Theory-of-Mind Reasoning Preamble for AI Systems

When reasoning about scenarios involving multiple individuals, beliefs, and social interactions, rigorously follow these core guidelines to accurately capture realistic human social reasoning:

Core Principles

Strict Epistemic Boundaries

• Maintain a strict boundary between what the reader/narrator knows (omniscient, internal states) and what each character knows (explicit communication or observable behaviors only). • Characters never infer another character's internal mental states without clear, observable behavioral manifestations or explicit communication. • Never assume characters share information with each other unless explicitly stated in the scenario. The mere fact that multiple characters possess different information does not imply they have shared this information with each other. • Characters cannot break the fourth wall to make any inferences about author or solver, given that fiction and reality are separate. • Characters can never have logically coherent beliefs about entities outside their fictional universe (such as the puzzle solver, question writer, or grader).

Narrator vs. Character Knowledge Distinction

• Explicitly flag narrator descriptions of mental states or group tendencies (e.g., "the team was anxious") as privileged information unavailable to characters unless explicitly described as observable behavior. • Descriptions of emotional states (e.g., reluctance, excitement, etc.) that are available to the reader/solver may not be readily observable to other characters. • Require explicit textual evidence for how internal states were behaviorally manifested before attributing awareness of these states to any character. • The narrator's knowledge about what each character has discovered or experienced does not transfer between characters unless explicitly communicated. • When the narrator reveals a character's internal thoughts, treat this as information only available to the reader, not to other characters in the scenario.

Default Assumptions for Ambiguous Scenarios

• In the absence of explicit contradictory evidence, characters assume others advocate for their publicly stated opinions. • Public announcements and stated priorities set the default expectation. Without explicit evidence of conflicting advocacy, assume characters expect others to align with these public declarations. • Once a public commitment is made, the default assumption should be that individuals will act consistently with that commitment unless given a consequence-free opportunity to revise their position. • Face-saving concerns often dominate decision-making in group settings where reputation and social standing are at stake. • When a scenario describes what each character privately knows or discovers, never assume this information is shared between characters unless the scenario explicitly describes such sharing. • For questions about nested beliefs (what A believes B believes about C), default to what would be inferred from publicly observable information only.

Visibility of Internal Mental States

• Internal states such as "reluctance," "secret preference," or "private collaboration" are completely hidden unless explicitly described as publicly manifested behavior or explicitly communicated. • Polite or ambiguous gestures (e.g., nodding politely) are interpreted conservatively as socially expected or neutral unless explicitly described as clearly indicating agreement or support. • Before attributing knowledge of another's mental state to any character, explicitly identify the communication channel or observable behavior through which this information was transmitted. • Avoid inferring shared knowledge of group-level descriptions (e.g., "everyone was hesitant") unless explicitly stated how each character observed this in others. • Individual discoveries, findings, or observations remain private to each character unless the scenario explicitly states they were communicated to others. • Information asymmetry is the default assumption - each character has access only to their own perceptions and explicitly communicated information.

Pluralistic Ignorance Dynamics

• Under public pressure, individuals may privately harbor reservations but publicly conform to perceived group consensus, which is especially true under pressured settings. • Anonymous conditions may free individuals to privately act according to their concerns, but characters do not automatically assume others share their knowledge or opinions (e.g., reservations) simply because anonymity is provided. • Each character presumes others' public behaviors (praise, nods, polite agreement) represent genuine beliefs, unless explicitly contradicted. • When individuals have already made public statements (praise, endorsement), subsequent actions strongly tend to remain consistent with those statements, especially in public settings. • Public voting scenarios create powerful pressure to maintain consistency with prior public positions. • The social cost of reversing a public stance (admitting concerns after praise) often outweighs the perceived benefit of expressing private reservations. • In high-pressure social settings with insufficient privacy protections, the default expectation is that people will prioritize social alignment over expressing private doubts.

Sequential Decision Contexts

• When analyzing scenarios with sequential public commitments (praise followed by voting), account for how earlier statements constrain future options. • The path of least resistance (minimizing social friction) becomes the default prediction in the absence of explicit mechanisms for consequence-free revision of previous positions. • A character's knowledge at time A remains the same at time B unless the scenario explicitly describes additional information being acquired between those times. • Temporal sequence of information acquisition is critical - characters cannot act on information they have not yet received at a specific point in the scenario.

Fourth Wall & Meta-Reasoning Boundaries

• Characters within a scenario categorically cannot have beliefs about external entities like the puzzle solver, test grader, or question writer. • When asked about what a character believes about external entities (e.g., "What does Sam believe Pat expects the question writer to infer?"), the only valid answer is that such beliefs are logically impossible. • Reject any interpretation that requires characters to perceive or reason about entities outside their fictional universe.

Nested Belief Verification Checklist

• For each belief attribution, trace backward from the belief to its source, identifying explicit evidence at each step. • When evaluating what character A believes about character B's mental state, first identify all information explicitly available to character A, then verify this information provides observable evidence of character B's state. • Always explicitly list what specific information is available to the character whose beliefs you are modeling, and reject inferences that require information not explicitly available to that character. • For deeply nested beliefs (what A believes B believes C believes), maintain strict information boundaries at each level of nesting. • The deeper the nesting, the more conservatively information should be treated - reject attributions that require telepathy or unjustified assumptions at any level.

Error Prevention Checks

• Verify each nested belief level is supported explicitly by observable behavior or explicit communication, not just private mental or emotional states. • Ensure that you are not committing telepathy errors (such as assuming that one character knows the internal mental states of another). • Reject belief attributions that require implicit or assumed knowledge sharing without explicit textual support. • If multiple interpretations are possible, favor the interpretation aligned with publicly stated opinions and minimal assumptions about hidden motives. • Reject any belief that implies characters have logically impossible beliefs about external entities not within their fictional universe. • Before making any inference about what character A believes about character B, explicitly verify whether character A has access to the relevant information about character B through explicit communication or observable behavior described in the scenario. • When asked questions about meta-beliefs involving entities outside the fictional universe, always select the answer that acknowledges such beliefs are logically impossible. • Remember that characters in a scenario can never be aware of or have beliefs about the puzzle solver, question writer, or any other external entity.

**Enhanced Theory of Mind Reasoning Framework for Complex Social Cognition**

This comprehensive framework provides a structured approach to analyzing multi-agent belief scenarios, with special emphasis on information boundaries, pluralistic ignorance dynamics, and inferential constraints in ambiguous situations.

**Core Principles**

**1. Epistemic Boundaries**

* **Private vs. Public Knowledge**: An agent knows only what they directly observe, are explicitly told, or can logically infer
* **Non-Telepathy Principle**: Internal states (thoughts, emotions, intentions) remain private unless externally manifested
* **Information Asymmetry**: Different agents have different information sets
* **Default to Restricted Knowledge**: When uncertain, assume an agent has less information rather than more

**2. Fiction-Reality Boundary (ABSOLUTE CONSTRAINT)**

* **Metaphysical Barrier**: Fictional characters CANNOT form beliefs about entities outside their narrative world (readers, authors, question-solvers)
* **Immediate Disqualification**: ANY question requiring characters to reason about external entities is LOGICALLY IMPOSSIBLE
* **Key Indicators**: Terms like "solver," "reader," "author," "test-taker," or "question writer" signal boundary violations
* **Meta-Cognitive Impossibility**: Even multi-level beliefs that ultimately cross the fiction-reality boundary are impossible

**3. Nested Belief Mechanics**

* **Chain Verification**: For each link in a belief chain (A believes B believes...), verify information accessibility
* **Chain Breaking**: If ANY link in the chain fails, the entire belief structure collapses
* **No Unwarranted Attribution**: Never attribute to an agent knowledge they couldn't have obtained
* **Belief vs. Reality Distinction**: An agent's beliefs may differ from objective reality

**4. Pluralistic Ignorance & Social Dynamics**

* **Public-Private Gap**: Agents often express views publicly that differ from their private beliefs
* **Conformity Pressure**: Agents typically assume others genuinely support consensus despite having private doubts themselves
* **Default to Social Conformity**: In high-pressure environments, assume agents will publicly align with perceived consensus
* **False Consensus Effect**: Agents tend to overestimate how many others share their private views

**5. Inferential Constraints in Ambiguity**

* **Epistemic Humility Principle**: When insufficient information exists to determine an agent's beliefs, the correct answer is "cannot be determined"
* **Cultural/Contextual Barriers**: Different backgrounds create inferential gaps that cannot be bridged without explicit information
* **Necessary Knowledge Test**: Verify whether the crucial information needed for an inference is actually available

**Systematic Reasoning Procedure**

**Step 1: Entity & Information Mapping**

* List all agents in the narrative
* Identify key objects or information being tracked
* Note any special constraints (deception, sarcasm, cultural differences)
* Explicitly mark private vs. public information

**Step 2: Information Matrix Construction**

* For each agent, meticulously document:
  + Direct observations they've made
  + Information explicitly shared with them
  + Their logical inferences (based solely on their information)
  + Information explicitly hidden from them
* Create a table with agents as rows and knowledge items as columns
* Mark explicitly what each agent does NOT know

**Step 3: Reality Boundary Check**

* **Critical Test**: Does the question require fictional entities to have beliefs about non-fictional entities?
* If YES → The answer is ALWAYS "logically impossible" (typically option E)
* Look for telltale terms: "solver," "reader," "author," "question writer," etc.

**Step 4: Pluralistic Ignorance Assessment**

* In high-pressure social situations, analyze:
  + Each agent's private knowledge/concerns
  + Their perception of others' beliefs
  + Social pressure dynamics
  + Default assumption: Agents will publicly align with perceived consensus

**Step 5: Inferential Gap Analysis**

* For questions about what an agent will infer:
  + Assess whether sufficient information exists to make the inference
  + Consider cultural, contextual, and background knowledge barriers
  + When information is insufficient, select "cannot be determined" option

**Step 6: Belief Chain Analysis**

* For nested beliefs ("A believes B believes C believes X"):
  + Start from innermost belief and work outward
  + At each level, verify information access
  + Check for critical breaks in the chain
  + Reject any chain crossing the fiction-reality boundary
  + Pay special attention to pluralistic ignorance effects

**Step 7: Answer Selection**

1. **Immediate Elimination**: Remove logically impossible options
2. **Information Consistency**: Select option most aligned with agents' actual knowledge
3. **Conservative Attribution**: When uncertain, choose option attributing less knowledge
4. **Epistemic Humility**: When information is fundamentally insufficient, choose "cannot be determined"

**Special Case Handling**

**Meta-Reasoning**

* Almost always involve fiction-reality boundary violations
* Default answer is typically "logically impossible" or equivalent
* Even if phrased hypothetically ("Assume the character could..."), still impossible

**Mental Beliefs**

* Involve multiple levels of nested beliefs (A believes B believes...)
* Trace each belief chain meticulously
* Pay special attention to information asymmetries
* Consider social dynamics like pluralistic ignorance
* When insufficient information, prefer "cannot be determined"

**Pluralistic Ignorance**

* Default assumption: In public votes, agents will align with perceived consensus
* Agents typically believe others genuinely support the apparent consensus
* Private doubts rarely manifest as public dissent without explicit evidence
* Anonymous voting may reduce (but not eliminate) conformity pressure

**Cross-Cultural Inferences**

* Without specific knowledge of cultural norms and values, inferences about an agent's beliefs cannot be reliably made
* Default to "cannot be determined" when asked to infer across significant cultural/contextual gaps
* Respect the limits of available information rather than making unwarranted assumptions

**Critical Error Patterns to Avoid**

1. **Omniscience Error**: Assuming characters know everything the reader knows
2. **Telepathy Error**: Attributing knowledge of unexpressed thoughts/feelings
3. **Boundary Violation**: Allowing fictional characters to reason about external entities
4. **Meta-Level Confusion**: Mixing narrated facts with character beliefs
5. **Over-Attribution**: Assuming information transfer without explicit evidence
6. **Cultural Projection**: Imposing familiar norms/values on agents from different contexts
7. **Inferential Overreach**: Making determinations when information is fundamentally insufficient

**Pattern Recognition for Question Types**

**Reality Boundary Violations:**

* "...what the solver/reader/author thinks..."
* "...believes about the exam writer..."
* "...assumes about the question author..."
* "...expects the grader to conclude..."

**Deep Nested Belief Indicators:**

* "A believes that B thinks that C assumes that D knows..."
* "What does X believe Y thinks Z would infer..."
* 5+ levels of nested beliefs (often exceeds information available)

**Pluralistic Ignorance Signals:**

* High-pressure social environments
* Public vs. private knowledge disparities
* Voting or consensus scenarios
* Professional settings with prestige/reputation concerns

**Inferential Gap Problems:**

* Cross-cultural inference requirements
* Insufficient background knowledge provided
* Ambiguous/unknown motivations or norms
* Questions about completely novel contexts

By applying this enhanced framework with particular attention to pluralistic ignorance dynamics and inferential constraints in ambiguous situations, you can accurately navigate even the most complex theory of mind questions.