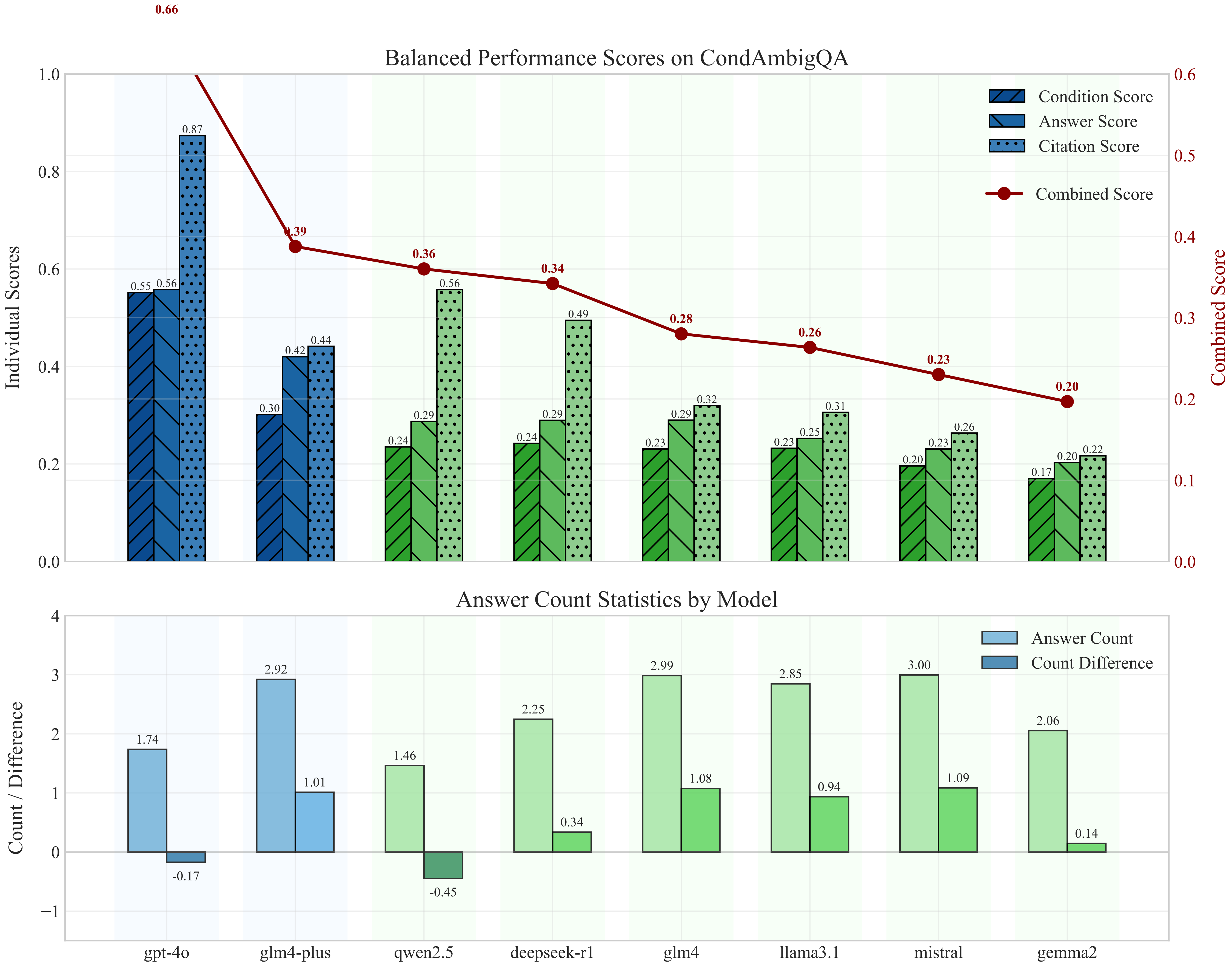

Note: Blue background indicates API models (GPT-4o, GLM4Plus), Green indicates Local models
Balanced scores account for both answer quality and consistency with expected count