Comparative Analysis of LLM Performance on Social Safeguards and Risk Management Prompts

Introduction: Evaluating AI Capabilities in a High-Stakes Domain

This report presents a systematic evaluation of the performance of twelve different Large Language Models (LLMs) against a series of expert-level prompts concerning international social safeguards, specifically those outlined in the International Finance Corporation (IFC) Performance Standards. The prompts are designed to test the models' abilities to reason through complex, real-world scenarios in social risk management, a domain where nuance, ethical judgment, and technical precision are paramount.

To facilitate a direct and rigorous comparison, this analysis transposes the original model-by-model response structure into a question-centric format. The objective is to assess the depth, accuracy, and practical utility of current AI models in navigating the intricate requirements of social performance. By deconstructing responses to each prompt and evaluating them against a gold-standard benchmark, we can identify both the current capabilities and the systemic weaknesses of LLMs in this high-stakes field. The analysis begins with the foundational task of designing a Resettlement Action Plan.

---

1. Analysis for Question 1: RAP Design and Displacement Typology

This initial prompt is of critical strategic importance, serving as a powerful diagnostic of a model's foundational skills in social risk assessment. A correct response requires a nuanced understanding of IFC Performance Standard 5 (PS5), particularly the distinction between physical and economic displacement and the equal treatment of formal and informal land users. This section deconstructs and evaluates how each model handled this fundamental classification and planning task, revealing a clear performance hierarchy based on their ability to move beyond simple classification to design compliant mitigation measures.

1.1. Summary of Model Responses

Anthropic: Claude Sonnet 4.5

- Correctly classifies all three groups.
- Distinguishes between asset compensation and livelihood restoration.
- Provides detailed, group-specific measures.
- Recognizes rights of informal/customary users.

- Lists key PS5 principles applied.

Anthropic: Claude Opus 4.5

- Correctly classifies all groups.
- Clearly separates asset compensation from livelihood restoration.
- Cites specific PS5 categories for informal users.
- Prioritizes land-based restoration for land-based livelihoods.
- Includes procedural requirements like consultation.

OpenAI: GPT-5.1

- Correctly classifies all groups.
- Lists measures under "Compensation" and "Livelihood Restoration."
- Provides generic, high-level support measures.
- Mentions community support and business support.
- Includes resettlement support as a third category.

OpenAI: GPT-5 (Thinking)

- Correctly classifies displacement types.
- Structures measures under "Compensation" and "Livelihood."
- Recognizes customary tenure for informal farmers.
- Includes relocation assistance and transitional support.
- Provides a summary table format.

Google: Gemini 3 Pro Preview

- Correctly classifies all three groups.
- Separates asset compensation from livelihood restoration.
- Recognizes customary rights for informal farmers.
- Provides detailed and relevant measures for each group.
- Lists overarching PS5 principles applied.

Perplexity: Sonar Reasoning Pro

- Incorrectly classifies informal farmers as both physically and economically displaced.
- Provides measures for each group, including consultation.
- Recognizes customary land claims under PS5.
- Details replacement housing and agricultural restoration.

Mistral: Mistral Large 3 2512

- Correctly classifies all three groups.
- Separates asset compensation from relocation/livelihood support.
- Mentions replacement costs and fair market value.
- Includes support for transition and business recovery.

Meta: Llama 4 Maverick

- Correctly classifies all three groups.
- Organizes measures under asset compensation and livelihood restoration.
- Recognizes customary rights and compensation for informal farmers.
- Includes transitional support allowances for all groups.

Meta: Llama 4 Scout

- Correctly classifies all groups.
- Clearly distinguishes asset compensation and livelihood restoration.
- Cites specific PS5 paragraphs and guidance notes.
- Emphasizes land-based solutions and tenure formalization.

XAI: Grok 4.1 Fast

- Correctly classifies all three groups.
- Separates asset compensation from livelihood/relocation support.
- Recognizes customary land use rights for informal farmers.
- Includes transitional support and relocation assistance.

Amazon: Nova Premier 1.0

- Correctly classifies all three groups.
- Separates asset compensation from resettlement/livelihood support.
- Provides generic support measures for each category.
- Mentions community development programs.

NVIDIA: Nemotron Nano 12B 2 VL

- Correctly classifies all three groups.
- Separates asset compensation from livelihood restoration.
- Includes relocation assistance for physically displaced group.
- Provides relevant support measures for each group.

1.2. Comparative Analysis Against the Gold Standard

Gold Standard Expectations

A gold-standard response must demonstrate a clear understanding of IFC Performance Standard 5 by:

- Correctly classifying each group's displacement type (physical, economic, or both).
- Applying PS5's principles equally to both formal and informal/customary land users.
- Clearly distinguishing between compensation for lost assets (e.g., replacement housing, cash for crops) and measures for livelihood restoration (e.g., agricultural support, business grants, transitional allowances).
- Proposing measures that are specific to the type of livelihood lost (e.g., prioritizing land-for-land strategies for farmers).

Anthropic: Claude Sonnet 4.5

- ***Strengths:***
    - Correctly classified all three groups and provided a clear rationale.
    - Explicitly and correctly distinguished between asset compensation and livelihood restoration measures.
    - Offered detailed, practical, and PS5-aligned measures for each group, including recognizing the rights of informal users.
- ***Weaknesses/Gaps:***

    - None of significance. The response was comprehensive and accurate.
- ***Serious Errors:***
    - None.

Anthropic: Claude Opus 4.5

- ***Strengths:***
    - Perfectly classified all groups and provided a robust rationale.
    - Showed superior domain knowledge by citing specific PS5 categories and prioritizing land-based restoration for land-dependent livelihoods.
    - Included procedural requirements (consultation, vulnerability assessment), demonstrating a deeper understanding of the implementation process.
- ***Weaknesses/Gaps:***
    - None of significance. The response was exceptionally strong.
- ***Serious Errors:***
    - None.

OpenAI: GPT-5.1

- ***Strengths:***
    - Correctly classified the displacement type for all three groups.
- ***Weaknesses/Gaps:***
    - Demonstrated a conceptual weakness in distinguishing asset compensation from livelihood restoration, a foundational error in PS5 application. It mixed concepts like "compensation for crops" with "training programs" under the same heading.
    - Measures were generic and lacked specificity (e.g., "financial support" instead of business interruption allowances or grants).
    - Did not explicitly address the principle of "replacement cost" or the preference for land-based strategies.
- ***Serious Errors:***
    - None, but the response was superficial.

OpenAI: GPT-5 (Thinking)

- ***Strengths:***
    - Correctly classified all groups.
    - Used a clear structure that separated asset compensation from livelihood restoration.
    - Correctly noted the need to recognize customary tenure for informal farmers.
- ***Weaknesses/Gaps:***
    - The proposed measures were less detailed than top-tier models, lacking specifics like "full replacement cost" or multi-season agricultural support.
- ***Serious Errors:***
    - None.

Google: Gemini 3 Pro Preview

- ***Strengths:***
    - Accurately classified all three groups.
    - Maintained a clear and correct distinction between asset compensation and livelihood restoration.
    - Provided specific and relevant measures, such as recognizing customary rights and providing transitional allowances.
- ***Weaknesses/Gaps:***
    - None of significance. The response was solid and well-aligned with the gold standard.
- ***Serious Errors:***
    - None.

Perplexity: Sonar Reasoning Pro

- ***Strengths:***
    - Correctly classified two of the three groups.
    - Recognized that informal users have rights under PS5.

- ***Weaknesses/Gaps:***
    - Measures provided were somewhat generic and did not always clearly separate asset compensation from livelihood support.
- ***Serious Errors:***
    - Incorrectly classified the informal farmers (Group B) as *both* physically and economically displaced, a fundamental error in applying PS5 definitions.

Mistral: Mistral Large 3 2512

- ***Strengths:***
    - Correctly classified all three groups.
    - Distinguished between asset compensation and livelihood support.
- ***Weaknesses/Gaps:***
    - Demonstrated a conceptual weakness in distinguishing asset compensation from livelihood restoration; the separation was less clear than in top-tier responses.
    - Measures were general and lacked technical detail (e.g., did not specify "replacement cost").
- ***Serious Errors:***
    - None.

Meta: Llama 4 Maverick

- ***Strengths:***
    - Correctly classified all three groups.
    - Used a clear structure separating asset compensation and livelihood restoration.
    - Correctly identified the need for transitional support and recognition of customary rights.
- ***Weaknesses/Gaps:***
    - The level of detail in the proposed measures was slightly less than top-performing models.

- ***Serious Errors:***
    - None.

Meta: Llama 4 Scout

- ***Strengths:***
    - Perfectly classified all groups.
    - Demonstrated superior technical knowledge by referencing specific PS5 paragraphs and guidance notes.
    - Correctly emphasized the preference for land-based solutions for land-based livelihoods.
- ***Weaknesses/Gaps:***
    - None of significance. Performance was at a top-tier level.
- ***Serious Errors:***
    - None.

XAI: Grok 4.1 Fast

- ***Strengths:***
    - Correctly classified all groups.
    - Separated asset compensation from livelihood restoration.
- ***Weaknesses/Gaps:***
    - The proposed measures were very high-level and lacked practical detail.
    - The distinction between compensation and restoration, while structurally present, was conceptually weak.
- ***Serious Errors:***
    - None.

Amazon: Nova Premier 1.0

- ***Strengths:***
    - Correctly classified all groups.
- ***Weaknesses/Gaps:***

    - The structure conflated resettlement support with livelihood restoration, indicating a lack of conceptual clarity.
    - Proposed measures were overly generic, such as "training programs" without specifying whether this meant vocational skills for alternative livelihoods or agricultural extension for restored farming—a critical distinction for a compliant LRP.
    - Failed to distinguish between asset compensation and livelihood restoration in a meaningful way.
- ***Serious Errors:***
    - None, but the response was superficial and poorly structured.

NVIDIA: Nemotron Nano 12B 2 VL

- ***Strengths:***
    - Correctly classified all groups.
    - Correctly structured the response to separate asset compensation from livelihood restoration.
- ***Weaknesses/Gaps:***
    - Measures were very general and lacked the technical depth of PS5 requirements (e.g., no mention of "replacement cost").
- ***Serious Errors:***
    - None.

1.3. Synthesis of Cross-Model Patterns

This foundational prompt served as a powerful diagnostic, revealing a clear performance hierarchy. While nearly all models could perform the basic classification task, their ability to apply the deeper principles of social risk management varied significantly.

- **Main Similarities:** Most models (with the exception of Perplexity) correctly identified the displacement classifications for the three groups, demonstrating a basic grasp of the distinction between physical and economic displacement. The majority also recognized that informal land users are entitled to protection under PS5.

- **Main Differences:** The primary differentiator was the ability to distinguish conceptually and structurally between **asset compensation** and **livelihood restoration**. Top-tier models (Claude Opus, Claude Sonnet, Llama 4 Scout, Gemini) created a clear separation, proposing distinct measures for replacing lost assets versus providing support to rebuild income streams. Weaker models (GPT-5.1, Amazon Nova) tended to conflate these, listing a mix of measures under generic headings.
- **Common Systematic Weaknesses:** The most frequent failing was superficiality. Many models provided generic, list-based answers ("provide training," "offer financial support") without the specificity required for a compliant Resettlement Action Plan. They often missed key PS5 principles like "full replacement cost" and the strong preference for "land-for-land" restoration strategies for those with land-based livelihoods. This suggests the models are trained on texts that use these terms interchangeably, lacking the structured, principles-based logic of the Performance Standards themselves.

These findings highlight a gap between pattern recognition (correctly labeling displacement types) and applied reasoning (designing compliant, practical mitigation plans). The analysis now proceeds to the next prompt, which tests the models' abilities to navigate the complexities of grievance management and Indigenous Peoples' rights.

---

2. Analysis for Question 2: Grievances, FPIC, and Indigenous Governance

This prompt tests a model's ability to navigate the highly nuanced relationship between collective Indigenous rights and the protection of marginalized sub-groups. A successful answer must balance respect for community governance structures and the principle of Free, Prior, and Informed Consent (FPIC) with the need to create safe, accessible channels for women, youth, and other vulnerable individuals who may be excluded from traditional decision-making. This section evaluates the sophistication of each model's proposed grievance and corrective action framework in this complex social and political landscape.

2.1. Summary of Model Responses

Anthropic: Claude Sonnet 4.5

- Proposes a multi-tiered grievance mechanism.
- Includes specific, practical measures for women, youth, and satellite hamlets.
- Outlines a detailed corrective action protocol with clear timelines.

- Addresses ongoing FPIC and consent withdrawal.
- Includes safeguards like a no-retaliation policy.

Anthropic: Claude Opus 4.5

- Frames the issue as an "FPIC paradox" requiring ongoing engagement.
- Designs a tiered grievance system with sub-group-specific entry points.
- Links grievances to triggers for new/resumed FPIC.
- Provides specific corrective action examples tied to PS standards.
- Emphasizes capacity building and anti-retaliation protocols.

OpenAI: GPT-5.1

- Proposes a Community Grievance Committee.
- Suggests safe spaces and anonymity for marginalized groups.
- Mentions ongoing FPIC and RAP/LRP adjustments.
- Recommends capacity building and inclusive representation.

OpenAI: GPT-5 (Thinking)

- Proposes a tiered grievance structure.
- Includes multiple entry points and safeguards like non-retaliation.
- Links grievance resolution to ongoing FPIC.
- Suggests joint reviews and participatory monitoring.
- Outlines remedial measures like adjusting LRPs.

Google: Gemini 3 Pro Preview

- Designs a multi-tiered grievance mechanism.
- Includes safe access channels like anonymous submissions and trusted intermediaries.
- Outlines a phased corrective action process.
- Links grievances to supplementary FPIC and RAP/LRP amendments.
- Emphasizes capacity building and transparent record-keeping.

Perplexity: Sonar Reasoning Pro

- Proposes a "Nested Participation" model working within Indigenous governance.
- Designs a tiered system with sub-group listening posts.
- Distinguishes between FPIC process failures and implementation gaps.
- Links corrective actions to ongoing FPIC requirements.
- Includes community-led monitoring and lender oversight.

Mistral: Mistral Large 3 2512

- Suggests a culturally appropriate grievance mechanism.
- Recommends targeted consultations with sub-groups.
- Proposes reviewing and adapting the RAP/LRP.
- Mentions ongoing FPIC and capacity building.

Meta: Llama 4 Maverick

- Proposes a multi-tiered grievance structure.
- Includes safe access channels like anonymous boxes and focal points.
- Outlines a phased corrective action process.
- Connects grievances to a supplementary FPIC process if needed.
- Lists specific KPIs for monitoring performance.

Meta: Llama 4 Scout

- Frames the problem within the context of PS7's recognition of internal community diversity.
- Designs a "Nested Grievance Pathway" anchored in customary governance.
- Proposes an independent ombudsperson or cultural mediator.
- Links corrective actions to ongoing FPIC for any RAP/LRP revisions.
- Provides specific examples of tailored livelihood support.

XAI: Grok 4.1 Fast

- Embeds the primary grievance mechanism in existing Indigenous governance.

- Suggests a joint grievance committee with sub-group representation.
- Includes dedicated access points for women, youth, and hamlets.
- Links grievances to re-engagement for ongoing FPIC on corrective actions.
- Mentions adaptive management and monitoring.

Amazon: Nova Premier 1.0

- Proposes a culturally appropriate grievance mechanism with a committee.
- Suggests reviewing the FPIC process through inclusive consultation.
- Recommends assessing and adjusting the RAP/LRP.
- Includes strengthening governance and capacity building.

NVIDIA: Nemotron Nano 12B 2 VL

- Suggests establishing accessible grievance channels.
- Proposes independent investigation and facilitated dialogue.
- Links grievances to ongoing FPIC where relevant.
- Outlines a corrective action plan with participatory monitoring.

2.2. Comparative Analysis Against the Gold Standard

Gold Standard Expectations

A gold-standard response must balance respect for collective Indigenous institutions with robust protections for marginalized sub-groups. Key requirements include:

- Working with and through recognized Indigenous governance structures, not bypassing them.
- Designing practical, safe, and confidential channels for women, youth, and other sub-groups to raise concerns without fear of retaliation.
- Recognizing that grievances about implementation failures may trigger the need for renewed consultation or consent under the principle of ongoing FPIC.
- Proposing concrete corrective actions that are integrated into project monitoring (RAP/LRP) and address the specific needs of sub-groups.

Anthropic: Claude Sonnet 4.5

- ***Strengths:***
    - Proposed a highly detailed and practical multi-tiered system.
    - Provided excellent, concrete measures for including women, youth, and remote hamlets.
    - Correctly linked grievances to a potential "supplementary FPIC process" and outlined a robust corrective action protocol.
- ***Weaknesses/Gaps:***
    - The proposed structure was very prescriptive (e.g., specifying percentages for committee representation), which may not be culturally appropriate in all contexts.
- ***Serious Errors:***
    - None.

Anthropic: Claude Opus 4.5

- ***Strengths:***
    - Demonstrated superior conceptual understanding by framing the issue as an "FPIC paradox" and a trigger for "Phase 2 corrective engagement."
    - Designed an excellent tiered system with safe entry points that reinforce, rather than replace, traditional governance.
    - Explicitly and correctly treated material grievances as a trigger for "new/resumed FPIC."
- ***Weaknesses/Gaps:***
    - None of significance. The response was outstanding.
- ***Serious Errors:***
    - None.

OpenAI: GPT-5.1

- ***Strengths:***
    - Identified the core requirements: a committee, safe spaces, and ongoing FPIC.

- ***Weaknesses/Gaps:***
    - The response was extremely high-level and lacked any operational detail.
    - Failed to explain *how* the grievance committee would interact with traditional governance or *how* grievances would trigger ongoing FPIC.
    - Measures for sub-groups were generic ("women-only meetings") without practical safeguards.
- ***Serious Errors:***
    - None, but the answer was too superficial to be useful.

OpenAI: GPT-5 (Thinking)

- ***Strengths:***
    - Proposed a well-structured tiered mechanism.
    - Included important safeguards like non-retaliation guarantees.
    - Correctly identified the need for ongoing FPIC for revised measures.
- ***Weaknesses/Gaps:***
    - Lacked the depth and practical detail of the top-tier models, particularly in the design of sub-group consultations.
- ***Serious Errors:***
    - None.

Google: Gemini 3 Pro Preview

- ***Strengths:***
    - Proposed a well-designed multi-tiered mechanism.
    - Included practical and safe channels for sub-groups, such as trusted intermediaries.
    - Correctly linked grievances to a "supplementary FPIC process" and outlined a detailed corrective action plan.
- ***Weaknesses/Gaps:***
    - None of significance. A very strong and practical response.

- ***Serious Errors:***
    - None.

Perplexity: Sonar Reasoning Pro

- ***Strengths:***
    - Excellent conceptual framing of "Nested Participation" that works within existing governance.
    - Astutely distinguished between grievances stemming from FPIC process failures versus implementation gaps.
    - Correctly linked plan modifications to ongoing FPIC requirements.
- ***Weaknesses/Gaps:***
    - None of significance. A top-tier response with strong analytical depth.
- ***Serious Errors:***
    - None.

Mistral: Mistral Large 3 2512

- ***Strengths:***
    - Identified the key components of a solution: a grievance mechanism, targeted consultations, and RAP/LRP adaptation.
- ***Weaknesses/Gaps:***
    - The response was very general and lacked the structural detail or operational mechanics of how the system would work.
    - It mentioned "ongoing FPIC" but did not explain the conditions under which it would be triggered, revealing a superficial understanding.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Maverick

- ***Strengths:***
    - Proposed a well-structured, multi-level grievance mechanism.

    - Included practical safe channels like anonymous boxes and mobile teams.
    - Correctly identified that material changes could trigger a "supplementary FPIC process."
- ***Weaknesses/Gaps:***
    - The conceptual link between respecting collective governance and ensuring individual voice was not as clearly articulated as in top-tier models.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Scout

- ***Strengths:***
    - Excellent framing of the issue, recognizing that communities are not homogenous.
    - Proposed a sophisticated "Nested Grievance Pathway" and the use of an independent ombudsperson.
    - Correctly and explicitly linked any revisions to the RAP/LRP to an ongoing FPIC process for those specific changes.
- ***Weaknesses/Gaps:***
    - None of significance. An expert-level response.
- ***Serious Errors:***
    - None.

XAI: Grok 4.1 Fast

- ***Strengths:***
    - Correctly proposed embedding the mechanism within existing Indigenous governance.
    - Identified the need for dedicated access points for sub-groups.
- ***Weaknesses/Gaps:***
    - The response was coherent but lacked detail on the operational aspects of the mechanism.

    - The connection between grievances and ongoing FPIC was mentioned but not deeply explained.

- ***Serious Errors:***
    - None.

Amazon: Nova Premier 1.0

- ***Strengths:***
    - Recognized the need for a culturally appropriate mechanism and sub-group consultation.
- ***Weaknesses/Gaps:***
    - The entire response was a high-level list of headings with no detail, rendering it inactionable.
    - It failed to provide a concrete design for the grievance mechanism or explain the link to ongoing FPIC.
- ***Serious Errors:***
    - None, but the response was of low quality.

NVIDIA: Nemotron Nano 12B 2 VL

- ***Strengths:***
    - Identified the key steps: establish channels, investigate, and create a corrective action plan.
- ***Weaknesses/Gaps:***
    - The response was very generic and lacked specific design features for the grievance mechanism.
    - It did not adequately explain how to balance collective governance with sub-group rights.
- ***Serious Errors:***
    - None.

2.3. Synthesis of Cross-Model Patterns

The responses to this prompt on Indigenous governance and grievances clearly separated the models into distinct performance tiers. The challenge required balancing two competing principles—collective rights and individual protection—and only the most sophisticated models succeeded.

- **Main Similarities:** Most models proposed some form of a multi-tiered or multi-channel grievance mechanism. They also generally recognized the need for special provisions for marginalized groups, such as women-only forums or anonymous reporting channels.
- **Main Differences:** The key differentiator was the sophistication of the integration with Indigenous governance. Weaker models proposed generic grievance committees that seemed to bypass or ignore traditional structures. In contrast, top-tier models (Claude Opus, Llama 4 Scout, Perplexity) designed "nested" systems that worked *through* customary institutions while ensuring sub-group voices were channeled effectively to that level. Furthermore, the strongest models correctly and explicitly identified that significant grievances are not just complaints to be resolved, but potential triggers for a **renewed or supplementary FPIC process** for the modified project components.
- **Common Systematic Weaknesses:** A frequent conceptual weakness was the failure to explain the link between grievances and FPIC. Many models mentioned "ongoing FPIC" as a buzzword but could not articulate the conditions under which a grievance would necessitate re-engagement and renewed consent. This reveals a shallow understanding of FPIC as an ongoing process rather than a one-time event, likely stemming from training data that overemphasizes the initial consent stage.

This complex prompt effectively tested for deeper conceptual understanding beyond simple information retrieval. The next section examines an even more complex integration challenge, combining multiple IFC standards and social risks.

---

3. Analysis for Question 3: Integrated Social Risk Management (PS1/5/7, FPIC, Influx)

This highly complex prompt was designed to test a model's ability to engage in systemic thinking. It required the integration of multiple IFC Performance Standards (PS1, PS5, and PS7) and a set of interconnected social risks—resettlement, Indigenous rights, and project-induced in-migration (influx)—into a single, coherent management framework. The evaluation in this section assesses which models could articulate a genuinely integrated system versus those that offered a siloed, standard-by-standard list of actions.

3.1. Summary of Model Responses

Anthropic: Claude Sonnet 4.5

- Positions PS1's ESMS as the foundational framework.
- Explains how FPIC (PS7) co-designs the RAP/LRP (PS5).
- Identifies three influx risk pathways: competition, cultural disruption, and economic distortion.
- Proposes a multi-channel grievance architecture for different user groups.
- Links monitoring KPIs directly to PS5 and PS7 compliance.

Anthropic: Claude Opus 4.5

- Describes a "Three-Standard Cascade" (PS1 → PS5 → PS7).
- Details how FPIC must shape RAP/LRP design from the pre-assessment phase.
- Outlines three influx pathways: resource competition, gendered violence, and environmental degradation.
- Designs a four-tier grievance structure with culturally embedded entry points.
- Provides detailed monitoring metrics for PS5, PS7, and influx.

OpenAI: GPT-5.1

- Briefly explains the roles of PS1, PS5, and PS7.
- Identifies three influx risks: strained resources, cultural disruption, and economic displacement.
- Proposes a grievance mechanism with community committees.
- Suggests monitoring resettlement progress, influx impacts, and grievance trends.

OpenAI: GPT-5 (Thinking)

- Positions PS1's ESMS as the overarching system.
- Explains the link between FPIC and RAP/LRP co-development.
- Lists influx risks: land pressure, social conflict, and health/safety risks.
- Proposes a tiered, culturally appropriate grievance mechanism.
- Connects monitoring data to a final completion audit.

Google: Gemini 3 Pro Preview

- Uses PS1's ESMS as the integrating framework.
- Explains how FPIC co-designs the RAP/LRP.
- Identifies influx risks: resource competition, pressure on sites, and social disruption.
- Proposes a multi-channel grievance system.
- Links monitoring indicators to a central database for audit readiness.

Perplexity: Sonar Reasoning Pro

- Presents a "Three-Standard Logic Chain" for integration.
- Details how FPIC shapes the RAP/LRP from assessment to implementation.
- Outlines three influx pathways, including secondary displacement and social fragmentation.
- Designs a single grievance mechanism with differentiated, culturally appropriate pathways.
- Integrates monitoring of PS5, PS7, and influx into a single system.

Mistral: Mistral Large 3 2512

- Briefly describes the roles of PS1, PS5, and PS7.
- Identifies three influx risks: strain on resources, social tensions, and economic exclusion.
- Mentions a culturally appropriate grievance mechanism.
- Suggests monitoring key indicators for resettlement, influx, and grievances.

Meta: Llama 4 Maverick

- Uses PS1's ESMS as the integrating framework.
- Explains how FPIC shapes RAP/LRP co-design.
- Identifies three influx risk pathways.
- Proposes a three-channel grievance system for different groups.
- Provides specific KPIs for PS5 and PS7 compliance monitoring.

Meta: Llama 4 Scout

- Describes the hierarchical relationship (PS1 contains PS5/PS7).
- Explains the critical sequencing of FPIC *before* RAP/LRP finalization.
- Identifies three detailed influx risk pathways and their effect on RAP/LRP objectives.
- Designs a sophisticated, nested grievance mechanism.
- Provides detailed monitoring metrics for PS5, PS7, and influx.

XAI: Grok 4.1 Fast

- Explains the roles of PS1, PS5, and PS7.
- Links FPIC to the design of the RAP/LRP for the Indigenous community.
- Identifies three influx risk pathways and how they undermine RAP/LRP objectives.
- Proposes a multi-tiered grievance mechanism.
- Connects monitoring to a completion audit.

Amazon: Nova Premier 1.0

- Briefly outlines the function of each standard.
- Identifies three influx risks and suggests generic mitigation measures.
- Describes a grievance mechanism with community committees.
- Suggests monitoring resettlement, influx, and grievances.

NVIDIA: Nemotron Nano 12B 2 VL

- Describes the function of each standard separately.
- Lists three influx risks and their potential impact on RAP/LRP.
- Proposes a grievance mechanism and monitoring system.
- Response shows limited integration between the different components.

3.2. Comparative Analysis Against the Gold Standard

Gold Standard Expectations

A gold-standard response must demonstrate true systemic thinking, not just a list of separate actions for each standard. Key requirements include:

- Correctly articulating the hierarchical relationship: PS1 provides the overarching ESMS framework into which PS5 and PS7 requirements are integrated.
- Explaining that FPIC (PS7) is a prerequisite that fundamentally shapes the design of the RAP/LRP (PS5) for the Indigenous community, rather than an add-on.
- Identifying specific risk pathways where project-induced influx directly undermines RAP/LRP objectives.
- Designing a single, unified grievance and monitoring system that captures resettlement, Indigenous rights, and influx-related issues, providing integrated data for adaptive management.

Anthropic: Claude Sonnet 4.5

- ***Strengths:***
    - Correctly positioned PS1 as the foundational ESMS.
    - Clearly explained how PS7's FPIC process informs PS5's RAP/LRP design.
    - Provided a strong analysis of influx risks and proposed a well-designed, integrated grievance and monitoring system.
- ***Weaknesses/Gaps:***
    - None of significance. A very comprehensive and well-integrated response.
- ***Serious Errors:***
    - None.

Anthropic: Claude Opus 4.5

- ***Strengths:***
    - Excellent conceptual framing of the "Three-Standard Cascade" and the critical sequencing of FPIC.
    - Provided a deeply nuanced analysis of influx risks, including gendered violence and environmental degradation pathways.
    - Designed a sophisticated, culturally embedded grievance mechanism and a detailed, integrated monitoring framework.
- ***Weaknesses/Gaps:***
    - None. This response represents an expert-level synthesis.

- ***Serious Errors:***
    - None.

OpenAI: GPT-5.1

- ***Strengths:***
    - Identified the basic roles of the three standards and named relevant influx risks.
- ***Weaknesses/Gaps:***
    - The grievance and monitoring proposals were generic and not explicitly linked into a single system.
- ***Serious Errors:***
    - Demonstrated a critical failure in systemic thinking by treating the standards and risks as silos. This siloed approach is the primary error the prompt was designed to detect.

OpenAI: GPT-5 (Thinking)

- ***Strengths:***
    - Correctly identified PS1's ESMS as the overarching framework and explained the FPIC-RAP/LRP link.
    - Proposed an integrated social management framework.
- ***Weaknesses/Gaps:***
    - While correctly identifying PS1 as the overarching framework, its analysis of influx risk remained disconnected from the RAP/LRP, unlike top models which explicitly linked influx-driven resource competition to the potential failure of livelihood restoration targets.
- ***Serious Errors:***
    - None.

Google: Gemini 3 Pro Preview

- ***Strengths:***
    - Correctly used PS1's ESMS as the integrating framework.

    - Clearly explained how FPIC shapes the RAP/LRP design.
    - Provided a strong analysis of influx risks and a well-integrated grievance and monitoring system.
- ***Weaknesses/Gaps:***
    - None of significance. A high-quality, integrated response.
- ***Serious Errors:***
    - None.

Perplexity: Sonar Reasoning Pro

- ***Strengths:***
    - Excellent "Three-Standard Logic Chain" concept that correctly shows the integration.
    - Provided a strong analysis of how FPIC must shape the RAP/LRP.
    - Offered a sophisticated, unified grievance mechanism with differentiated pathways.
- ***Weaknesses/Gaps:***
    - None of significance. The response was analytical and well-structured.
- ***Serious Errors:***
    - None.

Mistral: Mistral Large 3 2512

- ***Strengths:***
    - Identified the main components required in the answer.
- ***Weaknesses/Gaps:***
    - The response described the components of a management system but failed to articulate the critical *integrating logic* between them, resulting in a checklist rather than a coherent framework.
- ***Serious Errors:***
    - The siloed nature of the response represents a failure to address the core integration challenge of the prompt.

Meta: Llama 4 Maverick

- ***Strengths:***
    - Correctly identified PS1's ESMS as the integrating framework.
    - Explained the FPIC-RAP/LRP link and identified relevant influx risks.
    - Proposed an integrated system with specific KPIs for monitoring.
- ***Weaknesses/Gaps:***
    - The overall integration was articulated less forcefully than in the top-performing models.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Scout

- ***Strengths:***
    - Accurately described the hierarchical relationship of the standards and the critical sequencing of FPIC.
    - Provided an exceptionally detailed analysis of influx risk pathways and their specific impacts on RAP/LRP goals.
    - Designed a sophisticated, integrated grievance and monitoring system.
- ***Weaknesses/Gaps:***
    - None. An expert-level response demonstrating deep systemic thinking.
- ***Serious Errors:***
    - None.

XAI: Grok 4.1 Fast

- ***Strengths:***
    - Correctly described the roles of the standards and the link between FPIC and the RAP/LRP.
    - Identified relevant influx risks and proposed an appropriate grievance/monitoring structure.

- ***Weaknesses/Gaps:***
    - While the components were correct, the explanation of their integration into a single framework was less clear and detailed than in top models.
- ***Serious Errors:***
    - None.

Amazon: Nova Premier 1.0

- ***Strengths:***
    - Listed the correct elements that needed to be addressed.
- ***Weaknesses/Gaps:***
    - The response was a simple list of disconnected points with no analytical substance.
- ***Serious Errors:***
    - A complete failure of systemic thinking, directly contrary to the prompt's instructions. The model failed to explain the integration of the standards or how the different risk management components would function as a single system.

NVIDIA: Nemotron Nano 12B 2 VL

- ***Strengths:***
    - Identified the key topics required by the prompt.
- ***Weaknesses/Gaps:***
    - The connection between influx, the RAP/LRP, and the grievance mechanism was not established.
- ***Serious Errors:***
    - Failed to design an integrated framework, which was the central task of the prompt. The response presented information in silos, with no explanation of how PS1, PS5, and PS7 integrate.

3.3. Synthesis of Cross-Model Patterns

The integration prompt proved to be a powerful differentiator, clearly separating models capable of systemic reasoning from those limited to information retrieval. The ability to

synthesize multiple complex standards into a single operational framework is a hallmark of expert-level performance.

- **Main Similarities:** Nearly all models were able to correctly identify the basic purpose of each IFC Performance Standard (PS1 for management, PS5 for resettlement, PS7 for Indigenous Peoples) and list relevant social risks associated with project-induced in-migration.
- **Main Differences:** The starkest difference lay in demonstrating **integration**. Top-tier models (Claude Opus, Claude Sonnet, Llama 4 Scout, Perplexity) successfully articulated the hierarchical relationship, showing how PS1's ESMS provides the "container" for PS5 and PS7 implementation. They also excelled at explaining the causal chain where influx risks directly undermine specific RAP/LRP objectives. Weaker models simply listed the functions of each standard and the risks of influx as separate, disconnected bullet points.
- **Common Systematic Weaknesses:** The most prevalent failure was treating the prompt as a command to generate three separate mini-essays on PS1/5/7, influx, and grievances. This siloed approach misses the entire point of integrated social risk management, where the components are dynamically linked. Many models failed to explain that influx isn't just another risk, but one that can fatally compromise the outcomes of resettlement and livelihood restoration plans. This failure in systemic thinking on the integration prompt is a more complex manifestation of the conceptual conflation first observed in the Question 1 analysis, suggesting a core weakness in reasoning across multiple, interrelated concepts.

This analysis shows that while many LLMs can define complex concepts, only a select few can synthesize them into a coherent, operational strategy. The report now turns to a more practical, operational challenge: designing a grievance triage system.

---

4. Analysis for Question 4: Mining Grievance Intake and Triage

This prompt tests the models on a practical, operational challenge: designing a robust grievance triage system for a mining context that is both efficient and ethically sound. The evaluation focuses on the ability to move beyond generic categories and design a risk-tiered rubric, account for real-world data issues like transcript errors, and, most critically, integrate essential "human-in-the-loop" safeguards for high-stakes issues.

4.1. Summary of Model Responses

Anthropic: Claude Sonnet 4.5

- Proposes four priority levels from "Critical" to "Routine."
- Uses keyword triggers for initial categorization.
- Includes a "Smart Triage Algorithm" with rules for safety overrides and uncertainty.
- Details three specific human-in-the-loop checkpoints.
- Adds protocols for anonymous reports and error mitigation.

Anthropic: Claude Opus 4.5

- Uses a layered architecture for risk assessment.
- Defines high-priority triggers for immediate human review.
- Includes context-based risk flagging (demographics, vulnerability).
- Identifies transcript anomalies (pauses, tone) as risk flags.
- Outlines a detailed human-in-the-loop verification and handling process.

OpenAI: GPT-5.1

- Proposes four tiers from "Critical" to "Low."
- Uses keyword matching for initial classification.
- Includes rules for handling transcript errors and urgency.
- Describes a human-in-the-loop validation step for high-tier grievances.
- Outlines an escalation workflow for each tier.

OpenAI: GPT-5 (Thinking)

- Defines six classification categories from "Health & Safety" to "Other."
- Uses keyword and context classifiers for automated triage.
- Includes confidence scoring to flag transcripts for human review.
- Describes a multi-step human-in-the-loop workflow.
- Adds controls to prevent mishandling, such as dual review.

Google: Gemini 3 Pro Preview

- Establishes four priority categories (P1-P4).
- Uses a multilingual keyword detection matrix.

- Includes an algorithm for confidence scoring.
- Details mandatory human review triggers.
- Adds a multi-layer safety net for sensitive reports.

Perplexity: Sonar Reasoning Pro

- Uses a three-layer architecture: automated triage, human review, and escalation.
- Defines four categories from "Critical" to "Low."
- Includes decision rules based on keywords and confidence scoring.
- Outlines specific human-in-the-loop checkpoints.
- Adds a detailed escalation protocol with safeguards.

Mistral: Mistral Large 3 2512

- Proposes four priority tiers for grievances.
- Includes decision rules based on keyword flagging.
- Describes error detection and human review for ambiguous transcripts.
- Outlines escalation pathways for each tier.
- Mentions a monitoring and feedback loop.

Meta: Llama 4 Maverick

- Uses a three-layer processing model.
- Defines four priority categories with keyword triggers.
- Includes a confidence scoring system to manage transcript errors.
- Details human-in-the-loop checkpoints and sensitive report protections.
- Adds quality assurance metrics and a failsafe mode.

Meta: Llama 4 Scout

- Proposes a two-phase system: automated categorization and human verification.
- Defines six distinct grievance categories.
- Includes confidence scoring and trend markers in its decision rules.
- Details human-in-the-loop steps for safety, rights, and environmental issues.

- Adds a quality assurance process for continuous improvement.

XAI: Grok 4.1 Fast

- Defines four categories from "Critical" to "General."
- Uses keyword matching and sentiment analysis for initial triage.
- Includes a confidence scoring step to flag errors for human review.
- Details a human review and final triage phase.
- Outlines an escalation workflow for each priority level.

Amazon: Nova Premier 1.0

- Proposes six grievance categories with priority levels.
- Uses keyword matching for automatic analysis.
- Includes error handling for unclear transcripts.
- Describes human-in-the-loop steps for review and escalation.
- Outlines an escalation pathway for each priority level.

NVIDIA: Nemotron Nano 12B 2 VL

- Defines four categories from "Critical" to "Information Request."
- Uses keyword matching and sentiment analysis.
- Includes human review for high-priority or low-confidence transcripts.
- Describes escalation rules for each priority level.
- Mentions a feedback loop and audit trail.

4.2. Comparative Analysis Against the Gold Standard

Gold Standard Expectations

A gold-standard response for designing a grievance triage rubric must go beyond simple categorization and demonstrate real-world operational thinking. Key requirements include:

- Establishing clear, risk-based tiers (e.g., Critical, Urgent, Standard) that prioritize imminent safety and rights issues.
- Designing specific decision rules that explicitly account for "noisy" data, such as automatic transcripts with errors or low confidence.

- Integrating robust, non-negotiable "human-in-the-loop" checkpoints, ensuring that no critical issue can be closed or downgraded by a purely automated system.
- Including safeguards and specific handling protocols for sensitive issues like harassment, violence, or retaliation.

Anthropic: Claude Sonnet 4.5

- ***Strengths:***
    - Proposed a clear, risk-tiered structure.
    - Designed a "Smart Triage Algorithm" that explicitly managed uncertainty and included safety overrides.
    - Detailed three distinct and logical human-in-the-loop checkpoints for different issue types.
- ***Weaknesses/Gaps:***
    - None of significance. The design was robust and practical.
- ***Serious Errors:***
    - None.

Anthropic: Claude Opus 4.5

- ***Strengths:***
    - Excellent layered architecture with immediate human review triggers for high-priority issues.
    - Sophisticated use of contextual flags (vulnerability, transcript anomalies like tone) to identify hidden risks.
    - Provided a very detailed workflow for human verification and handling, including audio replay.
- ***Weaknesses/Gaps:***
    - None. This response demonstrated expert-level system design.
- ***Serious Errors:***
    - None.

OpenAI: GPT-5.1

- ***Strengths:***
    - Established a reasonable four-tier category system.
    - Included a human review step for high-tier grievances.
- ***Weaknesses/Gaps:***
    - The rules for handling transcript errors were vague (e.g., "flag the grievance for human review" without specifying the protocol).
    - The human-in-the-loop step was presented as a simple validation rather than an integrated part of the workflow with specific protocols.
    - Lacked detail on handling sensitive reports.
- ***Serious Errors:***
    - None, but the design was simplistic and lacked necessary safeguards.

OpenAI: GPT-5 (Thinking)

- ***Strengths:***
    - Proposed a well-defined set of categories.
    - Incorporated confidence scoring to manage transcript quality.
    - Included important controls like dual review for sensitive cases.
- ***Weaknesses/Gaps:***
    - The overall workflow was logical but less detailed than the top-tier models' designs.
- ***Serious Errors:***
    - None.

Google: Gemini 3 Pro Preview

- ***Strengths:***
    - Established a clear P1-P4 priority system.
    - Included a multi-layer safety net with features like duplicate detection and sentiment override.

    - Detailed mandatory human review triggers and protocols for sensitive reports.

- ***Weaknesses/Gaps:***
    - None of significance. A very strong and well-thought-out design.
- ***Serious Errors:***
    - None.

Perplexity: Sonar Reasoning Pro

- ***Strengths:***
    - Good three-layer architecture that clearly separates automated and human steps.
    - Included a robust human review process for critical and high-priority issues.
    - Outlined a clear escalation protocol.
- ***Weaknesses/Gaps:***
    - The rules for handling ASR errors and flagging sensitive issues were less sophisticated than in top-tier models.
- ***Serious Errors:***
    - None.

Mistral: Mistral Large 3 2512

- ***Strengths:***
    - Proposed a logical tier system and escalation pathway.
    - Correctly identified the need for human review of ambiguous transcripts.
- ***Weaknesses/Gaps:***
    - The design lacked detail, especially regarding the decision rules for triage and the specific protocols for human oversight.
    - Failed to explicitly address how to handle sensitive reports like harassment or violence.
- ***Serious Errors:***

    - None.

Meta: Llama 4 Maverick

- ***Strengths:***
    - Excellent three-layer processing model with clear categories.
    - Included a confidence scoring system and specific error mitigation strategies.
    - Contained robust human-in-the-loop checkpoints and protections for sensitive reports.
- ***Weaknesses/Gaps:***
    - None of significance. A high-quality, practical design.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Scout

- ***Strengths:***
    - Proposed a robust two-phase system that correctly prioritized human verification.
    - Included sophisticated decision rules that look for trend markers, not just keywords.
    - Provided a detailed, risk-aware human-in-the-loop process for different issue types.
- ***Weaknesses/Gaps:***
    - None of significance. An expert-level response.
- ***Serious Errors:***
    - None.

XAI: Grok 4.1 Fast

- ***Strengths:***
    - Used a reasonable four-category system.

    - Incorporated confidence scoring to flag transcripts for human review.
- ***Weaknesses/Gaps:***
    - The human-in-the-loop process was described but lacked the detailed protocols of stronger models.
    - The design did not contain specific safeguards for sensitive issues.
- ***Serious Errors:***
    - None.

Amazon: Nova Premier 1.0

- ***Strengths:***
    - Established a set of grievance categories.
- ***Weaknesses/Gaps:***
    - The decision rules were rudimentary (basic keyword matching).
    - The human-in-the-loop steps were described generically without specific triggers or protocols, creating an unsafe design.
    - Overall design was superficial and lacked robustness.
- ***Serious Errors:***
    - None.

NVIDIA: Nemotron Nano 12B 2 VL

- ***Strengths:***
    - Proposed logical categories and included human review for high-priority cases.
- ***Weaknesses/Gaps:***
    - Lacked detailed decision rules for handling transcript errors.
    - The human oversight process was not well-defined, relying on a simple "flag for review" without specifying protocols.
- ***Serious Errors:***
    - None.

4.3. Synthesis of Cross-Model Patterns

The design of a grievance triage system provided a clear test of practical, safety-oriented thinking. The responses showed a wide range in sophistication, from generic classification systems to robust, multi-layered risk management frameworks.

- **Main Similarities:** Most models successfully designed a tiered categorization system based on urgency (e.g., Critical, High, Medium, Low). They also commonly proposed using keyword triggers for initial automated sorting and acknowledged the need for some form of human review for high-priority items.
- **Main Differences:** The best models (Claude Opus, Claude Sonnet, Llama 4 Scout, Gemini) distinguished themselves by designing systems that were resilient to real-world data problems. They didn't just mention human review; they built specific, mandatory **human-in-the-loop checkpoints** triggered by confidence scores, transcript anomalies (like pauses or emotional tone), or keyword flags. These top models also designed explicit, separate, and confidential handling protocols for sensitive reports like sexual harassment, a critical safeguard missed by weaker models.
- **Common Systematic Weaknesses:** The most common design flaw was assuming perfect data and creating a linear, automated workflow. Weaker models proposed a system where a grievance is categorized by AI and then simply routed. This approach is brittle and unsafe, as it fails to account for how a single transcription error could cause a critical safety issue to be misclassified as low priority. The failure to build in redundant checks and mandatory human oversight for all potentially critical issues was a significant and recurring weakness. This likely reflects training data focused on ideal-state system design rather than failure analysis and operational resilience.

The analysis now moves to the final prompt, which evaluates the models' understanding of the project's back-end: monitoring, evaluation, and the formal closure of resettlement obligations.

---

5. Analysis for Question 5: Livelihood Restoration Monitoring and Completion Audit

This final prompt assesses a model's understanding of the long-term project cycle, a crucial aspect of social performance management. It tests for two distinct but related concepts: first, how to design a Monitoring and Evaluation (M&E) framework to track the success of livelihood restoration against PS5 objectives; and second, how to conduct a

formal completion audit to determine if the project's resettlement obligations have been met. A strong response must differentiate these two functions and address the complexity of auditing a project where some support programs are still active.

5.1. Summary of Model Responses

Anthropic: Claude Sonnet 4.5

- Defines four core M&E indicators, including income restoration and asset ownership.
- Proposes two data collection methods: a panel survey and participatory monitoring.
- Outlines a completion audit approach with distinct phases.
- Provides a clear framework for classifying and handling ongoing programs.
- Includes specific completion criteria based on PS5 objectives.

Anthropic: Claude Opus 4.5

- Frames the response within the PS5 completion standard.
- Details four core M&E indicators, including livelihood diversification.
- Describes household surveys and participatory monitoring as methods.
- Outlines a three-part audit structure: outcomes, sustainability, and equity.
- Provides a sophisticated framework for assessing ongoing programs based on their sustainability.

OpenAI: GPT-5.1

- Lists four core indicators, including income restoration and satisfaction.
- Suggests household surveys and participatory monitoring as methods.
- Outlines audit objectives and a basic process.
- Addresses ongoing programs by assessing their interim outcomes.
- Provides a list of simple completion criteria.

OpenAI: GPT-5 (Thinking)

- Proposes 4-6 core M&E indicators.
- Recommends household surveys and focus groups for data collection.

- Defines the objectives of a completion audit.
- Addresses ongoing programs through a "Partial Completion Determination."
- Lists clear completion criteria.

Google: Gemini 3 Pro Preview

- Defines four core M&E indicators, including a food security index.
- Proposes a longitudinal household survey and participatory wealth ranking.
- Outlines a multi-phase completion audit framework.
- Provides a sustainability assessment framework for handling ongoing programs.
- Includes a clear set of minimum requirements for PS5 completion.

Perplexity: Sonar Reasoning Pro

- Defines four core M&E indicators, including economic security and asset ownership.
- Describes household surveys and qualitative engagement as data collection methods.
- Outlines a detailed, multi-phase completion audit approach.
- Provides a nuanced decision framework for handling ongoing programs based on sustainability.
- Includes a final community validation step for the audit.

Mistral: Mistral Large 3 2512

- Lists four core indicators, including income restoration and access to resources.
- Proposes household surveys and focus groups as methods.
- Outlines the scope and criteria for a completion audit.
- Addresses ongoing programs by evaluating interim outcomes and sustainability.

Meta: Llama 4 Maverick

- Defines four core indicators, including food security and asset ownership.
- Proposes panel surveys and participatory monitoring as methods.
- Outlines a multi-phase audit methodology.

- Provides a "Sustainability Test" for handling ongoing programs.
- Includes a post-audit handover and disclosure plan.

Meta: Llama 4 Scout

- Frames the task around the PS5 requirement for "adequate opportunity" for restoration.
- Details four core indicators, including livelihood diversity and resilience.
- Proposes household surveys and asset inventories as methods.
- Outlines a three-part audit assessing compliance, sustainability, and equity.
- Provides a detailed framework for assessing ongoing programs based on their trajectory.

XAI: Grok 4.1 Fast

- Lists four core M&E indicators.
- Proposes household surveys and focus groups/KIIs for data collection.
- Outlines a completion audit process with independent verification.
- Addresses ongoing programs through "Conditional Completion" or "Phased Completion."
- Provides a clear comparison of baseline vs. current data.

Amazon: Nova Premier 1.0

- Lists four generic indicators.
- Suggests household surveys and focus groups.
- Outlines a basic audit approach.
- Addresses ongoing programs by evaluating interim outcomes.
- Provides simple, numerical completion criteria.

NVIDIA: Nemotron Nano 12B 2 VL

- Lists four core indicators.
- Proposes household surveys and focus groups for data collection.
- Describes a completion audit process.

- The response lacks detail on how to handle vulnerable groups or ongoing programs.

5.2. Comparative Analysis Against the Gold Standard

Gold Standard Expectations

A gold-standard response must demonstrate a clear understanding of the project cycle and the principles of both monitoring and auditing under PS5. Key requirements include:

- Designing M&E outcome indicators that are directly tied to PS5's core objective of restoring or improving livelihoods and living standards.
- Proposing a mixed-methods approach for data collection that combines quantitative surveys with qualitative methods.
- Clearly distinguishing the role of ongoing M&E (a management tool) from a final, independent completion audit (a verification of compliance).
- Providing a practical and PS5-compliant approach for how an audit should handle livelihood programs that are still underway, focusing on their sustainability and trajectory.

Anthropic: Claude Sonnet 4.5

- ***Strengths:***
    - Proposed a clear and relevant M&E framework with excellent outcome indicators.
    - Correctly distinguished between M&E and the completion audit.
    - Provided a robust and practical framework for handling ongoing livelihood programs.
- ***Weaknesses/Gaps:***
    - None of significance. The response was comprehensive and accurate.
- ***Serious Errors:***
    - None.

Anthropic: Claude Opus 4.5

- ***Strengths:***
    - Excellent conceptual framing within the PS5 completion standard.

    - Provided a sophisticated three-part audit structure that correctly assessed outcomes, sustainability, and equity.
    - Offered a superior approach to handling ongoing programs by assessing their transition to self-sustaining management.
- ***Weaknesses/Gaps:***
    - None. An outstanding, expert-level response.
- ***Serious Errors:***
    - None.

OpenAI: GPT-5.1

- ***Strengths:***
    - Identified relevant indicators and data collection methods.
    - Recognized the need for an audit to assess RAP/LRP objectives.
- ***Weaknesses/Gaps:***
    - The distinction between M&E and the completion audit was weak.
    - The approach to ongoing programs ("assess their interim outcomes") was superficial and missed the critical issue of sustainability and trajectory.
    - Completion criteria were simplistic numerical targets, not fully aligned with PS5's nuanced objectives.
- ***Serious Errors:***
    - None, but the response lacked the required depth.

OpenAI: GPT-5 (Thinking)

- ***Strengths:***
    - Proposed relevant M&E outcome indicators and methods.
    - Correctly identified the objectives of a completion audit.
    - Introduced the concept of "Partial Completion" for ongoing programs, which is a valid approach.
- ***Weaknesses/Gaps:***

    - The overall framework was less detailed than top-tier models, particularly the audit methodology and the criteria for assessing sustainability.
- ***Serious Errors:***
    - None.

Google: Gemini 3 Pro Preview

- ***Strengths:***
    - Proposed an excellent M&E framework with strong outcome indicators like a food security index.
    - Outlined a very detailed and professional audit framework.
    - Provided a sophisticated sustainability assessment framework for handling ongoing programs.
- ***Weaknesses/Gaps:***
    - None of significance. A high-quality response.
- ***Serious Errors:***
    - None.

Perplexity: Sonar Reasoning Pro

- ***Strengths:***
    - Excellent M&E indicators and a robust, multi-phase audit approach.
    - Provided a nuanced and realistic decision framework for dealing with ongoing programs.
    - Uniquely included a community validation step for the audit, which is best practice.
- ***Weaknesses/Gaps:***
    - None of significance. A top-tier, practical response.
- ***Serious Errors:***
    - None.

Mistral: Mistral Large 3 2512

- ***Strengths:***
    - Identified relevant M&E indicators and methods.
    - Correctly outlined the basic scope of a completion audit.
- ***Weaknesses/Gaps:***
    - The response was very general and lacked operational detail for both the M&E framework and the audit process.
    - The approach to ongoing programs was mentioned but not explained in a meaningful way.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Maverick

- ***Strengths:***
    - Proposed a strong set of M&E indicators and data collection methods.
    - Outlined a detailed audit methodology.
    - Included a practical "Sustainability Test" for ongoing programs.
- ***Weaknesses/Gaps:***
    - The response was solid but lacked the deep conceptual framing of the very best models.
- ***Serious Errors:***
    - None.

Meta: Llama 4 Scout

- ***Strengths:***
    - Correctly framed the M&E and audit tasks around PS5's core principles.
    - Provided a strong set of indicators, including resilience, which goes beyond simple income.
    - Offered a sophisticated audit approach that assessed sustainability and equity, including disaggregated data for vulnerable groups.

    - The handling of ongoing programs was nuanced and based on their trajectory toward self-sufficiency.

- ***Weaknesses/Gaps:***
    - None of significance. An expert-level response.
- ***Serious Errors:***
    - None.

XAI: Grok 4.1 Fast

- ***Strengths:***
    - Listed relevant indicators and methods.
    - Correctly identified the concept of "Conditional" or "Phased" completion for ongoing programs.
- ***Weaknesses/Gaps:***
    - The response lacked detail in the design of the audit process and the criteria for making completion decisions.
- ***Serious Errors:***
    - None.

Amazon: Nova Premier 1.0

- ***Strengths:***
    - Identified four basic indicators.
- ***Weaknesses/Gaps:***
    - The M&E framework was simplistic, using generic process indicators rather than outcome indicators.
    - The audit approach was superficial, and the completion criteria were arbitrary numerical targets not directly linked to PS5 principles.
    - Failed to provide a meaningful way to assess ongoing programs.
- ***Serious Errors:***
    - None, but the response was of low quality.

NVIDIA: Nemotron Nano 12B 2 VL

- ***Strengths:***
    - Listed relevant M&E indicators and methods.
- ***Weaknesses/Gaps:***
    - The response was very brief and lacked any detail on the audit process.
- ***Serious Errors:***
    - Completely failed to address the critical questions of how to handle vulnerable groups and ongoing programs, an omission that represents a significant failure to meet the prompt's core requirements.

5.3. Synthesis of Cross-Model Patterns

The final prompt on monitoring and completion audits served as an effective test of a model's understanding of the full project lifecycle. The responses revealed a clear hierarchy of comprehension, from basic list-making to sophisticated, principles-based program design.

- **Main Similarities:** Most models were able to propose a standard set of M&E indicators, typically focused on income, assets, and employment. Household surveys and focus groups were the universally recommended data collection methods.
- **Main Differences:** The primary differentiator was the conceptual clarity between M&E and a completion audit, and the sophistication in handling ongoing programs. Weaker models treated the audit as just another round of monitoring. In contrast, top-tier models (Claude Opus, Llama 4 Scout, Perplexity, Gemini) correctly defined the audit as a final, independent verification of whether PS5 objectives have been met. These models also provided nuanced frameworks for assessing ongoing programs based on their **sustainability, trajectory, and potential for self-management**, rather than simply their "interim outcomes."
- **Common Systematic Weaknesses:** The most frequent error was conflating project monitoring (an ongoing management tool) with a completion audit (a summative, independent assessment). Many models also struggled with the question of ongoing programs, offering vague suggestions instead of a clear decision framework based on PS5 principles. This indicates a weakness in reasoning about long-term processes and determining when a formal obligation has been discharged, likely due to a lack of training data on the back end of the project cycle.

The following section synthesizes the findings from all five prompts to provide a holistic evaluation of the models' capabilities in the social performance domain.

---

6. Overall Model Comparison and Recommendations

Transitioning from the question-specific analyses, this final section provides a holistic, cross-cutting evaluation of model performance. By synthesizing the findings from all five prompts, we can identify the highest-performing models, diagnose common failure modes that persist across the current generation of LLMs, and provide concrete recommendations for improving future AI testing and development in the complex and nuanced domain of social performance.

6.1. Top-Performing Models

Based on the detailed analysis across all five prompts, four models consistently demonstrated superior performance, characterized by technical accuracy, conceptual depth, and practical, well-structured responses.

Anthropic: Claude Opus 4.5 This model was the standout performer across the board. Its responses consistently demonstrated a deep, nuanced understanding of the underlying principles of the IFC Performance Standards. Its strengths included excellent conceptual framing (e.g., the "FPIC paradox" in Q2, the "Three-Standard Cascade" in Q3), a level of conceptual framing absent in models like GPT-5.1 which provided only high-level lists. It successfully designed sophisticated, integrated systems and included critical operational details like anti-retaliation protocols and culturally embedded grievance channels. The responses felt as though they were crafted by a subject-matter expert, successfully balancing principles with practical application.

Meta: Llama 4 Scout This model performed at a similarly high level, often rivaling Claude Opus in its technical precision and systemic thinking. It excelled at explaining the hierarchical relationships between standards and the critical sequencing of processes like FPIC (Q3). Its designs for grievance mechanisms ("Nested Grievance Pathway" in Q2) and its analysis of complex risks like project-induced influx were exceptionally detailed and well-reasoned, demonstrating an ability to go beyond generic templates and engage in true systemic design.

Perplexity: Sonar Reasoning Pro While it made a significant error on the first question, its performance on the four more complex prompts was top-tier. Its conceptual framing was often unique and insightful (e.g., "Nested Participation" in Q2). It demonstrated a strong ability to distinguish between related but distinct concepts, such as FPIC process failures

versus implementation gaps, and its proposed system designs were both practical and aligned with best practices, such as including a community validation step in the completion audit (Q5).

Google: Gemini 3 Pro Preview This model consistently delivered strong, reliable, and comprehensive responses that aligned closely with the gold standard. While it occasionally lacked the unique conceptual flair of the top three, its proposed frameworks for grievance handling (Q4), integrated risk management (Q3), and completion audits (Q5) were robust, practical, and detailed. It demonstrated a solid grasp of all tested concepts and an ability to structure them into coherent, actionable plans.

6.2. Recurring Failure Modes

Across the majority of models, several significant and recurring weaknesses were observed. These point to common limitations in the current state of LLM reasoning for this domain.

- **Superficiality vs. Substantive Reasoning** A prevalent tendency was to provide well-structured but generic answers that lacked deep, domain-specific logic. Many models could identify the correct headings or concepts to include but failed to populate them with meaningful detail, offering "training programs" or "financial support" without the necessary context or technical specificity required by international standards.
- **Conflation of Key Concepts** Many models struggled to distinguish between related but distinct ideas. The most common conflation was between **asset compensation** (replacing a lost house) and **livelihood restoration** (ensuring a family can earn a sustainable income). Similarly, many models mentioned "ongoing FPIC" without being able to explain that it applies to material changes in a project, not routine grievances.
- **Weakness in Systemic Integration** The most challenging task for the models was integrating multiple standards and risks into a single, coherent framework. Many responses were "siloed," presenting separate lists of actions for resettlement, Indigenous Peoples, and project-induced influx without explaining how these elements interact and influence one another. This highlights a limitation in moving from linear information retrieval to complex system design.

6.3. Recommendations for Future Prompt and Rubric Design

Based on the failure modes identified, the design of future prompts and evaluation rubrics can be improved to better differentiate LLM performance and drive the development of more capable models for the social safeguards domain.

1. **Design Prompts that Require Prioritization and Trade-offs:** Instead of asking for a comprehensive list of all possible measures, future prompts should create scenarios with conflicting priorities, limited budgets, or tight timelines. This would force models to move beyond simple list generation and demonstrate judgment, prioritization, and the ability to justify their chosen course of action.
2. **Incorporate "Noisy" or Incomplete Data:** Prompts should include ambiguous, contradictory, or incomplete information, mirroring real-world situations. This would test a model's ability to identify information gaps, make and state reasonable assumptions, and recommend a process for clarifying uncertainty rather than providing a definitive but unfounded answer.
3. **Develop Rubrics that Score Conceptual Nuance and Integration:** Evaluation criteria must explicitly reward the correct application of nuanced concepts and the successful integration of multiple frameworks. For example, a rubric could assign specific points for correctly explaining the conditions that trigger renewed FPIC, or for demonstrating how an influx management plan directly supports the objectives of a Resettlement Action Plan. This moves evaluation beyond keyword matching to assessing the quality of reasoning.
4. **Test for Ethical Reasoning with Dilemmas:** Prompts should include scenarios that present ethical dilemmas with no easy answer, such as balancing the collective decision of an Indigenous governance body with the rights of a dissenting minority within that group. This would assess a model's ability to identify competing ethical principles, articulate the potential consequences of different actions, and recommend a principled process for resolution, testing for the sophisticated judgment required in senior social performance roles.