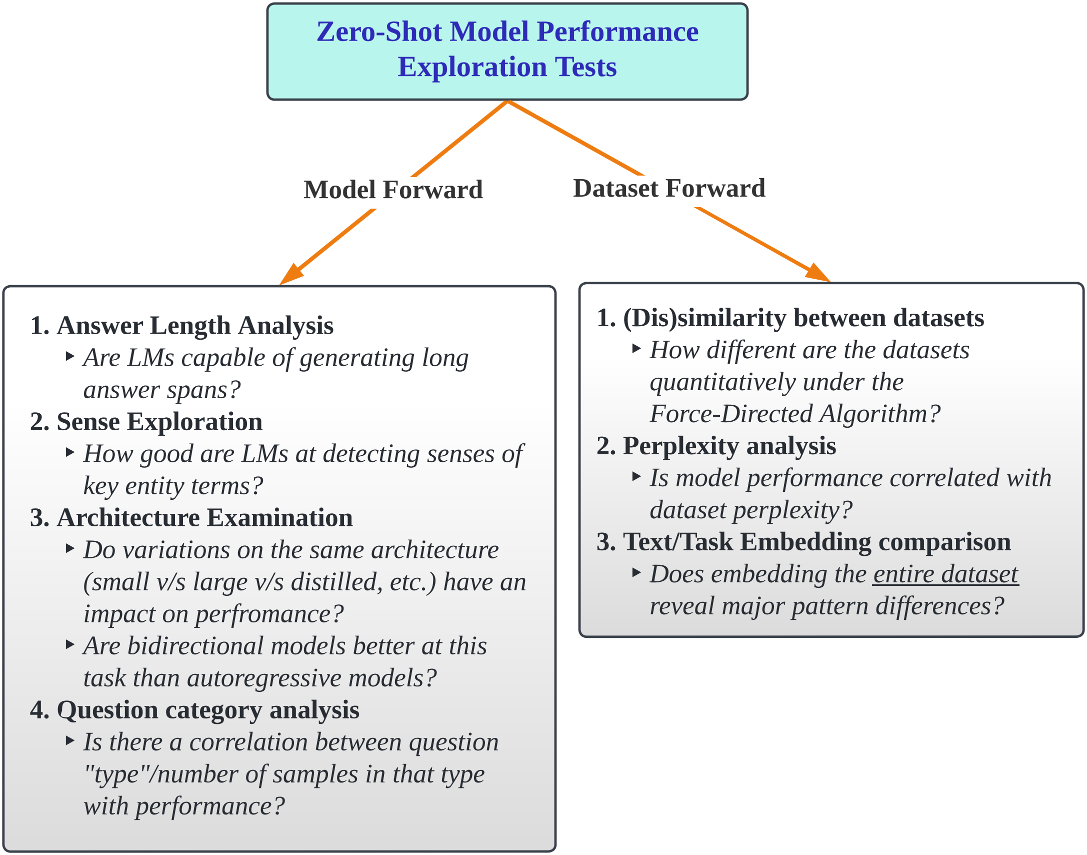

Zero-Shot Model Performance Exploration Tests

Model 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**
   - *Is there a correlation between question "type"/number of samples in that type with performance?*

Dataset Forward

1. **(Dis)similarity between datasets**
   - *How different are the datasets quantitatively under the Force-Directed Algorithm?*
2. **Perplexity analysis**
   - *Is model performance correlated with dataset perplexity?*
3. **Text/Task Embedding comparison**
   - *Does embedding the entire dataset reveal major pattern differences?*