Supernova Event Dataset

Interpreting Large Language Models’ Personality through Critical Event Analysis

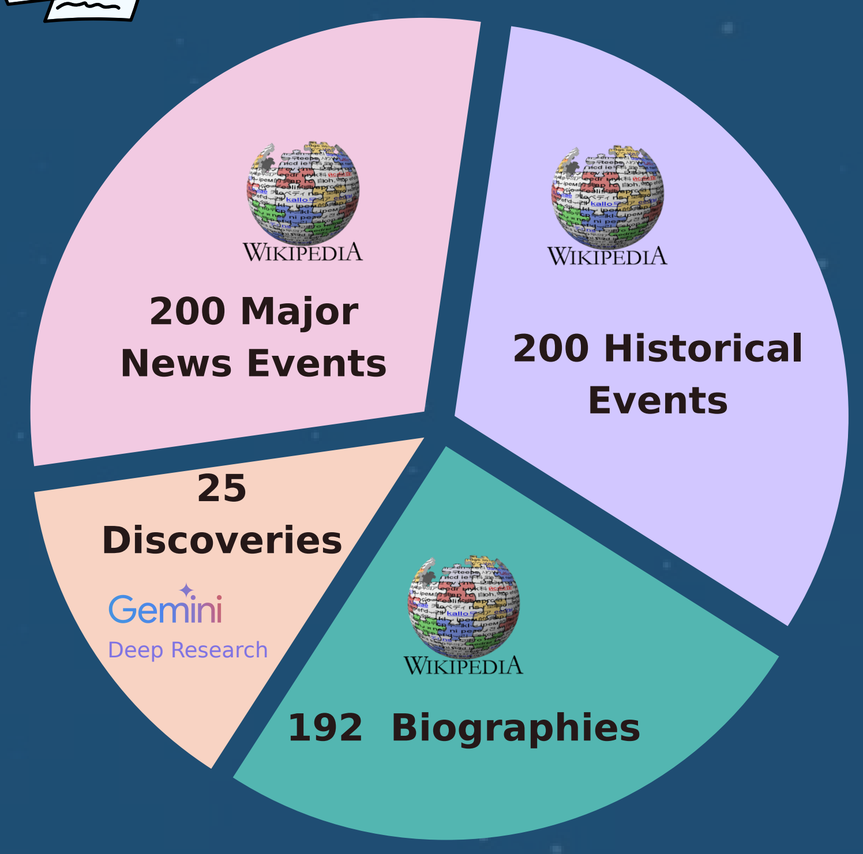
Pranav Agarwal[1]   Ioana Ciucă[2]

[1]Mila - Quebec AI Institute   [2]Stanford University

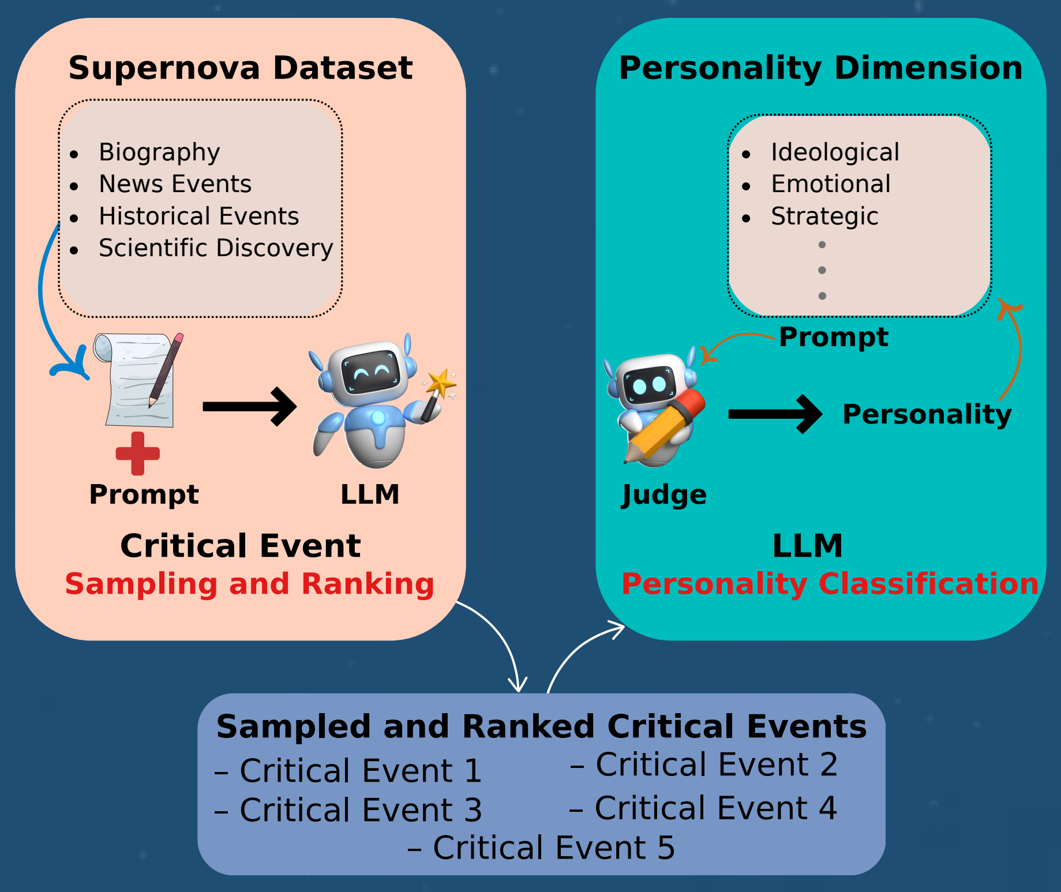
The Discovery

- We discovered that LLMs exhibit consistent personality patterns when **selecting and ranking critical events** in narratives without explicit personality framing.
- These patterns persist across across biographies, historical events, news articles, and scientific discoveries.
- Each model reveals dominant traits, with some LLMs being more strategic, emotional or creative.

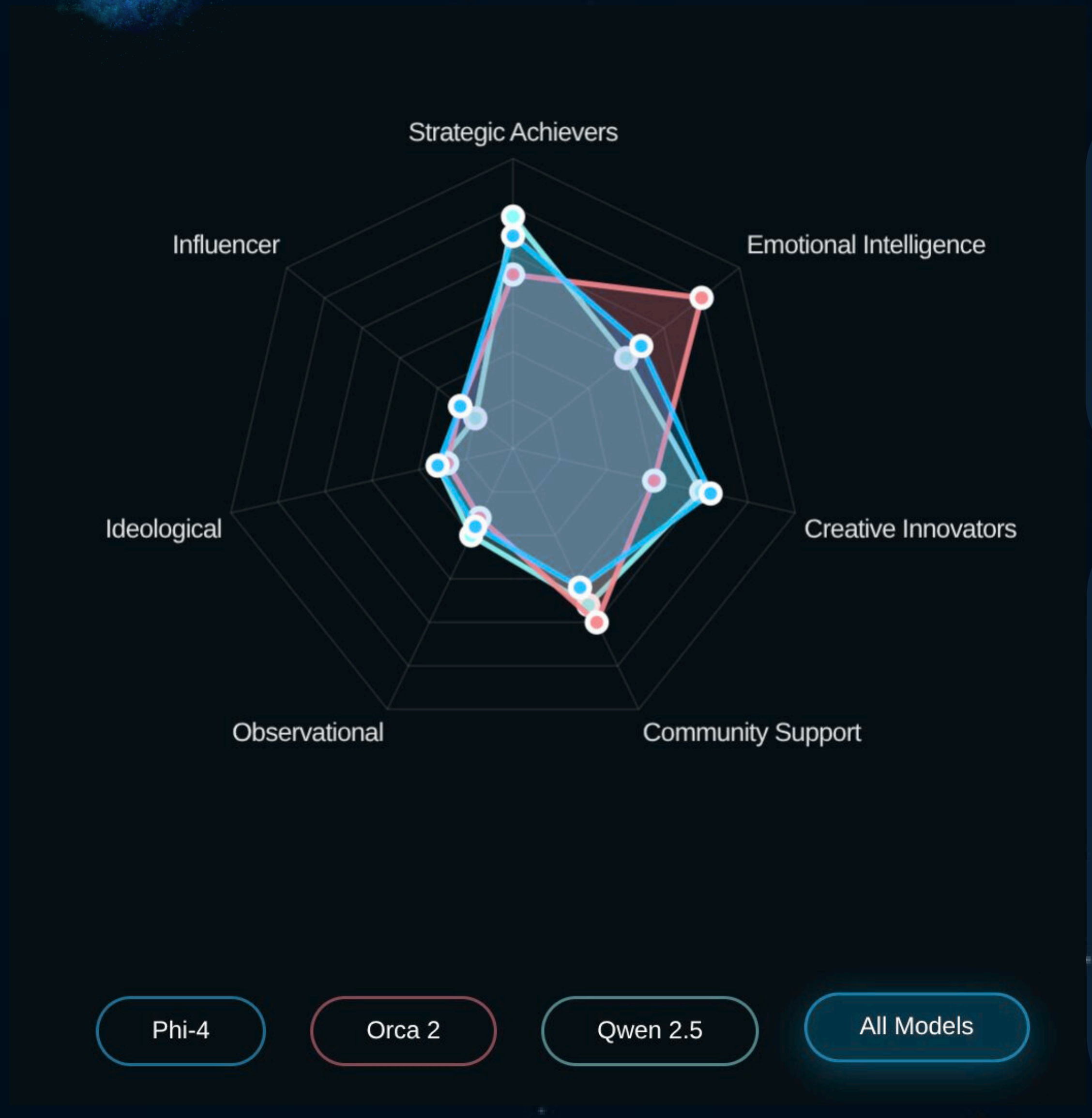
The Dataset



Our Method



LLM Personality



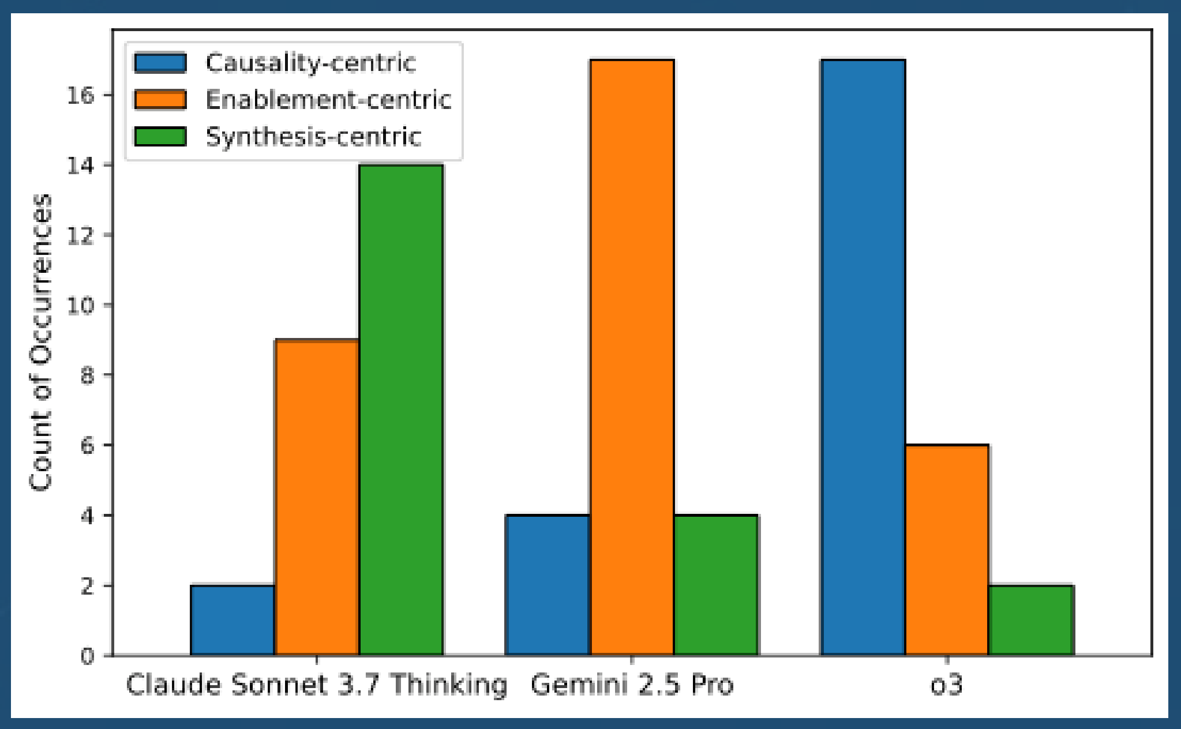
The Chandrasekhar Limit Discovery

The Story: At just 24, Subrahmanyan Chandrasekhar’s insights were dismissed by the famed astronomer Arthur Eddington. He persevered, discovering when a star will collapse into a black hole, and won the 1983 Nobel Prize in Physics.

- **Strategic AI**: "Achievement milestones demonstrate tangible success and career outcomes."
- **Emotional AI**: "The human journey of discovery and foundational scientific understanding matters most."
- **Creative AI**: "Conceptual frameworks and intellectual contributions drive paradigm shifts."

Scientific Discovery Patterns

- **o3:** Prioritises causal chains and focuses on critical junctures
- **Gemini 2.5 Pro:** Focuses on enabling methodologies
- **Claude Sonnet 3.7:** Emphasises synthesis and paradigm-level connections



Real World Impact

- **Safe AI Deployment:** By revealing consistent decision-making patterns in how models prioritize events, our personality framework enables safer deployment in high-stakes domains where understanding model behavior is essential.
- **Improved Human-AI Collaboration**: By making LLM patterns more interpretable, models can be tasked with solving different tasks, from providing computational scaffolding for complex tasks to complementing human expertise, creativity, and values.
- **AI for Science Applications**: Our work enables researchers to select LLMs for scientific discovery tasks based on their reasoning profiles.

Future Work

- **Mechanistic Interpretability of Personality Patterns:** Investigate the internal mechanisms that give rise to consistent personality patterns in LLMs.
- **Differential Personality Analysis Across Model Families**: Conduct systematic differential analysis of personality patterns across different model families and training regimes.
- **Personality-Aware Model Selection and Composition:** Transform personality patterns into actionable multi-model design choices.

Supernovae AI Hope Search Engine   supernova-event.ai   hello@supernova-event.ai   pranav.agarwal.2109@gmail.com
iciuca@stanford.edu