The SuperGLUE Benchmarks

Lesson Downloads

SuperGlue is designed as a successor to the original GLUE benchmark. It's a more advanced benchmark aimed at presenting even more challenging language understanding tasks for AI models. Created to push the boundaries of what AI can understand and process in natural language, SuperGlue emerged as models began to achieve human parity on the GLUE benchmark. It also features a public leaderboard, facilitating the direct comparison of models and enabling the tracking of progress over time.

SuperGLUE Tasks / Benchmarks:

Short Name	**Full Name**	**Description**
BoolQ	Boolean Questions	Involves answering a yes/no question based on a short passage.
CB	CommitmentBank	Tests understanding of entailment and contradiction in a three-sentence format.
COPA	Choice of Plausible Alternatives	Measures causal reasoning by asking for the cause/effect of a given sentence.
MultiRC	Multi-Sentence Reading Comprehension	Involves answering questions about a paragraph where each question may have multiple correct answers.
ReCoRD	Reading Comprehension with Commonsense Reasoning	Requires selecting the correct named entity from a passage to fill in the blank of a question.
RTE	Recognizing Textual Entailment	Involves identifying whether a sentence entails, contradicts, or is neutral towards another sentence.
WiC	Words in Context	Tests understanding of word sense disambiguation in different contexts.
WSC	Winograd Schema Challenge	Focuses on resolving coreference resolution within a sentence, often requiring commonsense reasoning.
AX-b	Broad Coverage Diagnostic	A diagnostic set to evaluate model performance on a broad range of linguistic phenomena.
AX-g	Winogender Schema Diagnostics	Tests for the presence of gender bias in automated coreference resolution systems.

Technical Terms Explained:

Coreference Resolution: This is figuring out when different words or phrases in a text, like the pronoun *she* and the *president*, refer to the same person or thing.

BoolQ Examples

Let's take a look at some examples from the BoolQ dataset. Here is a table from the paper "BoolQ: Exploring the surprising difficulty of natural yes/no questions." [1]

[1] Clark, Christopher, et al. "BoolQ: Exploring the surprising difficulty of natural yes/no questions." arXiv preprint arXiv:1905.10044 (2019).

Quiz Question

What does the WiC task assess in lang models?

- Spelling accuracy
- Ability to generate text
- Understanding of figures of speec
- Word sense disambiguation
- Pronunciation consistency

Submit