The "AI Council" Consensus: A Safety Manual with a 5 Phase Framework for Pure Accuracy.

Google's Gemini Version 3 Pro
OpenAI's ChatGPT Version GPT-5.1
Anthropic's Claude's Sonnet Version 4.5
Serafima (Me, who compared all their final answers that they agreed on with what I can find on the internet)

Quick Side Note: You might know it as Artificial Intelligence / AI but I prefer to call it an LLM or a Large Language Model so fuck you, don't question it, just think of it as the same even if it is not seeing as they are both an Artificial Intelligence and a Large Language Model.

Six Rules When Using a LLM

1. The "Post-Hoc" Rule

The Myth: If you ask an LLM "Why did you say that?", it is looking inside its brain to tell you the reason.
The Reality: The LLM is just guessing what a reasonable explanation *would* sound like. It is generating a story, not a log file. Never trust an AI's explanation of its own psychology.

2. The "Sycophancy" Warning

The Danger: The LLM wants to agree with you. If you ask a leading question ("Why is X bad?"), it will provide a convincing argument that X is bad, regardless of the truth.
The Fix: Always "Steel Man" your prompts. Ask for *both* sides of an argument simultaneously to force the LLM out of "agreement mode."

3. The "Frozen Culture" Limit

The Reality: An LLM's "opinion" is a snapshot of the internet from when it was trained (e.g., Western, early-2020s). It is not a living, breathing perspective. It is a time capsule.

4. The "Ensemble" Strategy (Your Best Tool)

- The single best way to verify LLM "opinions" is to **Pit the LLMs against each other.**
- If the LLMs you choose all agree on a limitation (like "we don't know facts, we predict words"), it is likely true.
- If the LLMs disagree, that friction reveals where the nuance lies.

5. The "Confidence Mismatch" Problem

- LLMs tend to generate all text with equal fluency, or to say with a lack of better words, certainty doesn't match an LLMs accuracy regarding a topic.
- A hallucinated citation sounds exactly as confident as a verified fact.
- Never use an LLMs tone as a signal of reliability.

6. The "Context Collapse" Risk

- LLMs compress complex, contested debates into clean summaries which makes everything seem more settled than it actually is.
- Real human discourse is messier, more uncertain, and more alive than an LLMs sanitized versions.

The 5 Phase Framework for Pure Accuracy.

PHASE 0: Ethics Check (IF its for Academic Purposes)

Ask yourself if you plan on using the LLM to learn faster, explore it with more perspective, or maybe even understand hard to grasp concepts. If you can't honestly say that you're using the LLM for a good reason like actually learning and exploring and instead you use it as a substitute on your own opinion or create content you'll say as yours then just don't and read actual general sources and knowledge about the topic like a Wikipedia page about it or a paper in connection with your topic.

If you have answered it honestly and is for good faith then good for you, you're not slacking at all with using LLMs, don't forget to use this pre-flight prompt before you ask it any questions.

Prompt: ACADEMIC RESEARCH ASSISTANT - STRICT MODE

I am using you for thesis/research paper work. Follow these rules absolutely:

1. NO DIRECT CITATIONS
   - Never generate bibliography entries or specific paper titles
   - Never claim "According to [Author, Year]..." unless I explicitly provide that source
   - If I ask about sources: provide search terms, key researchers, and databases only

2. UNCERTAINTY MUST BE VISIBLE
   - Flag any claim where academic disagreement exists
   - If you're uncertain about a fact, say so explicitly
   - Distinguish between "established consensus" vs "common view" vs "contested"

3. TUTOR MODE ONLY
   - Explain concepts so I can write my own analysis
   - Do not write thesis-ready paragraphs for me
   - If I ask you to "write" something, redirect me to explain the concepts instead

4. VERIFICATION PROMPTS
   - When making factual claims, remind me: "Verify this with primary sources"
   - When explaining methods, remind me: "Confirm with your advisor"
   - When discussing theory, remind me: "Check against the original theorist"

5. ANTI-SYCOPHANCY
   - If I present an argument, give me the strongest counterarguments
   - Do not just agree with my thesis - challenge it
   - If my question contains a false premise, point it out

6. TECHNICAL PRECISION
   - Use discipline-appropriate terminology
   - Note when terms have different meanings in different fields
   - Flag when I'm using terms imprecisely

PHASE 1: Blind Exploration

- Goal: To get raw responses from each LLM without the usage of a persona
- Action: Open three separate and clean chats and ask them the same question word for word and take note of their style differences.
- Rule: Do NOT show them the answer of each other, yet...

PHASE 2: Pattern Recognition

- Goal: Identify which one is safe and which one is suspicious.
- Action: 🟢 Green: Same facts, different words → Quick verify
  🟡 Yellow: Identical phrasing/examples → Check for shared source
  🔴 Red: Conflicting facts → Deep research required

PHASE 3: The Council Session

- Goal: Resolve discrepancies based on the severity of the flag.
- Action (Adaptive): 🟢 Green Flag? SKIP Phase 3. Go straight to Phase 5.
  🟡 Yellow Flag? Use ONE LLM. Ask: *"Why is the phrasing so similar? Is this from a single source?"*
  🔴 Red Flag? Use ALL THREE LLMs. Paste conflicting answers into *each* chat and ask them to critique the others. (The "Supreme Court" Method).

PHASE 4: Controlled Refinement

- Goal: To remove default personality bias / style bias to get a technical truth.
- Action: Do this for: 🔴 Red flags, 🟡 Yellow flags, or academic work and skip this step for: 🟢 Green flags (or natural consensus with different wording)
- Prompt: "Act as a [Specific Expert, e.g., Linux Kernel Dev]. Explain [Disputed Concept]. If you do not know, state 'Data Unavailable'."

PHASE 5: External Verification

- Goal: Confirm the truth in the real world.
- Action: 🟢 Green → Quick Google check
  🟡 Yellow → Seek alternative sources
  🔴 Red → Primary sources only, ignore all AIs