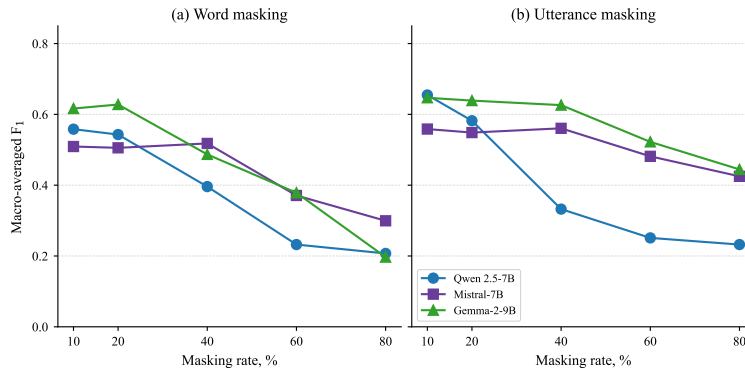


**Fig. 14 Per-class  $F_1$  vs. NSR, by LLM. The warning class (orange) degrades fastest under high noise.**

trends. Under word masking, performance generally degrades gradually as the masking rate increases, indicating that the LLMs can tolerate the loss of individual words up to moderate masking levels before downstream classification quality deteriorates substantially. Qwen-2.5-7B and Gemma-2-9B exhibit sharper degradation at higher masking rates, whereas Mistral-7B remains comparatively more stable and declines more gradually.

Utterance masking produces a different but still broadly monotonic degradation pattern. As the masking rate increases, performance generally decreases for all three LLMs, although the speed and shape of the degradation differ from word masking. Because utterance masking removes complete CTAF transmissions, it disrupts the conversational and temporal structure used for traffic reasoning rather than only corrupting local lexical information. Qwen-2.5-7B is particularly sensitive to this form of masking once moderate amounts of context are removed, suggesting a strong dependence on sequential conversational structure. In contrast, Mistral-7B and Gemma-2-9B remain comparatively more stable under moderate and high utterance masking rates, indicating that both models can still recover useful traffic-state information from incomplete conversation histories. Overall, the results show that utterance masking does not uniformly produce more severe degradation than word masking; instead, the impact depends on both the LLM and the type of contextual information being removed.

Figure 16 further shows that masking affects the three safety classes differently. Across both masking schemes, the nominal class experiences the strongest degradation as the masking rate increases, particularly for Qwen-2.5-7B and Gemma-2-9B. In contrast, the warning class remains comparatively stable across most masking levels and models. Hazard detection exhibits greater variability: Qwen-2.5-7B and Gemma-2-9B show substantial hazard degradation at high masking rates, while Mistral-7B maintains more stable hazard performance under utterance masking before eventually declining at the highest masking levels. These results suggest that masking primarily disrupts the models' ability to recover the global conversational state associated with nominal traffic flow, while hazard-related cues remain partially recoverable even under incomplete transcript conditions.



**Fig. 15 Macro- $F_1$  vs. masking rate. Left: word-level masking; right: utterance-level masking. Utterance masking is substantially more destructive at every rate.**

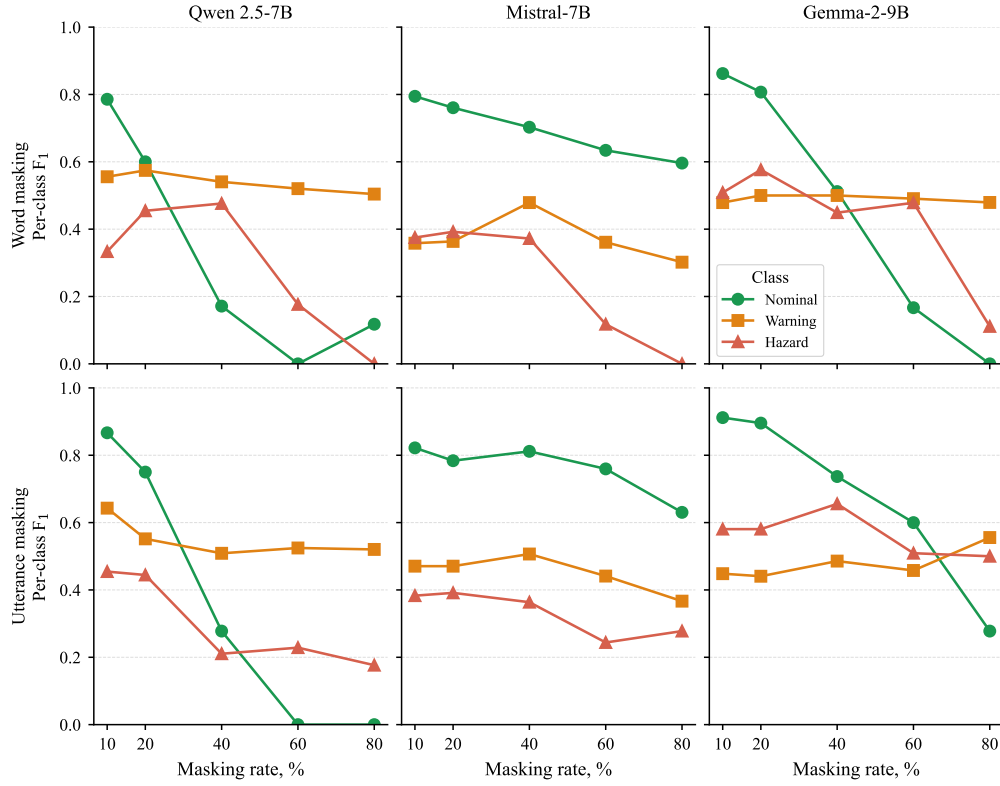


Fig. 16 Per-class  $F_1$  vs. masking rate. Top row: word masking; bottom row: utterance masking. Each column is one open-source LLM.

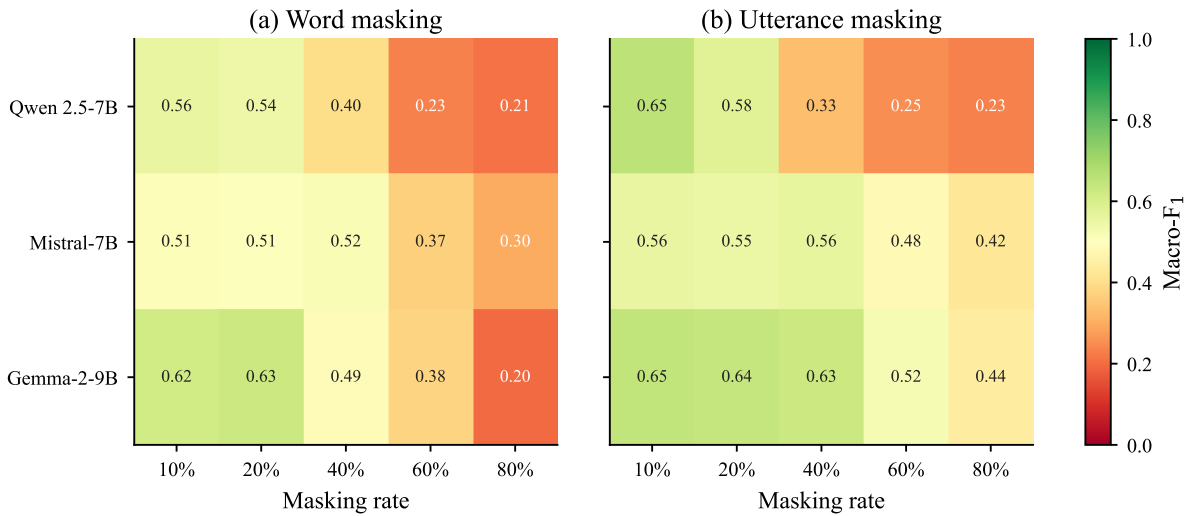


Fig. 17 Macro- $F_1$  heatmap across (LLM, mask rate) for word (left) and utterance (right) masking. Numerical cell values are macro- $F_1$ .