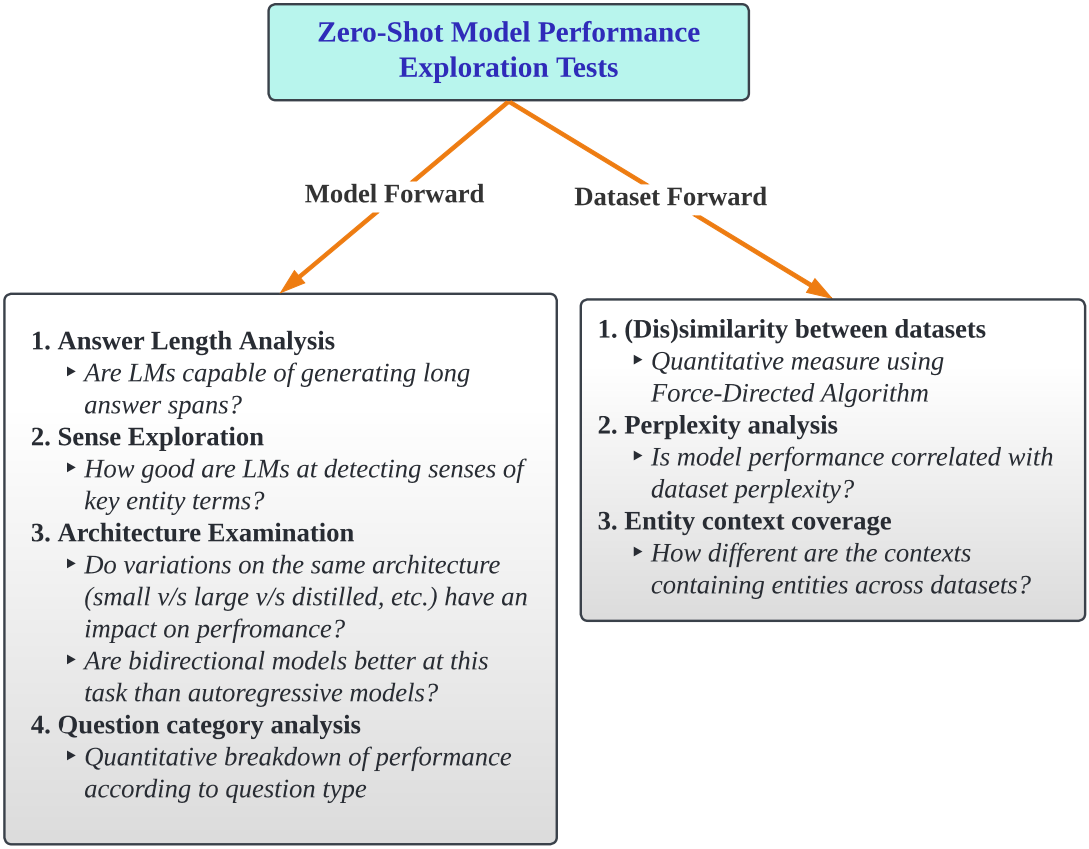
Zero-Shot Model Performance Exploration Tests
Model Forward
Dataset Forward
1. Answer Length Analysis
Are LMs capable of generating long answer spans?
2. Sense Exploration
How good are LMs at detecting senses of key entity terms?
3. Architecture Examination
Do variations on the same architecture (small v/s large v/s distilled, etc.) have an impact on perfromance?
Are bidirectional models better at this task than autoregressive models?
4. Question category analysis
Quantitative breakdown of performance according to question type
1. (Dis)similarity between datasets
Quantitative measure using Force-Directed Algorithm
2. Perplexity analysis
Is model performance correlated with dataset perplexity?
3. Entity context coverage
How different are the contexts containing entities across datasets?