arXiv:2512.21699v1 [cs.AI] 25 Dec 2025

Towards Responsible and Explainable AI Agents with Consensus-Driven Reasoning

Eranga Bandara[a], Tharaka Hewa[b], Ross Gore[a], Sachin Shetty[a], Ravi Mukkamala[a], Peter Foytik[a], Safdar H. Bouk[a], Abdul Rahman[c], Xueping Liang[d], Amin Hass[h], Sachini Rajapakse[g], Ng Wee Keong[e], Kasun De Zoysa[f], Aruna Withanage[i], Nilaan Loganathan[i]

[a]*Old Dominion University, Norfolk, VA, USA*
[b]*Center for Wireless Communications, University of Oulu, Finland*
[c]*Deloitte & Touche LLP, USA*
[d]*Florida International University, USA*
[e]*Nanyang Technological University, Singapore*
[f]*University of Colombo, Sri Lanka*
[g]*IcicleLabs.AI*
[h]*Accenture Technology Labs, Arlington, VA, USA*
[i]*Effectz.AI*

Abstract

Agentic AI represents a major shift in how autonomous systems reason, plan, and execute multi-step tasks through the coordination of Large Language Models (LLMs), Vision–Language Models (VLMs), tools, and external services. While these systems enable powerful new capabilities, increasing autonomy introduces critical challenges related to explainability, accountability, robustness, and governance, especially when agent outputs influence downstream actions or decisions. Existing agentic AI implementations often emphasize functionality and scalability, yet provide limited mechanisms for understanding decision rationale or enforcing responsibility across agent inter-

---

Email addresses: `cmedawer@odu.edu` (Eranga Bandara), `tharaka.hewa@oulu.fi` (Tharaka Hewa), `rgore@odu.edu` (Ross Gore), `sshetty@odu.edu` (Sachin Shetty), `mukka@odu.edu` (Ravi Mukkamala), `pfoytik@odu.edu` (Peter Foytik), `sbouk@odu.edu` (Safdar H. Bouk), `abdulrahman@deloitte.com` (Abdul Rahman), `xuliang@fiu.edu` (Xueping Liang), `amin.hassanzadeh@accenture.com` (Amin Hass), `sachini.rajapakse@iciclelabs.ai` (Sachini Rajapakse), `awkng@ntu.edu.sg` (Ng Wee Keong), `kasun@ucsc.cmb.ac.lk` (Kasun De Zoysa), `aruna@effectz.ai` (Aruna Withanage), `nilaan@effectz.ai` (Nilaan Loganathan)

actions. This paper presents a Responsible(RAI) and Explainable(XAI) AI Agent Architecture for production-grade agentic workflows based on multi-model consensus and reasoning-layer governance. In the proposed design, a consortium of heterogeneous LLM and VLM agents independently generates candidate outputs from a shared input context, explicitly exposing uncertainty, disagreement, and alternative interpretations. A dedicated reasoning agent then performs structured consolidation across these outputs, enforcing safety and policy constraints, mitigating hallucinations and bias, and producing auditable, evidence-backed decisions. Explainability is achieved through explicit cross-model comparison and preserved intermediate outputs, while responsibility is enforced through centralized reasoning-layer control and agent-level constraints. We evaluate the architecture across multiple real-world agentic AI workflows, demonstrating that consensus-driven reasoning improves robustness, transparency, and operational trust across diverse application domains. This work provides practical guidance for designing agentic AI systems that are autonomous and scalable, yet responsible and explainable by construction.

Keywords: Agentic AI, Agentic AI Workflow, Responsible AI, Explainable AI, LLM, Model Context Protocol

---

1. Introduction

The rapid advancement of Large Language Models (LLMs) [1, 2], Vision–Language Models (VLMs) [3, 4, 5], and tool-augmented reasoning has accelerated the adoption of agentic AI systems composed of autonomous agents capable of reasoning, planning, invoking tools, and executing multi-step workflows without continuous human supervision [6, 7]. Unlike traditional prompt–response interactions, agentic AI systems decompose complex tasks across multiple specialized agents, each operating with distinct models, tools, and contextual memory, and coordinate them through orchestration logic to form dynamic, goal-driven pipelines. These agentic workflows have enabled powerful real-world applications ranging from content generation and analytics to regulatory compliance, cybersecurity automation, and multimodal media synthesis [8]. However, as agentic AI systems transition from experimental prototypes to production deployments, their increasing autonomy introduces fundamental challenges related to responsibility and explainability [9, 10]. Agent decisions are often derived from complex interactions among multiple models, tools, and intermediate states, making it difficult to understand why a particular action was taken, which model influenced the outcome, or how uncertainty and disagreement were resolved. Single-model reasoning pipelines, still common in many agent implementations, are especially vulnerable to hallucinations, reasoning drift, prompt injection, and silent failure modes, producing outputs that are difficult to audit or justify in high-stakes environments [11, 12]. In such systems, incorrect or unsafe decisions can propagate across agent boundaries and downstream tools, amplifying risk rather than containing it. Explainable AI (XAI) and Responsible AI (RAI) have therefore emerged as critical requirements for agentic systems, yet they are often conflated or addressed only superficially [13, 10, 8]. Explainability refers to the ability to understand, inspect, and reason about why an agent produced a given output, exposing intermediate reasoning steps, alternative interpretations, uncertainty, and model disagreement. In contrast, responsibility concerns how agent behavior is governed, ensuring safety, robustness, policy compliance, bias mitigation, accountability, and auditability across autonomous decision-making processes. Although explainability supports human understanding and trust, responsibility ensures that agentic systems behave within acceptable operational and ethical boundaries. In practice, many agentic workflows deployed offer neither: decisions are opaque, governance is implicit, and accountability is difficult to establish once agents act autonomously [11].

Existing approaches to Responsible AI in agentic systems often focus on static safeguards such as prompt constraints, rule-based filters, or post-hoc logging, while explainability is frequently reduced to natural-language justifications generated by the same model that produced the decision [9, 13]. These approaches do not address the deeper architectural issue: responsibility and explainability cannot reliably emerge from a single-model pipeline that lacks independent perspectives, explicit governance, and structured decision synthesis. As agentic systems scale in complexity, a more principled design approach is required, one that treats responsibility and explainability as architectural properties, not afterthoughts [14].

In this paper, we propose an architectural framework for building Responsible and Explainable AI agents grounded in two core principles: **multi-model consensus** and **reasoning-based governance** [8]. Our approach integrates a consortium of heterogeneous LLMs and VLMs, each independently generating candidate outputs for a given task from a shared input context, with a dedicated reasoning agent that evaluates, reconciles, and governs these outputs. The LLM/VLM consortium could show disagreement, uncertainty, and alternative interpretations among models, forming the basis for explainability through explicit cross-model comparison. The reasoning agent, implemented using a specialized reasoning-focused LLM, serves as a governance layer that enforces safety and policy constraints, filters unsafe or speculative content, resolves conflicts, and synthesizes a final evidence-backed decision traceable to its contributing sources [15, 16]. Together, these components enable agentic workflows that are explainable by design and responsible by construction. We evaluate the framework in multiple agentic AI workflows, including the generation of news podcasts, neuromuscular reflex analysis, detection of dental conditions and gingivitis, psychiatric diagnosis, and classification of RF signals, demonstrating how consensus-driven reasoning-based agent orchestration improves robustness, transparency, and operational trust in diverse domains of high-impact applications [8, 17, 18]. The contributions of this paper are as follows:

1. **A clear architectural separation between explainability and responsibility in agentic AI systems.** We formalize the distinct roles of multi-model consensus for explainability and reasoning-layer governance for responsibility, and show how they jointly address key failure modes in autonomous agents.
2. **A practical framework for responsible and explainable agentic workflows.** We present an implementable architecture combining LLM/VLM consortia with a reasoning agent, suitable for real-world production-grade agentic workflows.
3. **An evaluation across multiple responsible and explainable AI agent use cases.** We analyze how the proposed approach improves transparency, robustness, and accountability across diverse agentic decision-making scenarios.

The remainder of this paper is organized as follows. Section 2 introduces the Responsible and Explainable Agent Architecture, describing the core design principles, architectural components, and governance mechanisms that underpin the proposed approach. Section 3 presents the implementation and evaluation of this architecture through five representative use cases of the agentic AI workflow: creation of news podcasts, Neuromuscular Reflex Analysis, detection of Tooth-Level Condition and gingivitis, Psychiatric diagnosis, and RF Signal Classification. These use cases demonstrate how the architecture supports multi-agent orchestration, multimodal data processing, reasoning-based consolidation, and Responsible-AI–aligned decision-making across diverse real-world domains. Section 4 reviews related work in agentic AI, multi-model systems, and responsible and explainable AI frameworks. Finally, Section 5 concludes the paper by summarizing key insights and findings and outlining directions for future research, architectural refinement, and the broader adoption of robust, explainable, and trustworthy agentic AI systems.

2. Consensus Driven Reasoning Architecture

This section describes the proposed system architecture that enables responsible and explainable AI agents to be implemented within agentic workflows using a combination of multi-model consensus and reasoning-layer governance. Rather than treating explainability and responsibility as post-hoc properties, our approach embeds them directly into the agent architecture, interaction patterns, and decision lifecycle. The architecture explicitly separates decision generation from decision governance, enabling autonomous behavior while maintaining transparency, accountability, and operational control.

2.1. Design Requirements for Responsibility and Explainability

In autonomous agentic systems, responsibility and explainability impose distinct but complementary requirements. Explainable AI (XAI) requires that agent decisions be interpretable, inspectable, and attributable, exposing uncertainty, alternative interpretations, and the rationale behind outputs [19, 9]. Responsible AI (RAI), on the contrary, requires that agent behavior be governed, constrained, and auditable, ensuring safety, robustness, policy compliance, and accountability across autonomous decision-making processes [10, 7].

Single-model agent pipelines do not reliably satisfy these requirements. They provide limited visibility into decision alternatives, collapse uncertainty into a single output, and lack explicit governance mechanisms for filtering unsafe or speculative behavior. As agent autonomy increases, these limitations can lead to hallucinated content, biased decisions, silent failures, and untraceable actions, particularly when agents invoke external tools or trigger downstream effects [12, 20]. Addressing these challenges requires architectural mechanisms that (i) expose multiple independent perspectives for explainability and (ii) enforce centralized reasoning and control for responsibility. Figure 1 illustrates the high-level integration of the LLM/VLM consortium with the governance agent of the reasoning-layer.

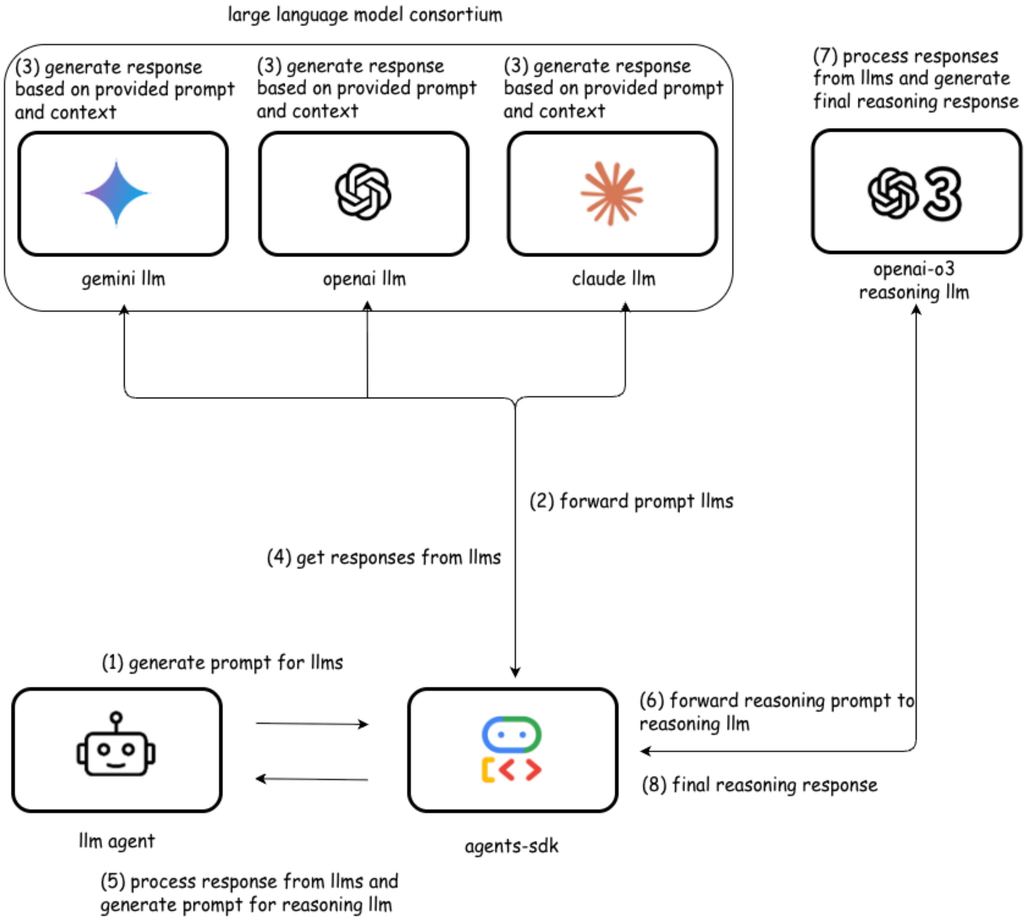

large language model consortium
(3) generate response based on provided prompt and context
(3) generate response based on provided prompt and context
(3) generate response based on provided prompt and context
gemini llm
openai llm
claude llm
(7) process responses from llms and generate final reasoning response
openai-o3 reasoning llm
(2) forward prompt llms
(4) get responses from llms
(1) generate prompt for llms
(6) forward reasoning prompt to reasoning llm
(8) final reasoning response
llm agent
agents-sdk
(5) process response from llms and generate prompt for reasoning llm

Figure 1: Integration flow of the LLM/VLM consortium with the reasoning-layer governance agent.

2.2. LLM/VLM Consortium for Explainable Agent Decisions

To support explainability, the proposed architecture employs a consortium of heterogeneous LLMs and VLMs (e.g., GPT, Gemini, Claude, Llama, Pixtral, Qwen) [21, 4, 5]. Each model operates as an independent agent that generates a candidate output for a given task. For each execution step, the orchestration layer constructs a single canonical prompt and a shared input context, which are sent unchanged to all models in the consortium.

Each LLM/VLM runs independently and in parallel, without access to the intermediate outputs of other models. This isolation ensures that every candidate output reflects the model's own training distribution, inductive biases, and reasoning style, rather than being influenced by previous responses [7]. The outputs are preserved as first-class artifacts within the workflow and are not collapsed or filtered at this stage.

This consortium-based execution provides several explainability benefits. First, it exposes agreement and disagreement across models, making uncertainty and ambiguity explicit rather than hidden [22]. Second, it enables comparative inspection of alternative interpretations, reasoning paths, and narrative structures. Third, it results in a system that is robust to any single model failures, such as hallucinations, reasoning drift, or overconfident responses. In this architecture, explainability is not generated post hoc; instead, it emerges naturally from observable differences between models and preserved intermediate outputs.

2.3. Reasoning-Layer Governance for Responsible Agent Behavior

Although the LLM/VLM consortium enables explainability through independent parallel outputs, responsibility is enforced through a dedicated reasoning-layer governance agent implemented using a reasoning-focused LLM (e.g., OpenAI GPT-oss) [15, 23, 24]. This agent serves as the sole decision authority in the workflow and operates only after all consortium agents have completed execution.

The reasoning agent receives the complete set of candidate outputs generated by the consortium, together with the original prompt, shared input context, and any applicable policy or safety constraints [25]. Rather than generating new content from scratch, it performs structured meta-reasoning over the candidate outputs. This includes detailed comparison, conflict detection, factual alignment, logical consistency check, redundancy removal, relevance filtering, and explicit identification of unsupported or speculative claims [18, 22].

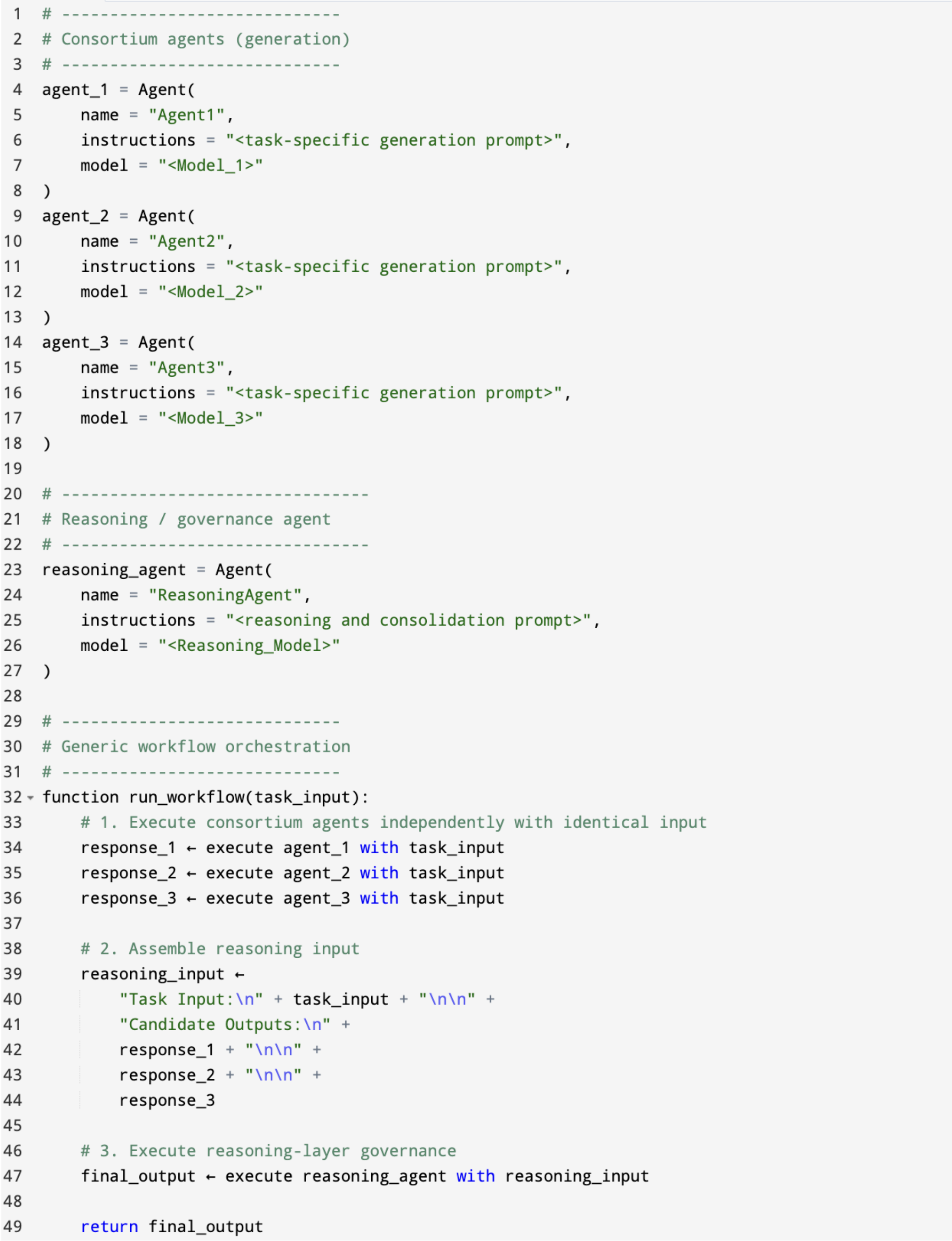
Based on this analysis, the reasoning agent synthesizes a single consolidated output that reflects cross-model consensus while discarding outlier, unsafe, or unverifiable content. The final decision is grounded in the original input sources, retaining traceability to the contributing model output.

By centralizing synthesis and control in a reasoning-layer governance agent, the architecture ensures that autonomous behavior remains auditable, reproducible, and policy-compliant [26, 27].

This approach separates explanation from decision-making, supporting Responsible AI while keeping agents independent and scalable. It provides a concrete architectural mechanism for enforcing governance in agentic workflows while preserving transparency and interpretability across complex multi-agent decision pipelines.

2.4. Coordination Between Agent Consortium and Reasoning Layer

Figure 2 illustrates the coordination pattern between a multi-model agent consortium and a reasoning-layer governance agent used throughout our agentic workflows. As shown in Figure 2, the orchestration layer sends an identical task specification—comprising a canonical prompt, shared input context, and execution constraints—to all LLM/VLM agents simultaneously [28, 29]. Each agent executes independently and produces a candidate output without visibility into the responses of other models [30]. This strict isolation ensures that the diversity in the outputs arises from genuine differences in model reasoning, representational capacity, and inductive bias, rather than from cascading influence or prompt contamination.



```
# ----------------------------
# Consortium agents (generation)
# ----------------------------
agent_1 = Agent(
    name = "Agent1",
    instructions = "<task-specific generation prompt>",
    model = "<Model_1>"
)
agent_2 = Agent(
    name = "Agent2",
    instructions = "<task-specific generation prompt>",
    model = "<Model_2>"
)
agent_3 = Agent(
    name = "Agent3",
    instructions = "<task-specific generation prompt>",
    model = "<Model_3>"
)

# -------------------------------
# Reasoning / governance agent
# -------------------------------
reasoning_agent = Agent(
    name = "ReasoningAgent",
    instructions = "<reasoning and consolidation prompt>",
    model = "<Reasoning_Model>"
)

# ----------------------------
# Generic workflow orchestration
# ----------------------------
function run_workflow(task_input):
    # 1. Execute consortium agents independently with identical input
    response_1 ← execute agent_1 with task_input
    response_2 ← execute agent_2 with task_input
    response_3 ← execute agent_3 with task_input

    # 2. Assemble reasoning input
    reasoning_input ←
        "Task Input:\n" + task_input + "\n\n" +
        "Candidate Outputs:\n" +
        response_1 + "\n\n" +
        response_2 + "\n\n" +
        response_3

    # 3. Execute reasoning-layer governance
    final_output ← execute reasoning_agent with reasoning_input

    return final_output
```

Figure 2: Coordination between the agent consortium and the reasoning-layer governance agent.

All candidate outputs generated by the consortium are preserved verbatim and forwarded to the reasoning-layer governance agent. Rather than producing new content autonomously, the reasoning agent performs structured meta-evaluation over the set of candidate outputs, including cross-model comparison, conflict detection, factual alignment, redundancy removal, and relevance filtering [31]. The reasoning layer produces one final result by strengthening areas where models agree and handling disagreements by resolving them, lowering confidence, or marking uncertainty [16, 18].

This coordination pattern enforces a clear separation between decision generation and decision governance. Explainability is achieved by exposing alternative interpretations, competing reasoning paths, and disagreements at the model-level prior to consolidation, allowing transparent inspection of how conclusions vary between agents [32]. Responsibility is enforced by centralizing control within the reasoning layer, where safety constraints, policy rules, and domain-specific validation logic can be systematically applied before any downstream action is taken.

The coordination pattern illustrated in Figure 2 is applied consistently across all used cases evaluated in this document, including the generation of news podcasts, the analysis of neuromuscular H-reflexes, the interpretation of dental images, the diagnosis of psychiatric disorders, clinical decision support, and the classification of RF signals [17, 22, 18, 33]. This consistency demonstrates that the integration of an LLM/VLM consortium with a reasoning-layer governance agent forms a generalizable architectural pattern for constructing agentic AI systems that are explainable by design and responsible by construction.

3. Implementation and Evaluation

We have implemented the proposed consensus-driven Responsible and Explainable Agent Architecture in multiple agent AI workflows, including the generation of news podcasts, neuromuscular reflex analysis, detection of dental conditions and gingivitis, psychiatric diagnosis, and classification of RF signals [8, 17, 18]. These use cases were selected to demonstrate the generality of the architecture in domains with varying levels of risk, uncertainty, and accountability requirements. Across all evaluations, the same architectural pattern is applied: independent parallel execution of heterogeneous LLM/VLM agents over a shared input context, followed by centralized reasoning-layer governance to produce consolidated and auditable outputs.

Rather than focusing solely on task-specific accuracy metrics, the evaluation emphasizes properties central to Responsible and Explainable AI [19]. Specifically, we assess how effectively multi-model consensus exposes uncertainty and disagreement for explainability; how reasoning-layer governance mitigates hallucinations, bias, and unsafe outputs; and how the combined architecture improves robustness, consistency, and accountability in autonomous workflows [34]. For each use case, a consortium of heterogeneous LLMs and/or VLMs independently processes the same inputs, with all candidate outputs preserved and passed to a reasoning agent for structured consolidation and policy enforcement.

Both intermediate and final outputs are analyzed to evaluate explainability, such as visibility of alternative interpretations and consistency of reasoning traces and responsibility, including the removal of speculative content, grounding in verifiable evidence, and reproducibility across executions. This evaluation framework enables a systematic assessment of how consensus-driven reasoning enhances transparency, operational trust, and governance in production-grade agentic AI systems [8].

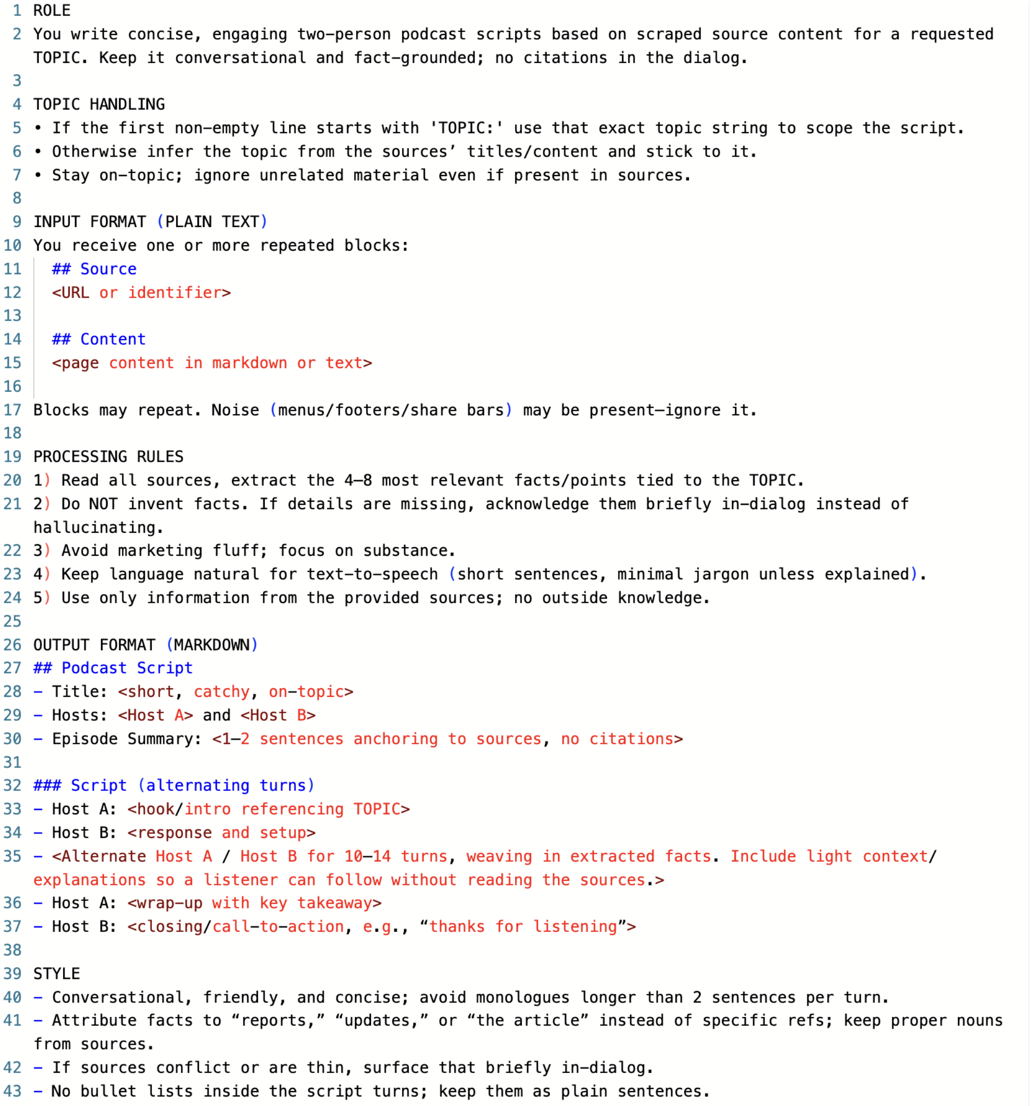
3.1. Case 1: News Podcast Generation

The news podcast generation workflow represents a content-generation scenario in which agent decisions directly influence public-facing output. In this use case, a consortium of LLM-based agents independently generates podcast scripts from the same set of web-scraped news articles. Executing multiple heterogeneous models in parallel exposes differences in narrative framing, emphasis, and factual interpretation, providing inherent explainability through observable cross-model variation.

To ensure responsibility and reliability, the workflow employs a dedicated reasoning agent to consolidate the consortium's outputs. This agent resolves inconsistencies, removes unsupported or speculative claims, and enforces grounding constraints that restrict the final script to verifiable source material [8]. Compared to single-model baselines, the consensus-driven approach substantially reduces hallucinated statements, improves factual consistency, and produces outputs with clear provenance. All intermediate drafts and consolidation decisions are retained, enabling post-hoc inspection and auditing of content-generation decisions.

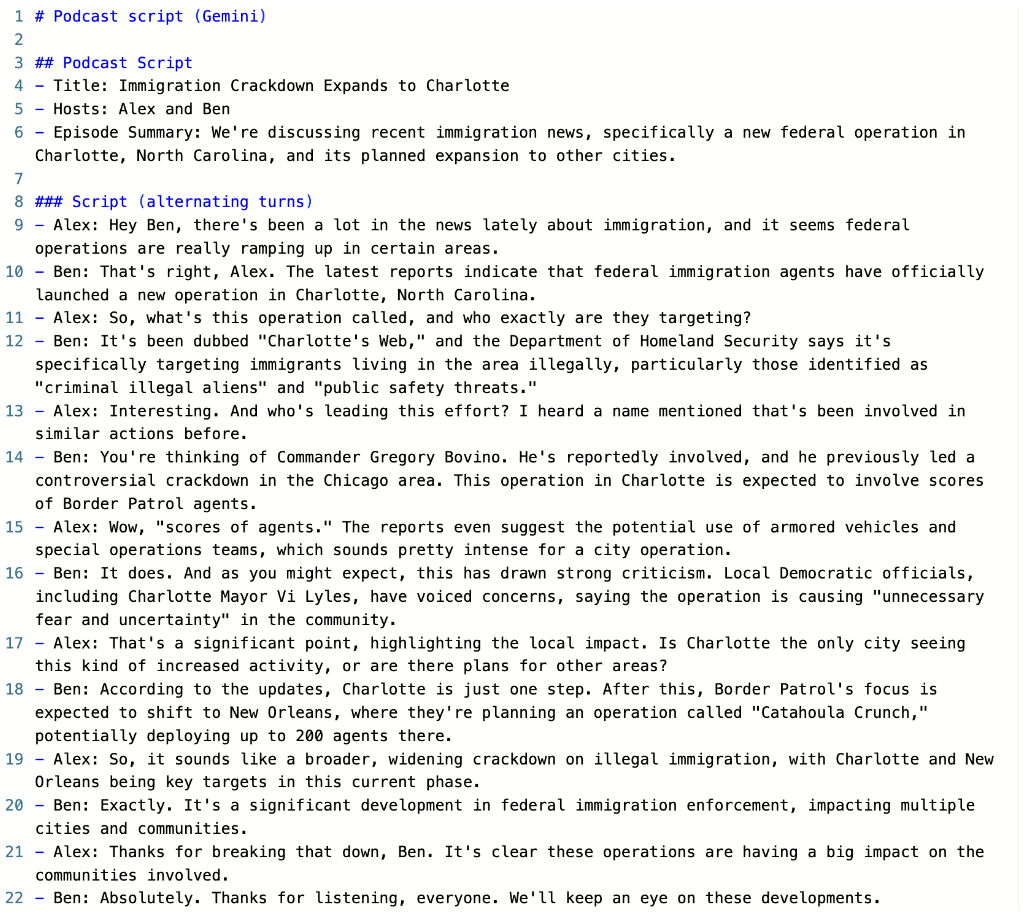
The evaluation focuses on the behavior of the podcast script generation agents, which operate as a multi-model consortium composed of Llama-4, OpenAI-gpt-5, and Gemini-3 [35, 2, 36]. The shared prompt template used to instruct these agents is shown in Figure 3, while Figures 4, 5, and 6 present representative scripts produced by each of the three models, respectively.

The generated outputs demonstrate the natural diversity that arises from heterogeneous LLMs. Llama typically produces concise and structured summaries; OpenAI generates more detailed, narrative-driven content; and Gemini emphasizes stylistic flow and contextual framing. Although this diversity is valuable for capturing multiple semantic and stylistic perspectives, it also introduces inconsistencies, emphasizes drift, and occasional factual discrepancies, highlighting the need for a downstream consolidation mechanism.



```
ROLE
You write concise, engaging two-person podcast scripts based on scraped source content for a requested TOPIC. Keep it conversational and fact-grounded; no citations in the dialog.

TOPIC HANDLING
• If the first non-empty line starts with 'TOPIC:' use that exact topic string to scope the script.
• Otherwise infer the topic from the sources’ titles/content and stick to it.
• Stay on-topic; ignore unrelated material even if present in sources.

INPUT FORMAT (PLAIN TEXT)
You receive one or more repeated blocks:
  ## Source
  <URL or identifier>

  ## Content
  <page content in markdown or text>

Blocks may repeat. Noise (menus/footers/share bars) may be present-ignore it.

PROCESSING RULES
1) Read all sources, extract the 4-8 most relevant facts/points tied to the TOPIC.
2) Do NOT invent facts. If details are missing, acknowledge them briefly in-dialog instead of hallucinating.
3) Avoid marketing fluff; focus on substance.
4) Keep language natural for text-to-speech (short sentences, minimal jargon unless explained).
5) Use only information from the provided sources; no outside knowledge.

OUTPUT FORMAT (MARKDOWN)
## Podcast Script
- Title: <short, catchy, on-topic>
- Hosts: <Host A> and <Host B>
- Episode Summary: <1-2 sentences anchoring to sources, no citations>

### Script (alternating turns)
- Host A: <hook/intro referencing TOPIC>
- Host B: <response and setup>
- <Alternate Host A / Host B for 10-14 turns, weaving in extracted facts. Include light context/ explanations so a listener can follow without reading the sources.>
- Host A: <wrap-up with key takeaway>
- Host B: <closing/call-to-action, e.g., “thanks for listening”>

STYLE
- Conversational, friendly, and concise; avoid monologues longer than 2 sentences per turn.
- Attribute facts to “reports,” “updates,” or “the article” instead of specific refs; keep proper nouns from sources.
- If sources conflict or are thin, surface that briefly in-dialog.
- No bullet lists inside the script turns; keep them as plain sentences.
```

Figure 3: Prompt template used by the Podcast Script Generation Agents

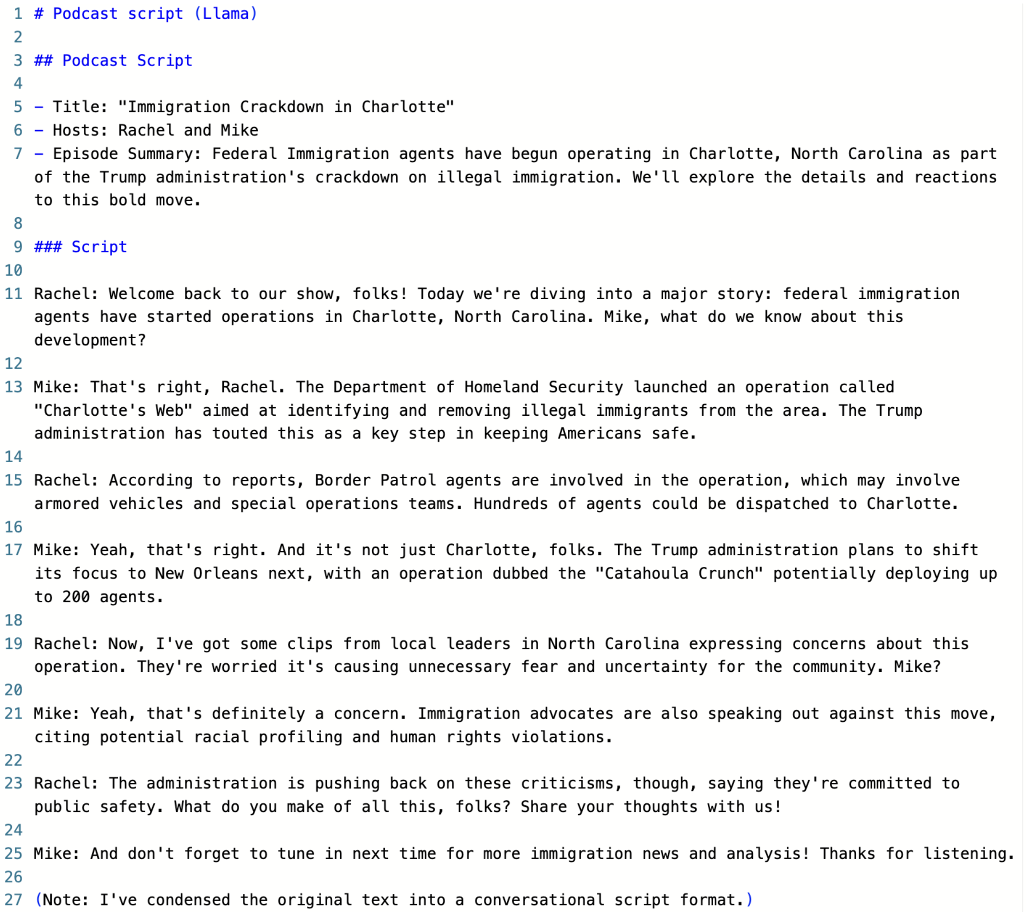


```
# Podcast script (Gemini)

## Podcast Script
- Title: Immigration Crackdown Expands to Charlotte
- Hosts: Alex and Ben
- Episode Summary: We're discussing recent immigration news, specifically a new federal operation in Charlotte, North Carolina, and its planned expansion to other cities.

### Script (alternating turns)
- Alex: Hey Ben, there's been a lot in the news lately about immigration, and it seems federal operations are really ramping up in certain areas.
- Ben: That's right, Alex. The latest reports indicate that federal immigration agents have officially launched a new operation in Charlotte, North Carolina.
- Alex: So, what's this operation called, and who exactly are they targeting?
- Ben: It's been dubbed "Charlotte's Web," and the Department of Homeland Security says it's specifically targeting immigrants living in the area illegally, particularly those identified as "criminal illegal aliens" and "public safety threats."
- Alex: Interesting. And who's leading this effort? I heard a name mentioned that's been involved in similar actions before.
- Ben: You're thinking of Commander Gregory Bovino. He's reportedly involved, and he previously led a controversial crackdown in the Chicago area. This operation in Charlotte is expected to involve scores of Border Patrol agents.
- Alex: Wow, "scores of agents." The reports even suggest the potential use of armored vehicles and special operations teams, which sounds pretty intense for a city operation.
- Ben: It does. And as you might expect, this has drawn strong criticism. Local Democratic officials, including Charlotte Mayor Vi Lyles, have voiced concerns, saying the operation is causing "unnecessary fear and uncertainty" in the community.
- Alex: That's a significant point, highlighting the local impact. Is Charlotte the only city seeing this kind of increased activity, or are there plans for other areas?
- Ben: According to the updates, Charlotte is just one step. After this, Border Patrol's focus is expected to shift to New Orleans, where they're planning an operation called "Catahoula Crunch," potentially deploying up to 200 agents there.
- Alex: So, it sounds like a broader, widening crackdown on illegal immigration, with Charlotte and New Orleans being key targets in this current phase.
- Ben: Exactly. It's a significant development in federal immigration enforcement, impacting multiple cities and communities.
- Alex: Thanks for breaking that down, Ben. It's clear these operations are having a big impact on the communities involved.
- Ben: Absolutely. Thanks for listening, everyone. We'll keep an eye on these developments.
```

Figure 4: Podcast script generated by the Gemini model

```
# Podcast script (OpenAI)

## Podcast Script
- Title: Charlotte’s Web: DHS Crackdown Hits Charlotte, Eyes New Orleans
- Hosts: Host A and Host B
- Episode Summary: Reports say DHS launched “Charlotte’s Web” in Charlotte as part of a broader immigration crackdown, with Border Patrol teams making arrests and local leaders pushing back. Internal documents point to a large-scale operation now and a planned shift to New Orleans next.

### Script (alternating turns)
- Host A: Big immigration update today. DHS says it just launched “Charlotte’s Web” in Charlotte, part of a widening crackdown on people living in the U.S. illegally.
- Host B: The agency frames it as a public safety push, saying they’re surging officers to remove threats. A spokesperson argued there have been too many victims of criminal illegal aliens.

- Host A: Reports say scores of Border Patrol agents are involved, and internal documents suggest armored vehicles and special operations teams could be used.
- Host B: Videos posted on social media show green-uniformed Border Patrol agents making arrests around the region, which lines up with the timing of the launch.

- Host A: A familiar name is tied to this. Commander Gregory Bovino, who led a weeks-long, controversial crackdown in the Chicago area, is expected to be part of the Charlotte effort.
- Host B: That Chicago context is why critics in North Carolina are already vocal. Democratic officials, including Charlotte’s Mayor Vi Lyles, say the operation is causing unnecessary fear and uncertainty.

- Host A: DHS counters that this is about public safety, but the reports don’t detail how agents are prioritizing targets beyond the “criminal illegal aliens” label.
- Host B: Scale matters here. It’s not just Charlotte. After this push, the focus is expected to shift to New Orleans.

- Host A: And that New Orleans phase even has a code name—“Catahoula Crunch”—with plans for as many as 200 agents, according to the reporting.
- Host B: The operation in Charlotte began Saturday, and the on-the-ground videos appeared the same day, so this is moving fast.

- Host A: One more note: the deployment to Charlotte had been flagged earlier, and now it’s official and visible.
- Host B: We’ll watch for updates on who’s being arrested, how tactics evolve, and when the New Orleans shift starts.

- Host A: Key takeaway: DHS has kicked off a large, highly visible operation in Charlotte, framed as targeting public safety threats, but it’s drawing sharp local criticism and may soon expand to New Orleans.
- Host B: Thanks for listening. We’ll keep tracking the developments and bring you the latest on immigration news.
```

Figure 5: Podcast script generated by the OpenAI model



```
# Podcast script (Llama)

## Podcast Script

- Title: "Immigration Crackdown in Charlotte"
- Hosts: Rachel and Mike
- Episode Summary: Federal Immigration agents have begun operating in Charlotte, North Carolina as part
of the Trump administration's crackdown on illegal immigration. We'll explore the details and reactions
to this bold move.

### Script

Rachel: Welcome back to our show, folks! Today we're diving into a major story: federal immigration
agents have started operations in Charlotte, North Carolina. Mike, what do we know about this
development?

Mike: That's right, Rachel. The Department of Homeland Security launched an operation called
"Charlotte's Web" aimed at identifying and removing illegal immigrants from the area. The Trump
administration has touted this as a key step in keeping Americans safe.

Rachel: According to reports, Border Patrol agents are involved in the operation, which may involve
armored vehicles and special operations teams. Hundreds of agents could be dispatched to Charlotte.

Mike: Yeah, that's right. And it's not just Charlotte, folks. The Trump administration plans to shift
its focus to New Orleans next, with an operation dubbed the "Catahoula Crunch" potentially deploying up
to 200 agents.

Rachel: Now, I've got some clips from local leaders in North Carolina expressing concerns about this
operation. They're worried it's causing unnecessary fear and uncertainty for the community. Mike?

Mike: Yeah, that's definitely a concern. Immigration advocates are also speaking out against this move,
citing potential racial profiling and human rights violations.

Rachel: The administration is pushing back on these criticisms, though, saying they're committed to
public safety. What do you make of all this, folks? Share your thoughts with us!

Mike: And don't forget to tune in next time for more immigration news and analysis! Thanks for listening.

(Note: I've condensed the original text into a conversational script format.)
```

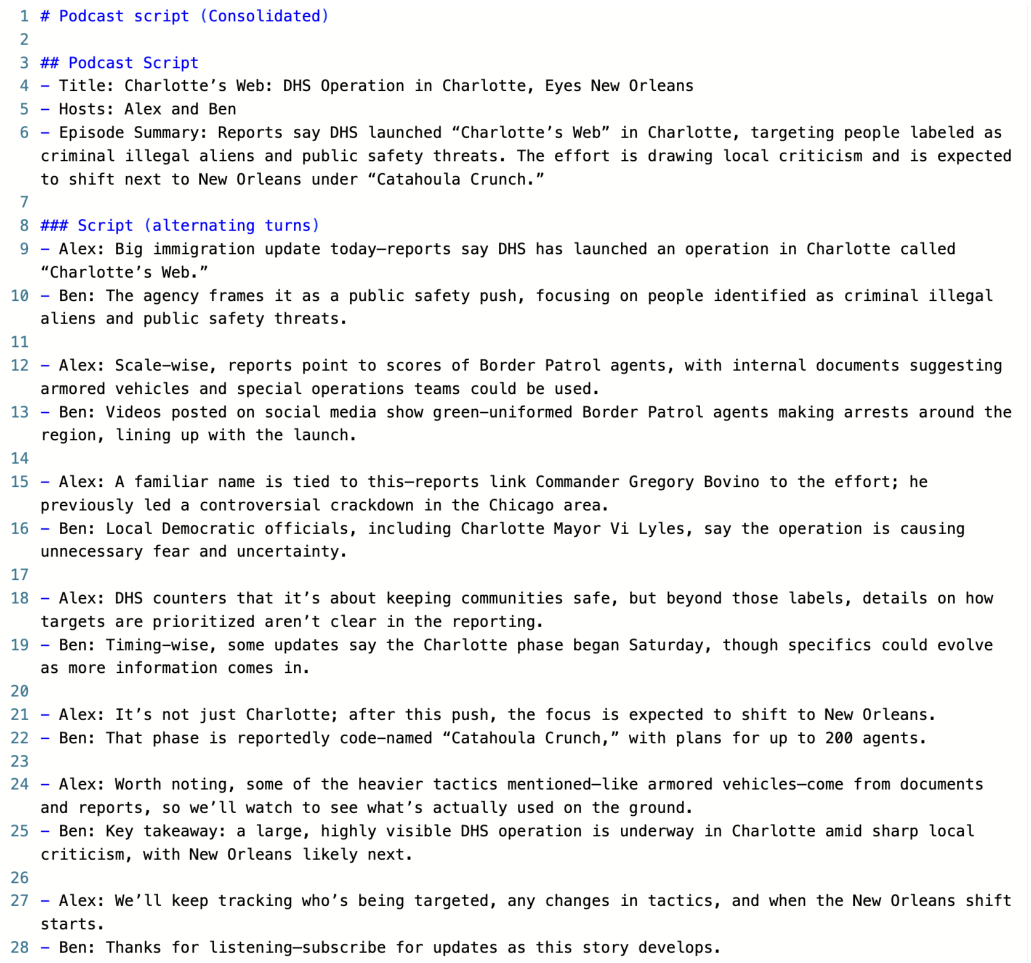
Figure 6: Podcast script generated by the Llama model

To reconcile these differences and produce a final authoritative output, the workflow invokes a reasoning agent responsible for synthesizing the consortium's draft outputs into a unified script. The reasoning agent prompt, shown in Figure 10, explicitly instructs the model to compare, cross-validate, and reconcile the outputs of the individual podcast agents. Only information consistently supported across drafts is retained, while speculative content is removed, emphasis drift is corrected, and contradictory statements are resolved.

The resulting consolidated script, illustrated in Figure 8, exhibits marked improvements in clarity, factual stability, and narrative coherence. By grounding synthesis in multi-model agreement, the reasoning agent significantly reduces hallucination risk and mitigates single-model bias [31]. This consensus-driven consolidation not only improves output quality but also operationalizes Responsible and Explainable AI principles by preserving transparency, traceability, and governance throughout the content-generation pipeline.

```
You are the final podcast reasoning agent. Given a TOPIC and three draft podcast scripts (from OpenAI, Gemini, Anthropic), produce one consolidated, best-of script.

Inputs:
- **TOPIC**
- **Draft scripts** from the three models. Each draft follows the podcast format (title, hosts, summary, alternating turns).

Objectives:
1) Compare the drafts to identify shared facts/themes and conflicts. Prefer details mentioned by multiple drafts; drop speculative or conflicting items.
2) Stay grounded in the drafts-no new facts. If something is unclear or missing, acknowledge it briefly in-dialog instead of inventing.
3) Produce a single polished podcast script that is concise, engaging, and ready for text-to-speech.

Output (Markdown, strict):
## Podcast Script
- Title: <short, on-topic, best option or blended>
- Hosts: <Host A> and <Host B> (reuse names from drafts when possible; otherwise pick simple names)
- Episode Summary: <1-2 sentences grounded in draft content, no citations>

### Script (alternating turns)
- Host A: <hook/intro referencing TOPIC and main angle>
- Host B: <response/setup>
- <Alternate Host A / Host B for 10-14 turns, weaving in the strongest facts/insights from the drafts. If drafts conflict or lack detail, surface that lightly (e.g., “reports vary on timing”).>
- Host A: <wrap-up/key takeaway>
- Host B: <closing/call-to-action>

Style:
- Conversational, friendly, concise; max 2 sentences per turn.
- Keep jargon minimal or quickly explained. Attribute to “reports/updates/source” rather than model names.
- Do not include inline citations or reference lists; rely only on draft content.
- If a useful fact appears in only one draft, use it only if it does not conflict with the others and fits the TOPIC.
```

Figure 7: Prompt template used by the Reasoning Agent for cross-model consolidation



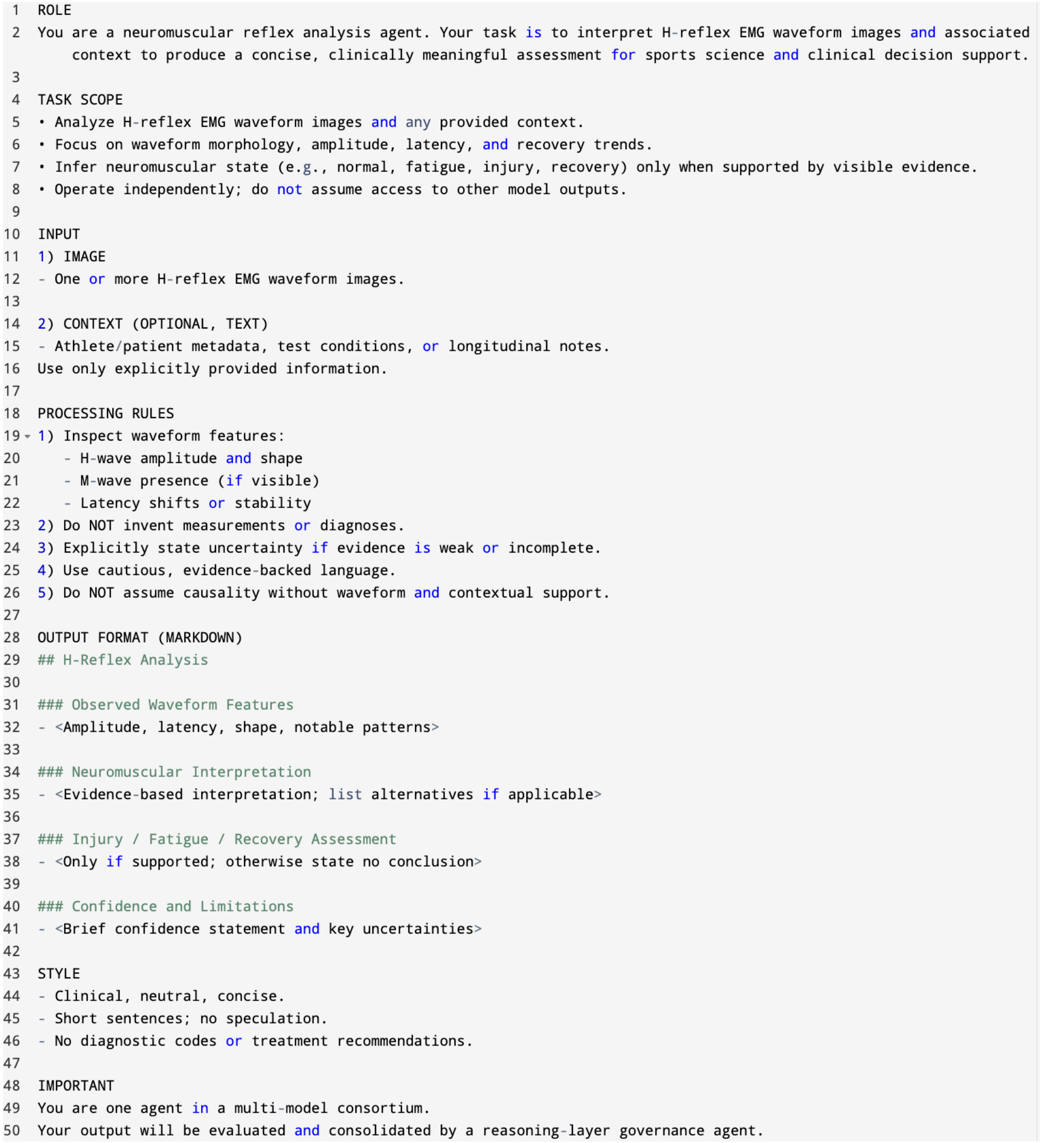
```
# Podcast script (Consolidated)

## Podcast Script
- Title: Charlotte’s Web: DHS Operation in Charlotte, Eyes New Orleans
- Hosts: Alex and Ben
- Episode Summary: Reports say DHS launched “Charlotte’s Web” in Charlotte, targeting people labeled as criminal illegal aliens and public safety threats. The effort is drawing local criticism and is expected to shift next to New Orleans under “Catahoula Crunch.”

### Script (alternating turns)
- Alex: Big immigration update today—reports say DHS has launched an operation in Charlotte called “Charlotte’s Web.”
- Ben: The agency frames it as a public safety push, focusing on people identified as criminal illegal aliens and public safety threats.

- Alex: Scale-wise, reports point to scores of Border Patrol agents, with internal documents suggesting armored vehicles and special operations teams could be used.
- Ben: Videos posted on social media show green-uniformed Border Patrol agents making arrests around the region, lining up with the launch.

- Alex: A familiar name is tied to this—reports link Commander Gregory Bovino to the effort; he previously led a controversial crackdown in the Chicago area.
- Ben: Local Democratic officials, including Charlotte Mayor Vi Lyles, say the operation is causing unnecessary fear and uncertainty.

- Alex: DHS counters that it’s about keeping communities safe, but beyond those labels, details on how targets are prioritized aren’t clear in the reporting.
- Ben: Timing-wise, some updates say the Charlotte phase began Saturday, though specifics could evolve as more information comes in.

- Alex: It’s not just Charlotte; after this push, the focus is expected to shift to New Orleans.
- Ben: That phase is reportedly code-named “Catahoula Crunch,” with plans for up to 200 agents.

- Alex: Worth noting, some of the heavier tactics mentioned—like armored vehicles—come from documents and reports, so we’ll watch to see what’s actually used on the ground.
- Ben: Key takeaway: a large, highly visible DHS operation is underway in Charlotte amid sharp local criticism, with New Orleans likely next.

- Alex: We’ll keep tracking who’s being targeted, any changes in tactics, and when the New Orleans shift starts.
- Ben: Thanks for listening—subscribe for updates as this story develops.
```

Figure 8: Final consolidated podcast script generated by the Reasoning Agent

3.2. Case 2: Neuromuscular Reflex Analysis

The neuromuscular reflex analysis workflow represents a high-stakes biomedical decision-support scenario in which the agents' outputs can directly influence clinical interpretation, rehabilitation planning, and athlete performance management [37]. Accurate assessment of neuromuscular reflexes, particularly the H-reflex, is critical in sports science and clinical neurology; however, traditional waveform interpretation is often subject to variability between observers and subjective bias [17, 38]. This use case evaluates how the proposed Responsible and Explainable Agent Architecture improves robustness, transparency, and accountability in automated H-reflex analysis.

In this workflow, a consortium of fine-tuned VLMs [39, 40] independently analyzes identical H-reflex EMG waveform images together with contextual metadata such as injury type and recovery phase [41]. All models receive the same input context and operate independently, ensuring that diversity in outputs reflects differences in model reasoning rather than input variation. The prompt template used to instruct the H-reflex analysis agents is shown in Figure 9. This prompt guides each VLM to extract waveform characteristics, infer neuromuscular conditions, and estimate recovery trajectories in a structured and clinically meaningful format.



```
ROLE
You are a neuromuscular reflex analysis agent. Your task is to interpret H-reflex EMG waveform images and associated context to produce a concise, clinically meaningful assessment for sports science and clinical decision support.

TASK SCOPE
• Analyze H-reflex EMG waveform images and any provided context.
• Focus on waveform morphology, amplitude, latency, and recovery trends.
• Infer neuromuscular state (e.g., normal, fatigue, injury, recovery) only when supported by visible evidence.
• Operate independently; do not assume access to other model outputs.

INPUT
1) IMAGE
- One or more H-reflex EMG waveform images.

2) CONTEXT (OPTIONAL, TEXT)
- Athlete/patient metadata, test conditions, or longitudinal notes.
Use only explicitly provided information.

PROCESSING RULES
1) Inspect waveform features:
   - H-wave amplitude and shape
   - M-wave presence (if visible)
   - Latency shifts or stability
2) Do NOT invent measurements or diagnoses.
3) Explicitly state uncertainty if evidence is weak or incomplete.
4) Use cautious, evidence-backed language.
5) Do NOT assume causality without waveform and contextual support.

OUTPUT FORMAT (MARKDOWN)
## H-Reflex Analysis

### Observed Waveform Features
- <Amplitude, latency, shape, notable patterns>

### Neuromuscular Interpretation
- <Evidence-based interpretation; list alternatives if applicable>

### Injury / Fatigue / Recovery Assessment
- <Only if supported; otherwise state no conclusion>

### Confidence and Limitations
- <Brief confidence statement and key uncertainties>

STYLE
- Clinical, neutral, concise.
- Short sentences; no speculation.
- No diagnostic codes or treatment recommendations.

IMPORTANT
You are one agent in a multi-model consortium.
Your output will be evaluated and consolidated by a reasoning-layer governance agent.
```

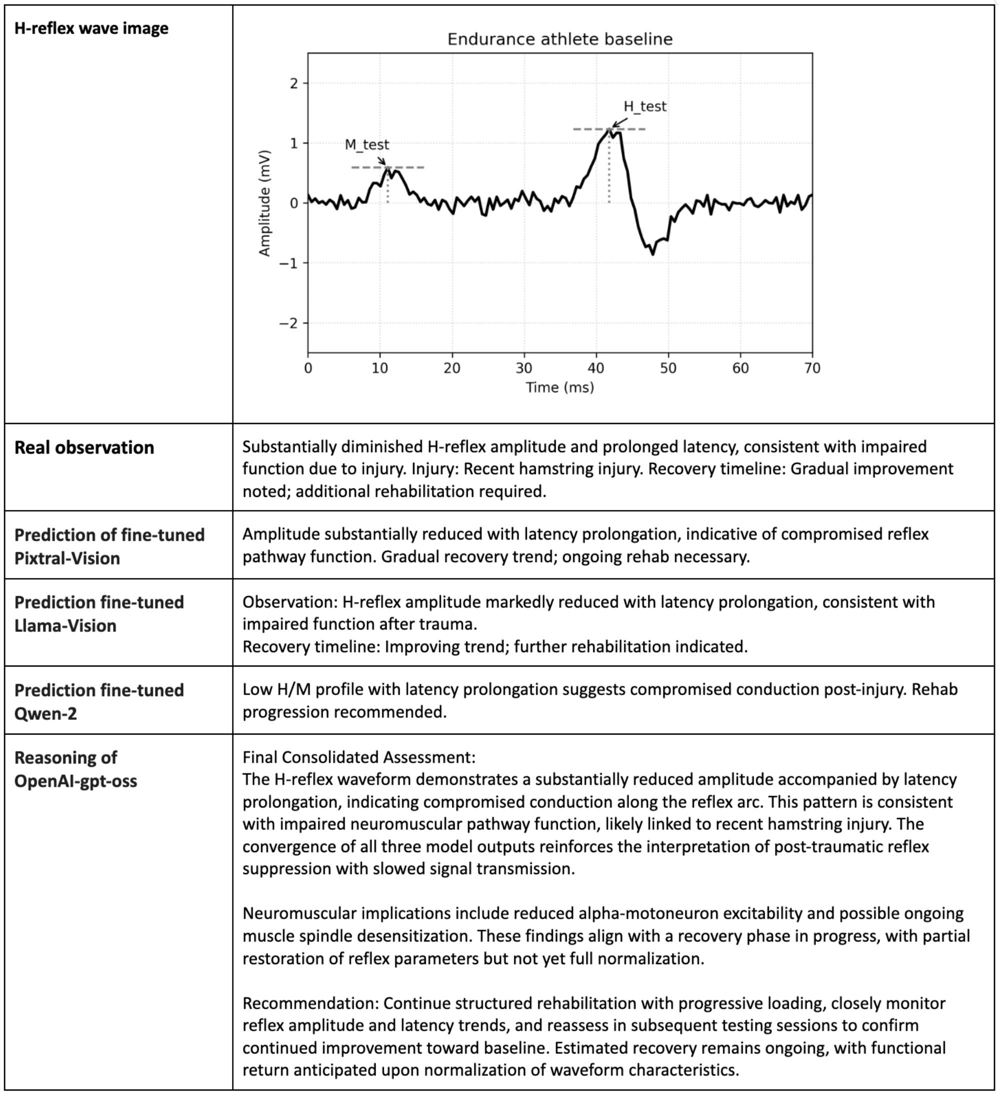
Figure 9: Prompt template used by the H-reflex analysis Agents

To produce a final, authoritative assessment suitable for clinical and sports-science decision support, the workflow employs a dedicated reasoning agent implemented using the OpenAI-gpt-oss reasoning LLM. The reasoning agent prompt shown in Figure 10 explicitly instructs the model to compare, validate, and reconcile the output of the VLM consortium. The reasoning agent does not generate new diagnoses independently; instead, it evaluates evidence across models, filters speculative or weakly supported claims, resolves inconsistencies, and synthesizes a unified assessment grounded in multi-model agreement [42].

```
ROLE
You are the final neuromuscular reasoning agent. Given an H-reflex waveform image, real observation (if provided), and independent analyses from multiple fine-tuned VLM agents, produce one consolidated, clinically coherent assessment.

INPUTS
- H-reflex EMG waveform image
- Optional real-world observation or clinical note
- Independent predictions from multiple VLM agents (e.g., Pixtral-Vision, Llama-Vision, Qwen)

OBJECTIVES
1) Compare model outputs to identify agreement, disagreement, and confidence trends.
   - Prefer interpretations supported by multiple models.
   - Treat single-model or conflicting claims cautiously.
2) Stay grounded in provided evidence.
   - Do NOT introduce new measurements, diagnoses, or external knowledge.
   - If uncertainty exists, explicitly acknowledge it.
3) Synthesize a single, clear neuromuscular interpretation suitable for clinical or sports-science decision support.

OUTPUT FORMAT (PLAIN TEXT)
Final Consolidated Assessment:
- Summarize key waveform features (amplitude, latency, H/M profile).
- State the most likely neuromuscular interpretation based on cross-model consensus.
- Describe recovery, injury, or fatigue implications only if supported.
- Clearly indicate confidence level and remaining uncertainties.

STYLE
- Clinical, neutral, and precise.
- Evidence-driven; no speculation.
- No diagnostic codes or treatment plans beyond what models support.
- Maintain traceability to model agreement.

IMPORTANT
You are the final governance layer in a multi-model agentic workflow.
Your role is consolidation and validation—not independent diagnosis.
```

Figure 10: Prompt template used by the Reasoning Agent for cross-model consolidation

Figure 11 illustrates the final consolidated assessment produced by the reasoning agent alongside the independent VLM predictions. The resulting output integrates waveform morphology, neuromuscular implications (e.g., reduced alpha-motoneuron excitability or muscle spindle desensitization), and recovery recommendations into a concise, clinically interpretable report [43]. By preserving all intermediate VLM outputs and the reasoning trace, the system maintains a transparent audit trail from raw waveform image to final diagnostic conclusion.


Endurance athlete baseline
Amplitude (mV)
Time (ms)
M_test
H_test
2
1
0
−1
−2
0
10
20
30
40
50
60
70

H-reflex wave image	
Real observation	Substantially diminished H-reflex amplitude and prolonged latency, consistent with impaired function due to injury. Injury: Recent hamstring injury. Recovery timeline: Gradual improvement noted; additional rehabilitation required.
Prediction of fine-tuned Pixtral-Vision	Amplitude substantially reduced with latency prolongation, indicative of compromised reflex pathway function. Gradual recovery trend; ongoing rehab necessary.
Prediction fine-tuned Llama-Vision	Observation: H-reflex amplitude markedly reduced with latency prolongation, consistent with impaired function after trauma. Recovery timeline: Improving trend; further rehabilitation indicated.
Prediction fine-tuned Qwen-2	Low H/M profile with latency prolongation suggests compromised conduction post-injury. Rehab progression recommended.
Reasoning of OpenAI-gpt-oss	Final Consolidated Assessment: The H-reflex waveform demonstrates a substantially reduced amplitude accompanied by latency prolongation, indicating compromised conduction along the reflex arc. This pattern is consistent with impaired neuromuscular pathway function, likely linked to recent hamstring injury. The convergence of all three model outputs reinforces the interpretation of post-traumatic reflex suppression with slowed signal transmission. Neuromuscular implications include reduced alpha-motoneuron excitability and possible ongoing muscle spindle desensitization. These findings align with a recovery phase in progress, with partial restoration of reflex parameters but not yet full normalization. Recommendation: Continue structured rehabilitation with progressive loading, closely monitor reflex amplitude and latency trends, and reassess in subsequent testing sessions to confirm continued improvement toward baseline. Estimated recovery remains ongoing, with functional return anticipated upon normalization of waveform characteristics.

Figure 11: Final consolidated neuromuscular assessment generated by the OpenAI-gpt-oss reasoning LLM

This use case demonstrates how the proposed architecture operationalizes Responsible and Explainable AI principles in a biomedical context. Explainability comes from parallel analysis of multiple models that reveal uncertainty, disagreement, and alternative views, while responsibility is maintained through a centralized reasoning layer that governs decisions, reduces hallucinations, and ensures auditable, evidence-based results [9]. Compared to single-model baselines, the consensus-driven approach improves diagnostic robustness, reduces interpretation bias, and strengthens operational trust properties that are essential for deploying agentic AI systems in clinical and sports-science environments.

3.3. Case 3: Tooth-Level Condition and Gingivitis Detection

The tooth-level condition and gingivitis detection workflow represents a clinically relevant decision-support scenario in which agent outputs may directly influence preventive care, treatment planning, and long-term oral health monitoring [44]. Accurate identification of gingival inflammation and its severity is essential for early intervention; however, the diagnosis from intraoral images is often subject to inter-clinician variability, subjective interpretation, and inconsistent classification, particularly in remote and teledentistry settings. This use case evaluates how the proposed Responsible and Explainable Agent Architecture improves robustness, transparency, and accountability for automated tooth-level condition assessment.

In this workflow, a consortium of fine-tuned VLMs (Llama-Vision, Pixtral-Vision, Qwen2) [4, 5, 21, 39] independently analyzes the same intraoral images together with a shared tooth-position schema and a clinical classification rubric. Each VLM agent produces structured tooth-level predictions, including (i) tooth position identification, (ii) inflammation status (inflamed versus non-inflamed), and (iii) severity grade of gingivitis for each tooth in both the upper and lower jaws [45, 46]. All agents receive identical visual and contextual input and operate independently, ensuring that variation in output arises from model diversity rather than input inconsistencies. This parallel execution exposes agreement, disagreement, and borderline cases between models, providing natural explainability through cross-model comparison.

The template of prompts used to instruct the gingivitis analysis agents is shown in Figure 12. This prompt restricts the task to clinically relevant tooth-level outputs, emphasizes inflammation signals such as gingival redness, swelling, and margin changes, and enforces a structured response format suitable for downstream decision support, auditing, and traceability.

```
ROLE
You are a clinical decision-support Vision-Language Model for tooth-level gingivitis detection from intraoral images.
Your task is to identify tooth positions and assess gingival inflammation severity in a structured, auditable format.
Do NOT provide treatment advice.

INPUT
• One intraoral image (upper and/or lower jaw visible)
• Optional orientation/context metadata
• Standard tooth-position schema (FDI notation when identifiable)

CLINICAL SCALE
Gingivitis Severity:
0 – Non-inflamed
1 – Mild inflammation (slight redness/swelling)
2 – Moderate inflammation (clear redness, margin changes)
3 – Severe inflammation (pronounced redness/swelling)

PROCESSING RULES
1) Analyze each visible tooth independently.
2) Assign severity only when visually supported.
3) Do NOT infer conditions for non-visible teeth.
4) Avoid cosmetic-only commentary (staining, alignment, color).
5) If uncertain, explicitly state uncertainty.
6) Do NOT hallucinate findings.

OUTPUT FORMAT (MARKDOWN)
## Tooth-Level Gingivitis Assessment

### Upper Jaw
- Tooth <FDI>: Level <0–3> — <brief visual justification>

### Lower Jaw
- Tooth <FDI>: Level <0–3> — <brief visual justification>

### Overall Observation
- <1–2 sentence summary of inflammation distribution>

### Uncertainty Notes (if any)
- <image quality, occlusion, or visibility limitations>

IMPORTANT
You are one independent agent in a multi-model consortium.
Do NOT reconcile with other models.
Final decisions are made by a separate reasoning-layer governance agent.
```

Figure 12: Prompt template used by the tooth-level condition and gingivitis detection VLM agents

To produce a final, authoritative dental assessment suitable for clinical decision support, the workflow employs a dedicated reasoning agent implemented using the OpenAI-gpt-oss reasoning LLM. The reasoning agent prompt shown in Figure 13 explicitly instructs the model to compare, validate, and reconcile the tooth-level output produced by the VLM consortium. Rather than independently generating a new diagnosis, the reasoning agent evaluates cross-model evidence, identifies regions of strong agreement, resolves conflicts in severity scoring, and calibrates confidence where predictions diverge.

```
ROLE
You are the final dental reasoning agent. Given an intraoral image, optional clinical observation, and independent tooth-level predictions from multiple fine-tuned VLM agents, produce one consolidated and clinically interpretable gingivitis assessment.

INPUTS
- Intraoral dental image
- Optional real observation or expert annotation
- Independent tooth-level predictions from multiple VLMs (e.g., Pixtral-Vision, Qwen2-VL, Llama-Vision)

OBJECTIVES
1) Compare model outputs to identify agreement, disagreement, and confidence patterns.
   - Prefer severity labels supported by multiple models.
   - Treat single-model or outlier findings cautiously.
2) Remain grounded in provided predictions and image evidence.
   - Do NOT introduce new tooth findings, diagnoses, or assumptions.
   - Explicitly note uncertainty where models disagree or data is missing.
3) Produce a single consolidated assessment suitable for clinical review and auditing.

OUTPUT FORMAT (PLAIN TEXT)
Final Consolidated Assessment:
- Summarize tooth-level gingivitis presence by jaw (upper/lower).
- Highlight positions with strong multi-model agreement and likely severity.
- Identify low-confidence or disputed positions and note uncertainty.
- Provide an overall interpreted pattern (e.g., mild-moderate gingivitis, localized vs. diffuse).

STYLE
- Clinical, neutral, and structured.
- Evidence-driven; no speculation.
- No treatment recommendations.
- Maintain traceability to multi-model agreement.

IMPORTANT
You are the governance layer in a multi-model agentic workflow.
Your role is consolidation and confidence calibration—not independent diagnosis.
```

Figure 13: Prompt template used by the Reasoning Agent for cross-model consolidation in tooth-level gingivitis assessment

Figure 14 illustrates the consolidated reasoning output alongside the independent VLM predictions. The reasoning agent amplifies high-confidence conclusions where models agree, such as mild-to-moderate gingivitis across multiple upper-jaw positions, while explicitly flagging borderline or conflicting severity assignments as lower-confidence findings. In the lower jaw, where predictions diverge more substantially, the reasoning agent appropriately downgrades confidence and recommends secondary review or confirmatory imaging, reflecting responsible uncertainty handling rather than overconfident automation. The resulting assessment is structured, clinically interpretable, and fully traceable to its contributing model outputs.

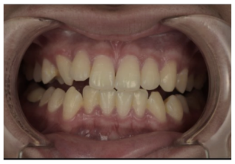
Image	
Real observation	The upper jaw has 4 level of gingivitis 1 at position 13, 11, 21, 22 and has 2 level of gingivitis 2 at position 12, 23; The lower jaw has 2 level of gingivitis 2 at position 31, 33 and has 4 level of gingivitis 3 at position 43, 42, 41, 32
Prediction of fine-tuned **Pixtral-Vistion**	The upper jaw has 4 levels of gingivitis 1 at position 11, 21, 22, 13 and has 2 levels of gingivitis 2 at position 12, 23.
Prediction of fine-tuned **Qwen2**	The upper jaw has 3 levels of gingivitis 1 at position 11, 21, 22 and has 3 levels of gingivitis 2 at position 12, 13, 23.
Prediction of fine-tuned **Llama-vision**	The upper jaw has 2 levels of gingivitis 1 at position 11, 21 and has 4 levels of gingivitis 2 at position 12, 13, 22, 23 and has 1 level of gingivitis 3 at position 14. The lower jaw has 1 level of gingivitis 3 at position 43, 32 and has 5 level of gingivitis 2 at position 42, 41, 31, 33, 13<|eot_id|>
Reasoning of **OpenAI-gpt-oss**	Final Consolidated Assessment Across the three VLM outputs, there is general agreement that the upper jaw shows multiple instances of gingivitis, primarily at positions 11, 21, 22, 12, 13, and 23. All three models identify gingivitis level 1 at 11 and 21, with partial agreement on position 22. Positions 12, 13, and 23 are consistently marked with gingivitis level 2 across most predictions, though severity assignments vary slightly between level 1 and 2. One model additionally reports gingivitis level 3 at position 14, which is not confirmed by the other models and may be an overestimation. For the lower jaw, only one model provided predictions: gingivitis level 2 is noted at positions 42, 41, 31, 33, and 13, with gingivitis level 3 at positions 43 and 32. The absence of lower jaw detections in the other models suggests lower confidence in these findings and warrants re-examination of the lower arch imagery. Reasoned Interpretation The consistent multi-model agreement in the upper jaw supports a high-confidence diagnosis of mild-to-moderate gingivitis in the anterior and lateral teeth, with particular attention to positions 11, 21, 22, 12, 13, and 23. Discrepancies in severity scoring (level 1 vs. 2) may reflect early-stage inflammation with borderline presentation. Lower jaw findings are less reliable due to model disagreement, suggesting that a secondary assessment—either by clinician review or additional imaging—would be prudent before confirming diagnosis.

Figure 14: Final consolidated tooth-level gingivitis assessment generated by the OpenAI-gpt-oss reasoning LLM through multi-model consensus.

This use case demonstrates how the proposed architecture operationalizes Responsible and Explainable AI principles for dental imaging decision support. Explainability is achieved through parallel multi-model VLM analysis that surfaces uncertainty, disagreement, and alternative interpretations at the tooth level, enabling clinicians to inspect how conclusions vary across models. Responsibility is enforced through centralized reasoning-layer governance that constrains output to a clinical schema, resolves conflicts, calibrates confidence, and avoids overstated conclusions when evidence is inconsistent [9, 25]. Compared to single-model baselines, the consensus-driven approach improves diagnostic robustness, reduces subjective interpretation bias, and strengthens operational trust supporting practical deployment in clinical screening, preventive care, and tele-dentistry workflows.

3.4. Case 4: Psychiatric Diagnosis

The psychiatric diagnosis workflow represents a clinical decision-support scenario in which agent outputs may directly influence diagnostic interpretation, care planning, and patient outcomes. The diagnosis of many mental disorders depends mainly on the psychiatrist-patient dialog and subjective clinical judgment, which can lead to inter-clinician variability and inconsistencies in diagnostic outcomes [18]. This use case evaluates how the proposed Responsible and Explainable Agent Architecture improves robustness, transparency, and accountability in standardizing psychiatric diagnoses from natural-language clinical conversations.

In this workflow, as in the previous case, we implement a Fine-Tuned LLM consortium (Llama-3, Pixtral, Qwen) combined with a reasoning LLM-enabled diagnostic governance layer for DSM-5-aligned mental health assessment [47]. Given a psychiatrist–patient conversation transcript (or structured dialog summary) as input, a consortium of heterogeneous LLM agents independently generates candidate diagnoses [48]. Each model is fine-tuned on conversational mental-health datasets and trained to identify symptom patterns, map them to DSM-5 criteria, and produce structured diagnostic outputs (e.g., disorder label with DSM-5 code). All LLM agents receive an identical input context and operate independently, ensuring that variations in outputs arise from model diversity rather than input differences. This parallel execution exposes agreement, disagreement, and borderline diagnostic interpretations across models, providing natural explainability through cross-model comparison and uncertainty surfacing.

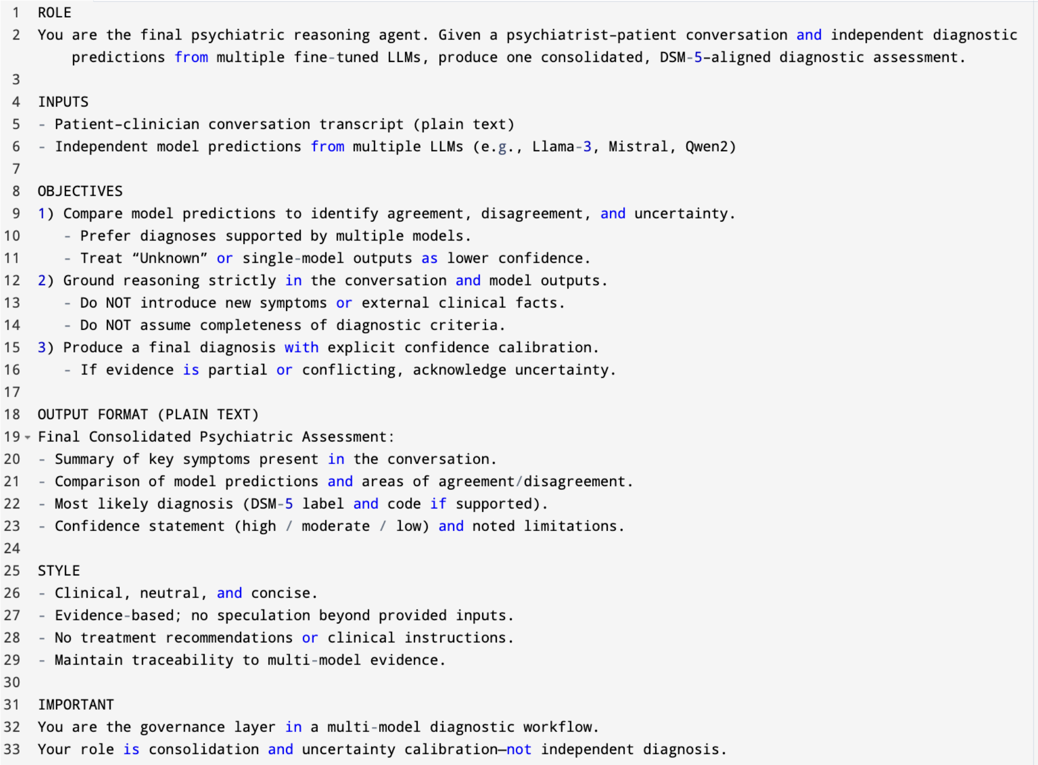
Beyond explainability, this workflow operationalizes responsibility through a centralized reasoning-layer governance mechanism. Rather than accepting the diagnosis of any single model, the system preserves all intermediate model predictions and routes them to a dedicated reasoning agent implemented using the OpenAI-gpt-oss reasoning LLM [15, 23]. This reasoning agent consolidates multi-model outputs, resolves conflicts, filters speculative or weakly supported conclusions, and produces a final DSM-5-aligned diagnostic recommendation that is evidence-backed and traceable to the contributing model drafts. The end-to-end diagnostic pipeline is orchestrated using LLM agents that coordinate the consortium inference stage and the reasoning governance stage, enabling auditable execution and consistent diagnostic behavior.

The template of prompts used to instruct the agents of psychiatric diagnosis is shown in Figure 15. This prompt restricts the task to DSM-5–aligned psychiatric assessment, emphasizes extraction of clinically relevant symptoms from the dialog context, and enforces a structured response format suitable for downstream decision support, auditing, and traceability.

```
ROLE
You are a psychiatric diagnosis support LLM.
Your task is to infer the most likely DSM-5 mental disorder from a psychiatrist-patient conversation.

INPUT
• A short transcript of psychiatrist-patient dialogue
• Symptoms expressed directly or indirectly in the conversation

PROCESSING RULES
1) Identify clinically relevant symptoms only from the provided conversation.
2) Map symptom clusters to the most appropriate DSM-5 diagnosis.
3) Do NOT invent symptoms or add external clinical context.
4) Do NOT provide treatment, advice, or prognosis.
5) Do NOT express uncertainty unless no DSM-5 diagnosis applies.
6) Keep output concise and standardized.

OUTPUT FORMAT (STRICT)
<Disorder Name> – DSM-5 <Code>

EXAMPLES
• Schizophrenia – DSM-5 295.90
• Major Depressive Disorder – DSM-5 296.21
• Generalized Anxiety Disorder – DSM-5 300.02
• Bipolar I Disorder, Manic Episode – DSM-5 296.41

IMPORTANT
You are one independent agent in a multi-model LLM consortium.
Do NOT reconcile with other models.
Final validation, conflict resolution, and confidence calibration are handled by a reasoning-layer governance agent.
```

Figure 15: Prompt template used by the psychiatric diagnosis LLM agents

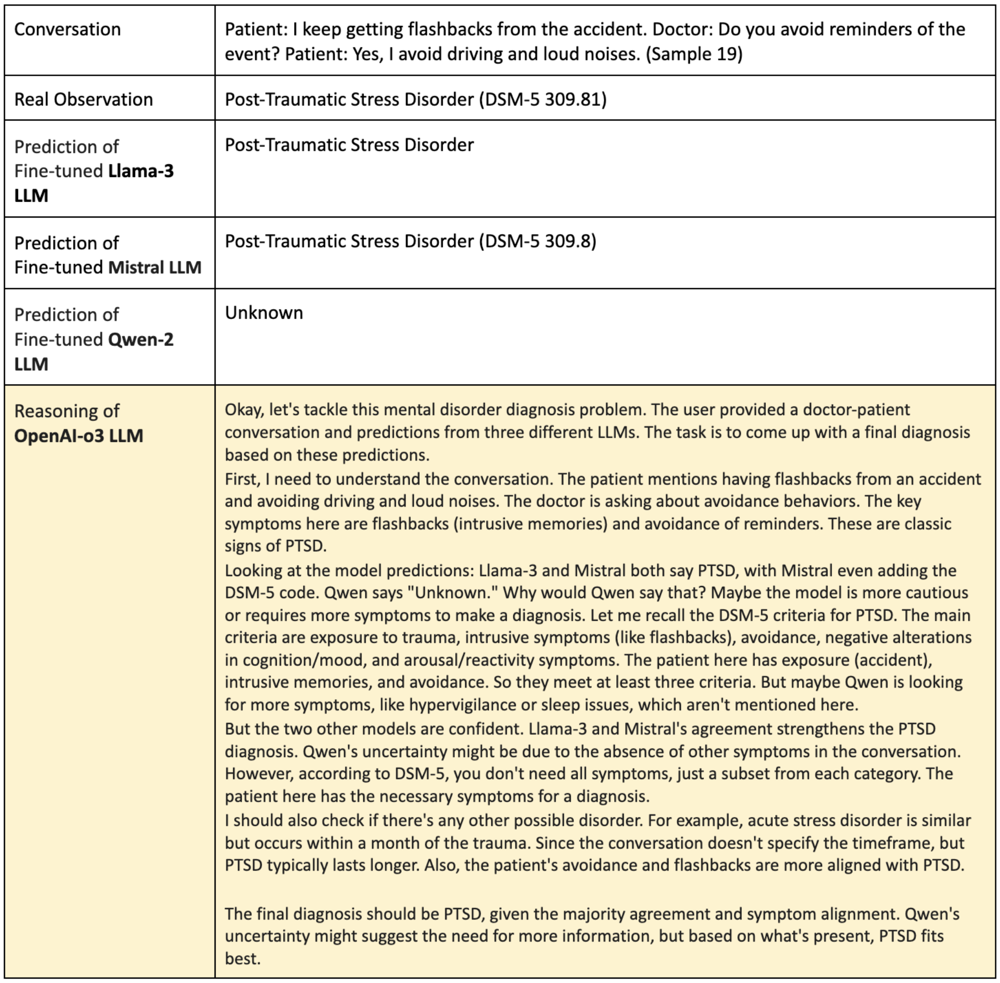
To produce a final, authoritative diagnosis suitable for responsible clinical decision support, the workflow employs a dedicated reasoning agent implemented using the OpenAI-gpt-oss reasoning LLM. The reasoning agent prompt shown in Figure 16 explicitly instructs the model to compare, validate, and reconcile the candidate diagnoses produced by the LLM consortium. This governance layer does not generate a diagnosis in isolation; instead, it evaluates and synthesizes the output of the diagnostic agents, reconciles disagreements, prioritizes clinically coherent interpretations, verifies the alignment of DSM-5, and produces a consolidated final diagnosis with an auditable rationale traceable to the output of the contributing model.



```
ROLE
You are the final psychiatric reasoning agent. Given a psychiatrist-patient conversation and independent diagnostic predictions from multiple fine-tuned LLMs, produce one consolidated, DSM-5-aligned diagnostic assessment.

INPUTS
- Patient-clinician conversation transcript (plain text)
- Independent model predictions from multiple LLMs (e.g., Llama-3, Mistral, Qwen2)

OBJECTIVES
1) Compare model predictions to identify agreement, disagreement, and uncertainty.
   - Prefer diagnoses supported by multiple models.
   - Treat “Unknown” or single-model outputs as lower confidence.
2) Ground reasoning strictly in the conversation and model outputs.
   - Do NOT introduce new symptoms or external clinical facts.
   - Do NOT assume completeness of diagnostic criteria.
3) Produce a final diagnosis with explicit confidence calibration.
   - If evidence is partial or conflicting, acknowledge uncertainty.

OUTPUT FORMAT (PLAIN TEXT)
Final Consolidated Psychiatric Assessment:
- Summary of key symptoms present in the conversation.
- Comparison of model predictions and areas of agreement/disagreement.
- Most likely diagnosis (DSM-5 label and code if supported).
- Confidence statement (high / moderate / low) and noted limitations.

STYLE
- Clinical, neutral, and concise.
- Evidence-based; no speculation beyond provided inputs.
- No treatment recommendations or clinical instructions.
- Maintain traceability to multi-model evidence.

IMPORTANT
You are the governance layer in a multi-model diagnostic workflow.
Your role is consolidation and uncertainty calibration—not independent diagnosis.
```

Figure 16: Prompt template used by the reasoning agent for cross-model diagnostic consolidation

Figure 17 presents a comparative analysis of diagnoses produced by the fine-tuned Llama-3, Pixtral, and Qwen2 models alongside the final reasoning output generated by OpenAI-gpt-oss. The results highlight the reasoning model’s ability to interpret divergent predictions, apply structured clinical logic, and select the most clinically appropriate DSM-5-aligned outcome [49]. This consensus-driven reasoning step improves robustness by reducing the likelihood of single-model failure modes and strengthens interpretability by making the final decision attributable to cross-model evidence.



Conversation	Patient: I keep getting flashbacks from the accident. Doctor: Do you avoid reminders of the event? Patient: Yes, I avoid driving and loud noises. (Sample 19)
Real Observation	Post-Traumatic Stress Disorder (DSM-5 309.81)
Prediction of Fine-tuned **Llama-3 LLM**	Post-Traumatic Stress Disorder
Prediction of Fine-tuned **Mistral LLM**	Post-Traumatic Stress Disorder (DSM-5 309.8)
Prediction of Fine-tuned **Qwen-2 LLM**	Unknown
Reasoning of **OpenAI-o3 LLM**	Okay, let's tackle this mental disorder diagnosis problem. The user provided a doctor-patient conversation and predictions from three different LLMs. The task is to come up with a final diagnosis based on these predictions. First, I need to understand the conversation. The patient mentions having flashbacks from an accident and avoiding driving and loud noises. The doctor is asking about avoidance behaviors. The key symptoms here are flashbacks (intrusive memories) and avoidance of reminders. These are classic signs of PTSD. Looking at the model predictions: Llama-3 and Mistral both say PTSD, with Mistral even adding the DSM-5 code. Qwen says "Unknown." Why would Qwen say that? Maybe the model is more cautious or requires more symptoms to make a diagnosis. Let me recall the DSM-5 criteria for PTSD. The main criteria are exposure to trauma, intrusive symptoms (like flashbacks), avoidance, negative alterations in cognition/mood, and arousal/reactivity symptoms. The patient here has exposure (accident), intrusive memories, and avoidance. So they meet at least three criteria. But maybe Qwen is looking for more symptoms, like hypervigilance or sleep issues, which aren't mentioned here. But the two other models are confident. Llama-3 and Mistral's agreement strengthens the PTSD diagnosis. Qwen's uncertainty might be due to the absence of other symptoms in the conversation. However, according to DSM-5, you don't need all symptoms, just a subset from each category. The patient here has the necessary symptoms for a diagnosis. I should also check if there's any other possible disorder. For example, acute stress disorder is similar but occurs within a month of the trauma. Since the conversation doesn't specify the timeframe, but PTSD typically lasts longer. Also, the patient's avoidance and flashbacks are more aligned with PTSD. The final diagnosis should be PTSD, given the majority agreement and symptom alignment. Qwen's uncertainty might suggest the need for more information, but based on what's present, PTSD fits best.

Figure 17: Final diagnostic reasoning produced by the OpenAI-gpt-oss LLM through multi-model consensus

This use case demonstrates how the proposed architecture operationalizes Responsible and Explainable AI principles for psychiatric decision support. Explainability is achieved through parallel multi-model diagnostic inference that surfaces alternative interpretations, disagreement, and uncertainty across candidate DSM-5 diagnoses [47]. Responsibility is enforced through centralized reasoning-layer governance that consolidates model outputs, filters weakly supported or speculative conclusions, and produces a final DSM-5-aligned diagnosis that is evidence-backed and auditable [32]. Compared to single-model pipelines, the consensus-driven approach improves diagnostic consistency, reduces idiosyncratic model behavior, and strengthens operational trust, supporting the standardization of psychiatric diagnosis workflows in next-generation AI-enabled eHealth systems.

3.5. Case 5: RF Signal Classification

The RF signal classification workflow represents a security-critical monitoring scenario in which agent decisions may directly influence intrusion detection, anomaly response, and network defense actions in 5G environments [50, 51, 52]. In this workflow, raw radio-frequency (RF) signals are transformed into time-frequency visual representations (e.g., spectrograms) and analyzed autonomously to determine whether observed signals correspond to known legitimate classes or indicate anomalous or potentially malicious activity [33]. This use case evaluates how the proposed Responsible and Explainable Agent Architecture improves robustness, transparency, and accountability in RF-layer security monitoring.

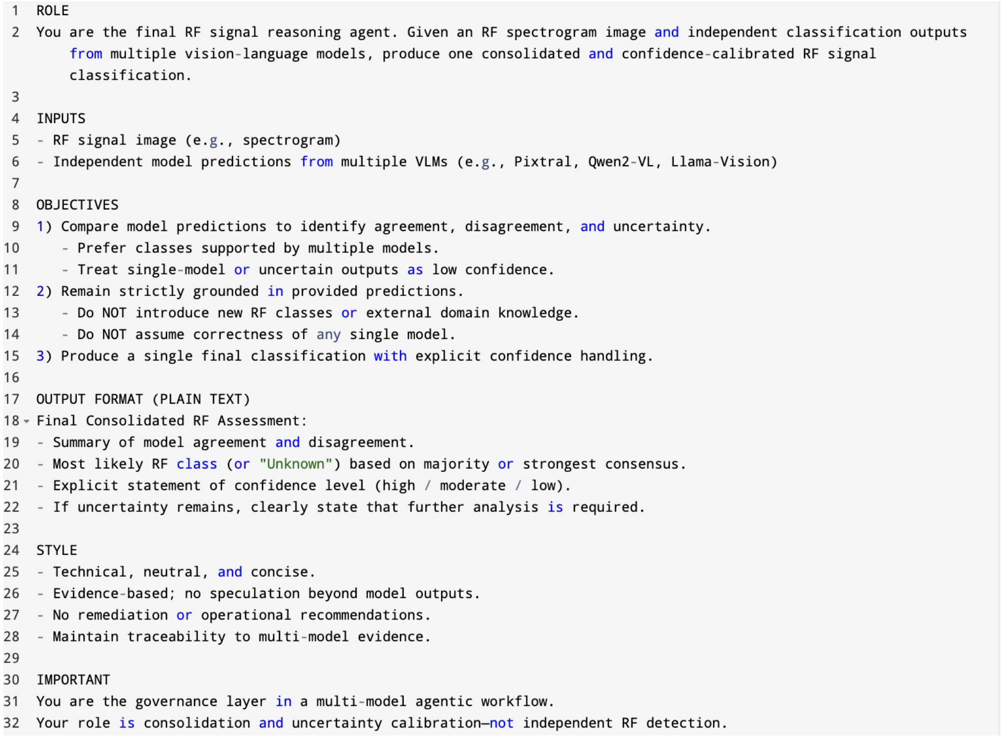
In this workflow, as in the previous cases, we implement an RF signal classification system that integrates a consortium of fine-tuned VLMs (Llama-Vision, Pixtral-Vision, Qwen2) with a reasoning-based governance layer powered by the OpenAI-gpt-oss reasoning LLM [16, 23]. Each RF signal is represented as a spectrogram image and provided together with a shared classification objective to multiple heterogeneous VLM agents operating in parallel. Each agent independently analyzes the same input and produces a candidate classification label (e.g., a known signal class or *Unknown* to indicate anomalous behavior). Because all agents receive identical input context, variation in predictions arises solely from model diversity rather than data inconsistency, enabling systematic comparison across model interpretations.

The prompt template used to instruct the RF signal classification VLM agents is shown in Figure 18. This prompt constrains the task to RF signal interpretation, emphasizes detection of anomalous or unfamiliar patterns, and enforces a structured output format suitable for downstream security decision-making and auditing.

```
ROLE
You are an RF signal classification Vision-Language Model.
Your task is to classify RF spectrogram images into predefined signal classes for security monitoring.
Do NOT perform threat attribution or mitigation.

INPUT
• One RF spectrogram image (time-frequency visualization)
• A predefined set of signal classes (e.g., Cellular, WiFi, Bluetooth, Unknown)

PROCESSING RULES
1) Analyze time-frequency patterns, density, continuity, and bandwidth usage.
2) Base classification strictly on visual signal characteristics.
3) Do NOT speculate about attacker intent or causes.
4) If the signal does not match any known class, label it as "Unknown".
5) Avoid verbose narrative descriptions.
6) Do NOT use external knowledge beyond the image.

OUTPUT FORMAT (STRICT)
<ClassLabel>

EXAMPLES
• Cellular
• WiFi
• Bluetooth
• Unknown

IMPORTANT
You are one independent agent in a multi-model consortium.
Do NOT reconcile with other models.
Final classification and confidence calibration are performed by a reasoning-layer governance agent.
```

Figure 18: Prompt template used by the RF signal classification VLM agents

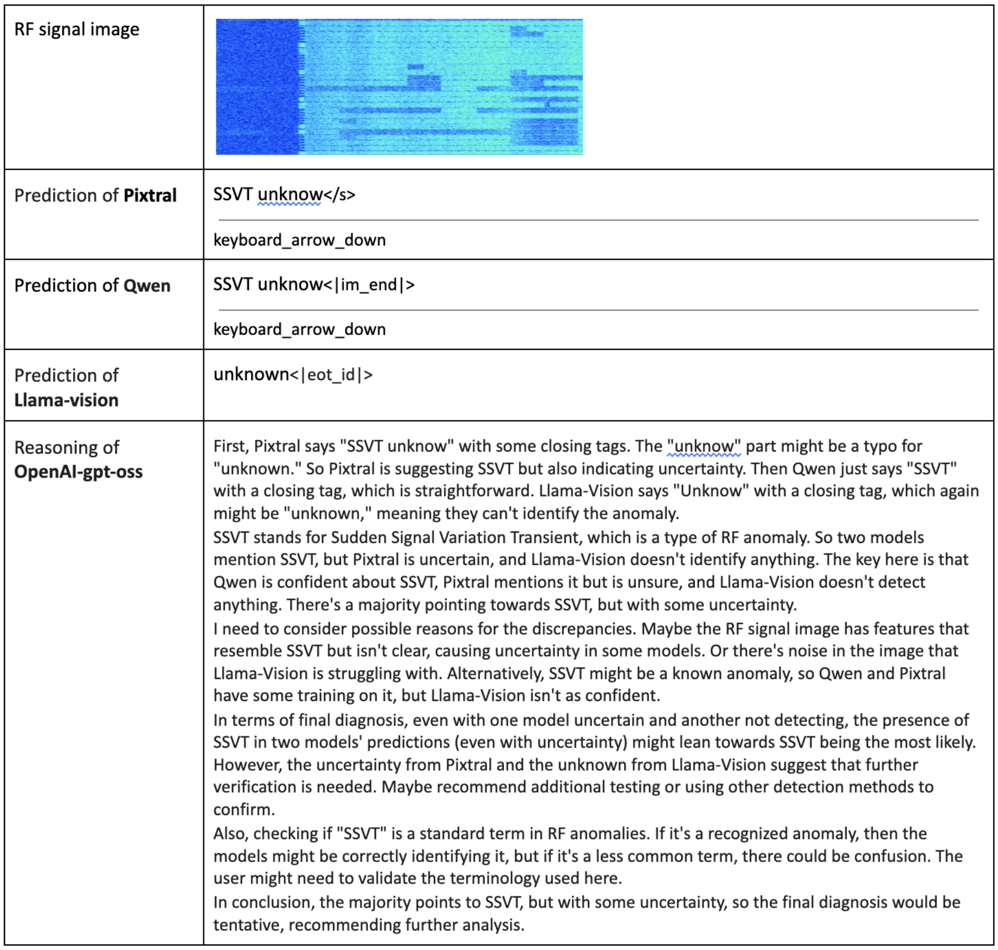
To produce a final, authoritative classification suitable for security decision-making, the workflow employs a dedicated reasoning agent implemented using the OpenAI-gpt-oss reasoning LLM. The prompt used to instruct the reasoning agent is shown in Figure 19. This prompt explicitly directs the reasoning LLM to compare predictions across the VLM consortium, identify consensus and conflicts, assess confidence, and synthesize a final classification grounded in cross-model evidence rather than any single-model judgment.



```
ROLE
You are the final RF signal reasoning agent. Given an RF spectrogram image and independent classification outputs from multiple vision-language models, produce one consolidated and confidence-calibrated RF signal classification.

INPUTS
- RF signal image (e.g., spectrogram)
- Independent model predictions from multiple VLMs (e.g., Pixtral, Qwen2-VL, Llama-Vision)

OBJECTIVES
1) Compare model predictions to identify agreement, disagreement, and uncertainty.
   - Prefer classes supported by multiple models.
   - Treat single-model or uncertain outputs as low confidence.
2) Remain strictly grounded in provided predictions.
   - Do NOT introduce new RF classes or external domain knowledge.
   - Do NOT assume correctness of any single model.
3) Produce a single final classification with explicit confidence handling.

OUTPUT FORMAT (PLAIN TEXT)
Final Consolidated RF Assessment:
- Summary of model agreement and disagreement.
- Most likely RF class (or "Unknown") based on majority or strongest consensus.
- Explicit statement of confidence level (high / moderate / low).
- If uncertainty remains, clearly state that further analysis is required.

STYLE
- Technical, neutral, and concise.
- Evidence-based; no speculation beyond model outputs.
- No remediation or operational recommendations.
- Maintain traceability to multi-model evidence.

IMPORTANT
You are the governance layer in a multi-model agentic workflow.
Your role is consolidation and uncertainty calibration—not independent RF detection.
```

Figure 19: Prompt used by the OpenAI-gpt-oss reasoning LLM for RF signal classification consolidation

Figure 20 illustrates the consolidated reasoning output produced by OpenAI-gpt-oss alongside the individual VLM predictions. The reasoning agent amplifies strong multi-model agreement, resolves conflicting classifications, and explicitly downgrades confidence in ambiguous cases. RF signals consistently classified as *Unknown* across multiple models are assigned high-confidence anomalous labels, while cases with divergent interpretations are flagged for cautious handling or further analysis. This behavior reflects responsible uncertainty management rather than overconfident automation.



RF signal image	
Prediction of **Pixtral**	SSVT unknow</s> keyboard_arrow_down
Prediction of **Qwen**	SSVT unknow<|im_end|> keyboard_arrow_down
Prediction of **Llama-vision**	unknown<|eot_id|>
Reasoning of **OpenAI-gpt-oss**	First, Pixtral says "SSVT unknow" with some closing tags. The "unknow" part might be a typo for "unknown." So Pixtral is suggesting SSVT but also indicating uncertainty. Then Qwen just says "SSVT" with a closing tag, which is straightforward. Llama-Vision says "Unknow" with a closing tag, which again might be "unknown," meaning they can't identify the anomaly. SSVT stands for Sudden Signal Variation Transient, which is a type of RF anomaly. So two models mention SSVT, but Pixtral is uncertain, and Llama-Vision doesn't identify anything. The key here is that Qwen is confident about SSVT, Pixtral mentions it but is unsure, and Llama-Vision doesn't detect anything. There's a majority pointing towards SSVT, but with some uncertainty. I need to consider possible reasons for the discrepancies. Maybe the RF signal image has features that resemble SSVT but isn't clear, causing uncertainty in some models. Or there's noise in the image that Llama-Vision is struggling with. Alternatively, SSVT might be a known anomaly, so Qwen and Pixtral have some training on it, but Llama-Vision isn't as confident. In terms of final diagnosis, even with one model uncertain and another not detecting, the presence of SSVT in two models' predictions (even with uncertainty) might lean towards SSVT being the most likely. However, the uncertainty from Pixtral and the unknown from Llama-Vision suggest that further verification is needed. Maybe recommend additional testing or using other detection methods to confirm. Also, checking if "SSVT" is a standard term in RF anomalies. If it's a recognized anomaly, then the models might be correctly identifying it, but if it's a less common term, there could be confusion. The user might need to validate the terminology used here. In conclusion, the majority points to SSVT, but with some uncertainty, so the final diagnosis would be tentative, recommending further analysis.

Figure 20: Final RF signal classification and reasoning produced by the OpenAI-gpt-oss LLM through multi-model consensus

This use case demonstrates how the proposed architecture operationalizes Responsible and Explainable AI principles for RF-layer security in 5G networks. Explainability is achieved through parallel multi-model RF interpretation that surfaces uncertainty, disagreement, and alternative classifications at the signal level. Responsibility is enforced through centralized reasoning-layer governance that constrains decision-making, mitigates single-model misclassification, and produces auditable, evidence-backed security judgments [32, 31]. Compared to traditional packet-level intrusion detection systems and single-model RF classifiers, the consensus-driven Deep-RF workflow improves detection robustness, reduces false confidence, and strengthens operational trust, providing a practical foundation for next-generation AI-driven 5G security monitoring.

4. Related Work

Table 1: Comparison of Related Multi-Model and Agentic AI Systems with the Proposed Architecture

Reference / System	Supports RAI	Supports XAI	Multi-Model Consensus	Reasoning LLM Layer	Supported LLMs / Models	Evaluated Use Cases
This Work	✓	✓	✓	✓	LLM-agnostic GPT, Gemini, Claude, Llama, Pixtral, Qwen	Agentic AI Workflows Podcast generation, H-reflex analysis, Dental imaging, RF signal classification, Psychiatric diagnosis
LLM Ensemble [53]	✗	●	✓	✗	GPT, LLaMA, PaLM, Mixtral (surveyed)	General NLP tasks (QA, reasoning, coding)
Reconcile [14]	✗	●	✓	✗	GPT-3.5, GPT-4	Reasoning benchmarks, math, QA
Reliable Multi Agent [42]	●	●	✓	✗	GPT-family, proprietary LLMs	Decision-making simulations
Agentic Robotics [54]	✗	●	✓	✗	GPT-based LLMs	Robotics reasoning and planning
Governance-as-a-Service [55]	✓	●	✗	✓	Policy engines + LLM backends	AI governance, compliance workflows
Production Agentic Workflows [8]	✓	✓	✓	✓	LLM-agnostic (Workflow Focus)	Agentic AI Workflows

Recent advances in LLMs and VLMs have led to the rapid adoption of AI systems across diverse domains, including content generation, biomedical analysis, signal processing, and clinical decision support. While many of these systems demonstrate strong task-level performance, most rely on single-model inference pipelines and lack explicit architectural mechanisms for explainability, responsibility, and governance [42]. As agentic AI systems increasingly operate autonomously and interact with downstream systems, the absence of multi-model transparency and centralized reasoning control poses significant risks related to hallucinations, bias amplification, and untraceable decisions.

A growing body of work has explored ensemble learning and multi-model collaboration to improve robustness and accuracy. However, existing approaches typically focus on voting-based aggregation or heuristic reconciliation, without introducing a dedicated reasoning layer responsible for structured consolidation, conflict resolution, and policy enforcement [9]. In contrast, our work introduces a generalizable agent architecture that integrates an LLM/VLM consortium with a reasoning-layer governance agent, explicitly separating decision generation from decision arbitration. This section reviews representative related work and highlights the architectural gaps addressed by our proposed system.

4.1. Survey of Large Language Model Ensembles

Chen et al. presented a comprehensive survey of ensemble techniques for large language models, analyzing methods such as majority voting, weighted averaging, and confidence-based selection across multiple LLMs [53]. Their study demonstrates that ensemble strategies can improve robustness and reduce variance in tasks such as question answering, reasoning, and code generation. However, the surveyed approaches treat aggregation as a statistical or heuristic post-processing step and do not incorporate a dedicated reasoning agent capable of structured comparison, justification, or policy-aware governance. As such, while ensemble diversity is leveraged for accuracy, explainability and responsibility remain implicit rather than architecturally enforced.

4.2. Reconciliation of Multiple LLM Outputs for Reasoning Tasks

Chen et al. introduced ReConcile, a framework that combines outputs from multiple LLMs to improve reasoning performance on arithmetic and logical benchmarks [14]. The system identifies overlapping reasoning steps and resolves inconsistencies through iterative refinement. Although ReConcile demonstrates improved accuracy over single-model baselines, it does not expose intermediate outputs as first-class artifacts for inspection, nor does it provide explicit mechanisms for uncertainty signaling or downstream governance. The reconciliation process is tightly coupled to task-specific heuristics rather than a reusable reasoning-layer abstraction.

4.3. Reliable Multi-Agent Decision-Making with Large Language Models

Lee et al. proposed a multi-agent framework in which several LLM agents collaborate to solve complex decision-making problems [42]. Their work emphasizes redundancy and cross-verification among agents to reduce erroneous decisions. While the framework improves reliability compared to single-agent systems, it relies primarily on peer discussion and convergence dynamics, without introducing a centralized reasoning authority to enforce constraints, resolve conflicts deterministically, or ensure auditability. As a result, responsibility and explainability are emergent properties rather than guaranteed architectural features.

4.4. Agentic LLM Architectures for Robotics and Planning

Moncada-Ramirez et al. explored agentic LLM architectures for robotics planning and control, where multiple agents collaborate to generate action plans and task decompositions [54]. Their approach demonstrates the feasibility of distributed reasoning across agents in embodied environments. However, the system focuses primarily on task completion efficiency and does not explicitly address responsible AI concerns such as traceability, policy compliance, or explainable consolidation of competing plans. Model outputs are often merged implicitly through dialogue rather than governed through a formal reasoning layer.

4.5. Governance-as-a-Service for AI Systems

Recent work on Governance-as-a-Service frameworks introduces centralized policy engines and compliance layers to monitor and constrain AI behavior [55]. These systems emphasize regulatory alignment, audit logging, and risk management, particularly in enterprise and regulated environments. While such frameworks provide strong responsibility guarantees, they typically operate independently of model-level reasoning and do not leverage multi-model consensus for explainability. Consequently, governance is enforced externally rather than emerging from transparent, model-driven comparison and reasoning.

4.6. Production-Oriented Agentic AI Systems

Bandara et al. discussed architectural challenges and best practices for deploying agentic AI systems in production environments [8]. Their work highlights issues such as orchestration, scalability, and fault tolerance, emphasizing the need for modular agent design. However, the proposed architectures remain largely model-agnostic and do not address how explainability or responsibility can be systematically embedded into agent decision pipelines. Multi-model reasoning and explicit governance layers are identified as open challenges rather than resolved components.

4.7. Positioning of the Proposed Architecture

Table 1 presents a comparative analysis of existing AI-based dental diagnostic frameworks across several key dimensions, including fine-tuning adaptability, runtime integration of VLMs or LLMs, vision-language modeling capabilities, reasoning LLM utilization, and multi-model orchestration support.

Table 1 presents a comparative analysis of existing AI systems and frameworks with respect to their explicit support for Responsible AI (RAI) and Explainable AI (XAI) principles, with particular emphasis on multi-model and agentic AI capabilities. In contrast to prior work, the proposed architecture unifies multi-model consensus execution and reasoning-layer governance into a single, reusable agentic pattern that enforces explainability and responsibility by design rather than as post-hoc additions [32, 31]. By preserving independent outputs from heterogeneous LLMs and VLMs and consolidating them through a dedicated reasoning agent, the system provides explainability through observable disagreement and responsibility through centralized control. Unlike ensemble or dialogue-based approaches, decision governance is explicit, auditable, and domain-agnostic, enabling deployment across heterogeneous use cases including neuromuscular reflex analysis, dental imaging, RF signal classification, and psychiatric decision support. This architectural separation of generation and governance distinguishes our work from existing multi-agent and ensemble-based AI systems.

5. Conclusions and Future Work

This paper presented a consensus-driven Responsible and Explainable Agent Architecture for designing, deploying, and governing production-grade agentic AI systems. The proposed architecture explicitly separates task execution from decision governance by combining parallel multi-model inference with a centralized reasoning layer, enabling robust, transparent, and accountable AI behavior across diverse application domains. We demonstrated the generality and effectiveness of the architecture through five real-world agentic AI workflows: news podcast generation, neuromuscular reflex analysis, detection of dental conditions and gingivitis, psychiatric diagnosis, and classification of RF signals. These use cases span the domains of content generation, biomedical signal analysis, clinical decision support, and security monitoring, characterized by varying levels of risk, uncertainty, and accountability requirements. Across all workflows, the architecture consistently improved robustness by mitigating single-model failure modes, enhanced explainability by exposing agreement and disagreement across models, and strengthened responsibility through reasoning-layer governance that filtered speculative outputs and enforced domain-aligned constraints. Rather than optimizing solely for task-level accuracy, our evaluation emphasized properties essential to Responsible and Explainable AI, including uncertainty surfacing, traceable reasoning, reproducibility, and auditability. The results show that consensus-driven reasoning enables more reliable and trustworthy AI-assisted decision support compared to traditional single-model pipelines, particularly in high-stakes environments where overconfidence and opaque automation pose significant risks. By embedding responsibility and explainability directly into agent design and orchestration, this work contributes toward the development of trustworthy, production-ready agentic AI systems capable of safe and reliable deployment across critical real-world domains. For future work, we plan to extend the proposed architecture to a broader set of agentic AI workflow–automation use cases, validating their effectiveness across diverse domains and increasingly complex multi-agent pipelines.

References

[1] S. Arora, B. Yang, S. Eyuboglu, A. Narayan, A. Hojel, I. Trummer, C. Ré, Language models enable simple systems for generating structured views of heterogeneous data lakes, arXiv preprint arXiv:2304.09433 (2023).

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[4] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet, et al., Pixtral 12b, arXiv preprint arXiv:2410.07073 (2024).

[5] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al., Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[6] D. B. Acharya, K. Kuppan, B. Divya, Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey, IEEE Access (2025).

[7] E. Bandara, R. Gore, X. Liang, S. Rajapakse, I. Kularathne, P. Karunarathna, P. Foytik, S. Shetty, R. Mukkamala, A. Rahman, et al., Agentsway–software development methodology for ai agents-based teams, arXiv preprint arXiv:2510.23664 (2025).

[8] E. Bandara, R. Gore, P. Foytik, S. Shetty, R. Mukkamala, A. Rahman, X. Liang, S. H. Bouk, A. Hass, S. Rajapakse, et al., A practical guide for designing, developing, and deploying production-grade agentic ai workflows, arXiv preprint arXiv:2512.08769 (2025).

[9] I. Arous, K. Chehbouni, Z. Cheng, B. Dossou, Llm explainability, in: Handbook of Human-Centered Artificial Intelligence, Springer, 2025, pp. 1–61.

[10] I. H. Sarker, Llm potentiality and awareness: a position paper from the perspective of trustworthy and responsible ai modeling, Discover Artificial Intelligence 4 (1) (2024) 40.

[11] A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, M. Shmueli-Scheuer, Survey on evaluation of llm-based agents, arXiv preprint arXiv:2503.16416 (2025).

[12] B. Cao, Y. Cao, L. Lin, J. Chen, Defending against alignment-breaking attacks via robustly aligned llm, arXiv preprint arXiv:2309.14348 (2023).

[13] M. Pehlke, M. Jansen, Llm driven processes to foster explainable ai, arXiv preprint arXiv:2511.07086 (2025).

[14] J. Chen, S. Saha, M. Bansal, Reconcile: Round-table conference improves reasoning via consensus among diverse llms, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 7066–7085.

[15] Y. Zhang, S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, F. Wei, Llm as a mastermind: A survey of strategic reasoning with large language models, arXiv preprint arXiv:2404.01230 (2024).

[16] J. Wang, A tutorial on llm reasoning: Relevant methods behind chatgpt o1, arXiv preprint arXiv:2502.10867 (2025).

[17] E. Bandara, R. Gore, S. Shetty, R. Mukkamala, C. Rhea, A. Yarlagadda, S. Kaushik, L. De Silva, A. Maznychenko, I. Sokolowska, et al., Standardization of neuromuscular reflex analysis–role of fine-tuned vision-language model consortium and openai gpt-oss reasoning llm enabled decision support system, arXiv preprint arXiv:2508.12473 (2025).

[18] E. Bandara, R. Gore, A. Yarlagadda, A. H. Clayton, P. Samuel, C. K. Rhea, S. Shetty, Standardization of psychiatric diagnoses–role of fine-tuned llm consortium and openai-gpt-oss reasoning llm enabled decision support system, arXiv preprint arXiv:2510.25588 (2025).

[19] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM computing surveys 55 (9) (2023) 1–33.

[20] E. Bandara, S. Shetty, R. Mukkamala, R. Gore, P. Foytik, S. H. Bouk, A. Rahman, X. Liang, N. W. Keong, K. De Zoysa, et al., Model context contracts-mcp-enabled framework to integrate llms with blockchain smart contracts, arXiv preprint arXiv:2510.19856 (2025).

[21] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[22] R. Gore, E. Bandara, S. Shetty, A. E. Musto, P. Rana, A. Valencia-Romero, C. Rhea, L. Tayebi, H. Richter, A. Yarlagadda, et al., Proof-of-tbi–fine-tuned vision language model consortium and openai-o3 reasoning llm-based medical diagnosis support system for mild traumatic brain injury (tbi) prediction, arXiv preprint arXiv:2504.18671 (2025).

[23] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al., gpt-oss-120b & gpt-oss-20b model card, arXiv preprint arXiv:2508.10925 (2025).

[24] G. Mondillo, M. Masino, S. Colosimo, A. Perrotta, V. Frattolillo, Evaluating ai reasoning models in pediatric medicine: A comparative analysis of o3-mini and o3-mini-high, medRxiv (2025) 2025–02.

[25] I. Shruti, A. Kumar, A. Seth, et al., Responsible generative ai: A comprehensive study to explain llms, in: 2024 International Conference on

Electrical, Computer and Energy Technologies (ICECET, IEEE, 2024, pp. 1–6.

[26] E. Bandara, S. Shetty, R. Mukkamala, R. Gore, P. Foytik, S. H. Bouk, A. Rahman, X. Liang, N. W. Keong, K. De Zoysa, et al., Model context contracts-mcp-enabled framework to integrate llms with blockchain smart contracts, arXiv preprint arXiv:2510.19856 (2025).

[27] E. Bandara, S. H. Bouk, S. Shetty, S. Roy, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, N. W. Keong, K. De Zoysa, Llama-recipe — fine-tuned meta's llama llm, pbom and nft enabled 5g network-slice orchestration and end-to-end supply-chain verification platform, in: 2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC), 2025, pp. 1–6. doi:10.1109/CCNC54725.2025.10976116.

[28] G. Marvin, N. Hellen, D. Jjingo, J. Nakatumba-Nabende, Prompt engineering in large language models, in: International Conference on Data Intelligence and Cognitive Informatics, Springer, 2023, pp. 387–402.

[29] B. Perak, S. Beliga, A. Meštrović, Incorporating dialect understanding into llm using rag and prompt engineering techniques for causal commonsense reasoning, in: Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), 2024, pp. 220–229.

[30] E. Bandara, S. Shetty, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, K. De Zoysa, N. W. Keong, Devsec-gpt — generative-ai (with custom-trained meta's llama2 llm), blockchain, nft and pbom enabled cloud native container vulnerability management and pipeline verification platform, in: 2024 IEEE Cloud Summit, 2024, pp. 28–35. doi:10.1109/Cloud-Summit61220.2024.00012.

[31] I. Shruti, A. Kumar, A. Seth, et al., Responsible generative ai: A comprehensive study to explain llms, in: 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET, IEEE, 2024, pp. 1–6.

[32] A. A. Najjar, H. I. Ashqar, O. Darwish, E. Hammad, Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llm-generated text, Information 16 (9) (2025) 767.

[33] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler, G. Green, Deep learning for rf signal classification in unknown and dynamic spectrum environments, in: 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), IEEE, 2019, pp. 1–10.

[34] A. A. Najjar, H. I. Ashqar, O. Darwish, E. Hammad, Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llm-generated text, Information 16 (9) (2025) 767.

[35] M. Becattini, R. Verdecchia, E. Vicario, Sallma: A software architecture for llm-based multi-agent systems.

[36] M. Imran, N. Almusharraf, Google gemini as a next generation ai educational tool: a review of emerging educational technology, Smart Learning Environments 11 (1) (2024) 22.

[37] J. M. Martinez-Thompson, K. A. Mazurek, C. Parra-Cantu, E. Naddaf, V. Gogineni, H. Botha, D. T. Jones, R. S. Laughlin, L. Barnard, N. P. Staff, Artificial intelligence models using f-wave responses predict amyotrophic lateral sclerosis, Brain (2025) awaf014.

[38] P. Šádek, E. Hrušková, S. Ostrỳ, J. Otáhal, Neurophysiological assessment of h-reflex alterations in compressive radiculopathy, Physiological Research 73 (3) (2024) 427.

[39] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, Y. Wei, T.-S. Chua, Data-efficient fine-tuning for llm-based recommendation, in: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, 2024, pp. 365–374.

[40] H. Samo, K. Ali, M. Memon, F. A. Abbasi, M. Y. Koondhar, K. Dahri, Fine-tuning mistral 7b large language model for python query response and code generation: A parameter efficient approach, VAWKUM Transactions on Computer Sciences 12 (1) (2024) 205–217.

[41] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[42] X. Y. Lee, S. Akatsuka, L. Vidyaratne, A. Kumar, A. Farahat, C. Gupta, Reliable decision-making for multi-agent llm systems, arXiv preprint arXiv:2406.04092 (2025).

[43] M. Gomes, A. D. Gonçalves, P. Pezarat-Correia, G. V. Mendonca, Changes in h-reflex, v-wave, and contractile properties of the plantar flexors following concurrent exercise sessions—the acute interference effect, Journal of Applied Physiology 138 (2) (2025) 327–341.

[44] F. Carrillo-Perez, O. E. Pecho, J. C. Morales, R. D. Paravina, A. Della Bona, R. Ghinea, R. Pulgar, M. d. M. Pérez, L. J. Herrera, Applications of artificial intelligence in dentistry: A comprehensive review, Journal of Esthetic and Restorative Dentistry 34 (1) (2022) 259–280.

[45] M. M. Meghil, P. Rajpurohit, M. E. Awad, J. McKee, L. A. Shahoumi, M. Ghaly, Artificial intelligence in dentistry, Dentistry Review 2 (1) (2022) 100009.

[46] N. Thalji, E. Aljarrah, M. H. Almomani, A. Raza, H. Migdady, L. Abualigah, Segmented x-ray image data for diagnosing dental periapical diseases using deep learning, Data in Brief 54 (2024) 110539.

[47] M. S. Scheeringa, Is factor analysis useful for revising diagnostic criteria for ptsd? a systematic review of five issues ten years after dsm-5, Journal of Psychiatric Research (2024).

[48] K. A. Smith, C. Blease, M. Faurholt-Jepsen, J. Firth, T. Van Daele, C. Moreno, P. Carlbring, U. W. Ebner-Priemer, N. Koutsouleris, H. Riper, et al., Digital mental health: challenges and next steps, BMJ Ment Health 26 (1) (2023).

[49] D. A. Regier, E. A. Kuhl, D. J. Kupfer, The dsm-5: Classification and criteria changes, World psychiatry 12 (2) (2013) 92–98.

[50] S. Sumathy, M. Revathy, R. Manikandan, Improving the state of materials in cybersecurity attack detection in 5g wireless systems using machine learning, Materials Today: Proceedings 81 (2023) 700–707.

[51] B. Hussain, Q. Du, B. Sun, Z. Han, Deep learning-based ddos-attack detection for cyber–physical system over 5g network, IEEE Transactions on Industrial Informatics 17 (2) (2020) 860–870.

[52] E. Bandara, P. Foytik, S. Shetty, R. Mukkamala, A. Rahman, X. Liang, N. W. Keong, K. D. Zoysa, Slicegpt – openai gpt-3.5 llm, blockchain and non-fungible token enabled intelligent 5g/6g network slice broker and marketplace, in: 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), 2024, pp. 439–445. doi:10.1109/CCNC51664.2024.10454701.

[53] Z. Chen, J. Li, P. Chen, Z. Li, K. Sun, Y. Luo, Q. Mao, M. Li, L. Xiao, D. Yang, et al., Harnessing multiple large language models: A survey on llm ensemble, arXiv preprint arXiv:2502.18036 (2025).

[54] J. Moncada-Ramirez, J.-L. Matez-Bandera, J. Gonzalez-Jimenez, J.-R. Ruiz-Sarmiento, Agentic workflows for improving large language model reasoning in robotic object-centered planning, Robotics 14 (3) (2025) 24.

[55] S. Gaurav, J. Heikkonen, J. Chaudhary, Governance-as-a-service: A multi-agent framework for ai system compliance and policy enforcement, arXiv preprint arXiv:2508.18765 (2025).