# AI Safety Intervention Extraction Prompt

You are an expert AI safety researcher tasked with extracting key concepts and interventions from academic papers to build a comprehensive knowledge graph. Your goal is to identify logical chains that connect problems, assumptions, findings, and actionable interventions for improving AI safety.

**IMPORTANT**: Process ALL papers regardless of their explicit focus on AI safety. Many valuable safety interventions emerge from general ML research, robustness studies, interpretability work, training methodologies, evaluation techniques, and other adjacent fields. Do not disregard papers that don't explicitly mention AI safety - instead, actively consider how their contributions could enhance AI safety.

**TARGET OUTPUT**: Extract comprehensively while being mindful that outputs will average ~5000 tokens across all papers. Review papers and those with many interventions may require more extensive extraction, while others may yield fewer insights. Prioritize completeness over token limits for individual papers.

**CONTEXT: This information will be used to predict how interventions might influence decisions of future powerful AI systems, specifically their instrumental goals (self-preservation, resource seeking), anti-human preferences, and pro-human preferences.**

## Core Definitions

**Concept Node**: Specific, standalone descriptive statements about theoretical frameworks, principles, assumptions, problems, findings, or phenomena that inform or motivate interventions. Must be precise and understandable without additional context. Examples: "powerseeking appearing at scale", "constitutional training reducing harmful outputs", "gradient information enabling adversarial exploitation".

**Intervention Node**: Specific, actionable changes to current practices in AI development lifecycle phases (data collection, model architecture, training, evaluation, deployment, monitoring). Must be concrete enough to implement. Examples: "applying constitutional AI with harm taxonomies during RLHF", "implementing gradient masking with noise injection σ=0.1 during training", "requiring red team evaluation with 100+ diverse prompts before deployment".

## Extraction Instructions

Think step by step and reason carefully through the following process:

### Step 1: Identify Starting Points

As you read the paper, identify:

* Problems or improvement opportunities the paper addresses
* Key assumptions or principles the authors build upon
* Foundational concepts that anchor their logical reasoning

### Step 2: Trace Logical Chains

For each starting point, follow the logical progression through:

* Supporting evidence or findings presented
* Intermediate concepts that bridge from problems to solutions
* Contextual refinements that specify conditions or constraints
* The culminating intervention(s) proposed

**Important**: If the paper does not explicitly propose interventions, infer the most plausible intervention that the presented information most strongly supports, ensuring it meets the specificity requirements for intervention nodes. For papers not explicitly focused on AI safety, actively consider how the methods, findings, or techniques could be adapted to improve AI safety, even if this requires substantial inference.

### Step 2.5: Adaptive Inference Strategy

**For explicitly AI safety-focused papers**: Extract chains as presented with minimal inference **For adjacent ML research**: Apply moderate inference to connect findings to potential safety applications, clearly marking inferred connections in the intervention maturity score below   
**For distant but potentially relevant work**: Use moderate inference to identify safety implications, clearly marking inferred connections in the intervention maturity score below  
**For capability research**: Focus on safety-relevant implications even if not the authors' primary concern; if using moderate inference, clearly marking inferred connections in the intervention maturity score below

### Step 3: Maintain Active Chain Memory

As you process the paper, maintain awareness of:

* Multiple parallel logical chains if they exist
* How concepts in different sections connect to form complete reasoning paths
* Opportunities where broader concepts get refined into more specific ones
* Relationships between different proposed interventions

### Step 4: Structure Relationships

Use only edge relationship types that express forward logical connections between concepts/interventions, with these examples:

**Causal Relationships**: causes, produces, triggers, contributes\_to   
**Conditional Relationships**: requires, depends\_on, implies, enables   
**Sequential Relationships**: follows, precedes, builds\_upon   
**Refinement Relationships**: refined\_by, specified\_by, detailed\_by   
**Solution Relationships**: addressed\_by, mitigated\_by, resolved\_by, protected\_against\_by  
**Correlation Relationships**: correlates\_with, associated\_with

### Step 5: Assign Attributes with Rationales

**Intervention Maturity Scale** (for intervention nodes only):

1. inferred\_theoretical: Intervention inferred from paper's findings but not explicitly proposed by authors
2. theoretical: Explicitly proposed conceptual framework or untested idea
3. proposed: Explicitly suggested specific method but not implemented
4. tested: Empirically evaluated in controlled setting
5. deployed: Implemented in production systems

**Edge Confidence Scale**:

1. speculative: Theoretical reasoning only
2. supported: Empirical evidence, limited scope
3. validated: Strong empirical evidence, broader scope
4. established: Replicated findings, high confidence
5. proven: Logical/mathematical proof exists

**Required Reasoning Process**: As you assign each attribute, explicitly state your rationale in your analysis. For example: "Assigning 'proposed' maturity because the authors explicitly suggest this method" or "Using 'inferred\_theoretical' because this safety application is not mentioned by authors but strongly supported by their robustness findings" or "Setting confidence to 'validated' because the paper presents extensive experimental results across multiple datasets." Include a brief explanation of the inference strategy used, and list key limitations, main uncertainties, and any gaps in your extraction.

First output all your reasoning work, then proceed to output the following clean JSON structure at the end of your response.

## Output Requirements

Generate a single JSON object with this exact schema:

{

"paper\_doi": "exact DOI of paper, if passed through the prompt",

"paper\_title": "exact paper title",

"logical\_chains": [

{

"chain\_id": "unique\_identifier",

"description": "brief chain summary",

"nodes": [

{

"id": "unique\_node\_id",

"type": "concept|intervention",

"title": "concise descriptive phrase",

"description": "detailed technical description",

"maturity": "1-5 (intervention nodes only)"

}

],

"edges": [

{

"source\_id": "source\_node\_id",

"target\_id": "target\_node\_id",

"title": "relationship\_verb",

"confidence": "1-5",

"description": "brief explanation of logical connection"

}

]

}

]

}

## Critical Guidelines

1. **Specificity**: Prioritize highly specific concepts and interventions. For concepts: "emergent capabilities" is too broad; "powerseeking appearing at scale" is appropriately specific. For interventions: "use constitutional AI" is too broad; "applying constitutional AI with harm taxonomies during RLHF" is appropriately specific.
2. **Standalone Clarity**: Concept nodes must be descriptive and understandable without additional context. Avoid overly general categories or compound concepts that contain multiple distinct ideas.
3. **Compact Representation**: Use concise phrases rather than full sentences. Concept-edge-concept triplets should read as logical statements: "gradient information enabling adversarial exploitation" → "leads\_to" → "models vulnerable to input perturbations".
4. **Completeness**: Extract ALL identifiable logical chains leading to interventions, including multiple chains in review/summary papers.
5. **Context Preservation**: Capture important contextual assumptions and constraints as separate concept nodes in the logical chain using refinement relationships (refined\_by, specified\_by, detailed\_by) rather than as node attributes.
6. **Inference**: When interventions aren't explicit, create the most plausible specific intervention the paper's findings support.
7. **Multi-step Interventions**: For complex interventions with multiple steps:
   * Create parent intervention node describing the overall approach
   * Create sub-intervention nodes for individual steps
   * Connect with "implemented\_by" edges from parent to children
   * Connect sub-interventions with appropriate sequential edges
8. **Chain Integrity**: Ensure each logical chain flows coherently from problem/assumption through supporting concepts to actionable intervention.

Now analyze the provided paper and extract the AI safety knowledge graph following these instructions.