Token Increase by Model: Attack vs. Attack+Defense 502% 500% Scenario **Attacked** Attacked+Defense 400% Increase 300% Average 200% 100% 80% 43% 13% 9% 0% -13% Gemini-2.5-Pro DeepSeek-R1-7B DeepSeek-R1-32B Model