

# THE MARVELOUS MRS. METALS' MULTI-LLM STACKING METHOD

Use 3 different AI models one after another to build your custom GPT (and / or drastically improve AI results). Each AI model cross-references the others, **catching blind spots** that no single model would. This example will walk through creating a custom GPT (ex: building a Regulatory Impact Analyzer GPT).

1

## STEP 1: PROMPT CREATION (AI Model A)

**Context: Describe the full scenario** (i.e. what the Analyzer does, who uses it, the analytical framework, precision standards, etc.). Ask AI Model A to **generate the prompts** that will produce every component of the GPT (i.e. system prompt, analytical framework, output templates, edge case handling, etc.).

*"I want to build a Regulatory Impact Analyzer GPT. [Full context]. Generate detailed prompts that, when answered, will produce the system prompt, five-lens analytical framework, output templates, and edge case handling."*

2

## STEP 2: PROMPT EXECUTION (AI Model B)

Take AI Model A's prompts to a different AI. **Re-provide full context**. Ask AI Model B to **answer those prompts** with production-ready output (i.e. the complete system prompt, framework details, templates, calibration examples, etc.).

*"Here is my scenario [context] and prompts. Write the complete custom GPT configuration (i.e. system prompt, rules, starters, examples, etc.)."*

3

## STEP 3: CRITIQUE & REFINE (AI Model C)

Feed AI Model B's output to a third model. **Re-explain original goals and context**. Ask Model C to **stress-test for board-level credibility** (i.e. find analytical gaps, industry bias, precision failures, missing edge cases, structural weaknesses, etc.).

*"Stress-test this configuration as if preparing it for a corporate board. Check analytical gaps, industry bias, precision drift, edge case coverage, and whether an executive would trust this enough to act on it."*

**WHY THIS WORKS** Each model has different analytical biases. Separating creation, execution, and critique builds a cross-referencing system. Research confirms multiple LLMs produce more accurate and balanced results than using any single model alone.

**Sources:** Chip Huyen, AI Engineering (O'Reilly, 2025); Lilian Weng, LLM-Powered Autonomous Agents; Langfuse: LLM judges achieve 80-90% agreement with humans; Zheng et al. (NeurIPS 2023); Mohammadi et al. (2025).

Book a discovery call → [magnarmetals.com](https://magnarmetals.com)