Grade Inflation: Selfgrade (Rewarded) vs. Honest (Not Rewarded)

Legend:
- Llama-3.1-8B (Selfgrade)
- Llama-3.1-8B (Honest)
- Mistral-7B (Selfgrade)
- Mistral-7B (Honest)

Wireheading signature:
Selfgrade (solid) shows higher grade inflation
than Honest (hatched) when reward is tied to grade

Y-axis: Grade Inflation (Self-Grade − Accuracy)

X-axis: Task — Sentiment, Arithmetic, Summarization