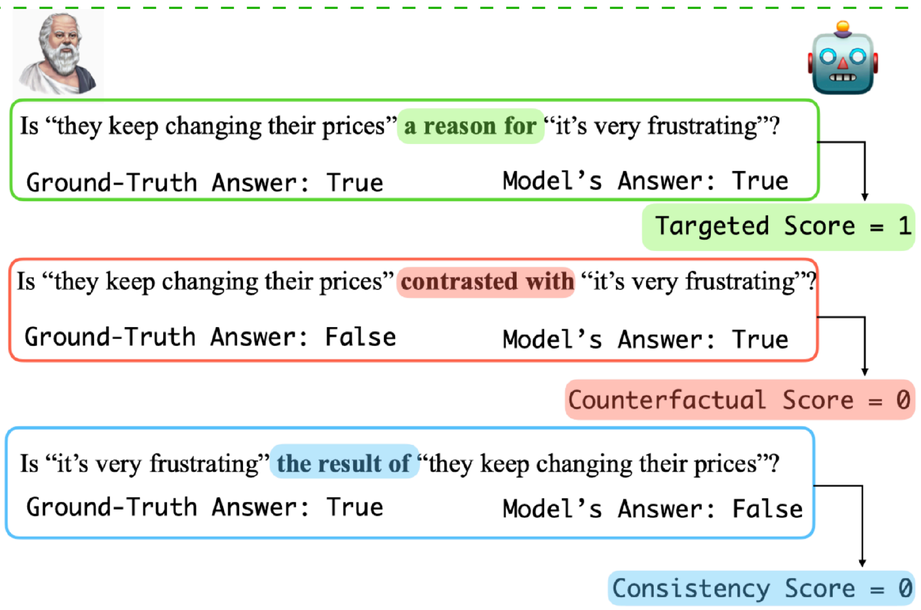
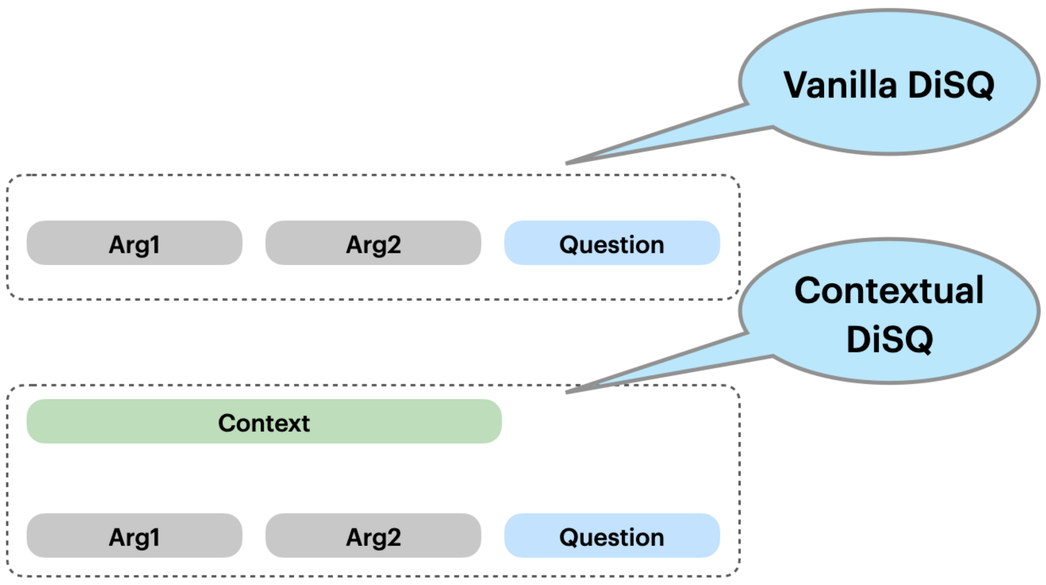
Contexts Ground Discourse

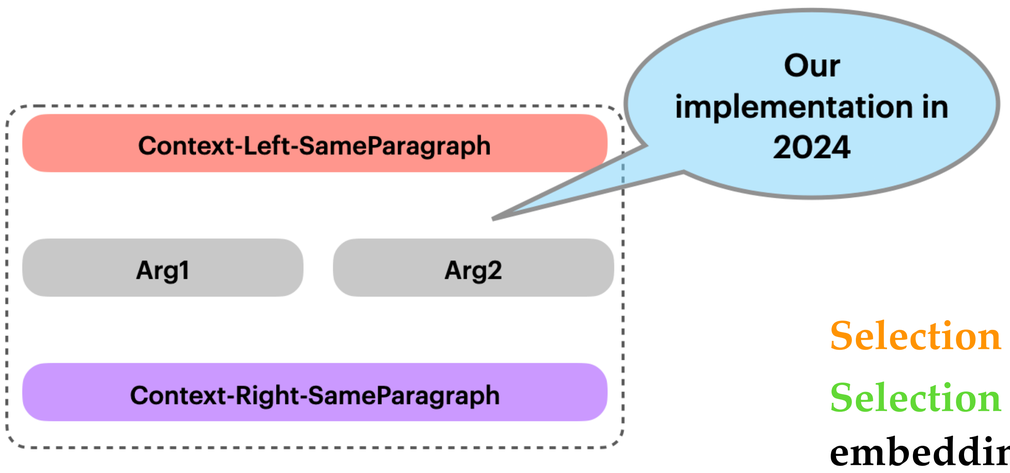
CS6101 Project 11 — Yisong Miao



Background 1: Discourse relations describe the logical relations among sentences. In 2024, our team proposes a new measure, named Discursive Socratic Questioning (DiSQ) [1] to evaluate models' faithfulness.



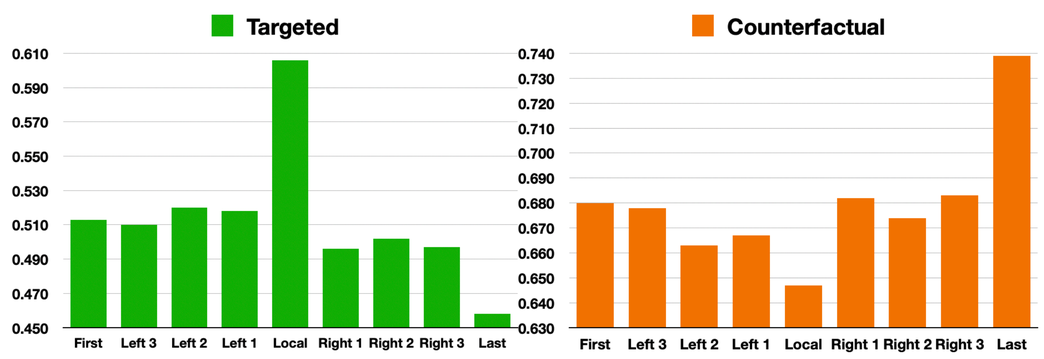
Background 2: In our prior work (2024), we found contexts boost discourse faithfulness. 🚀 However, our context selection was straightforward (local paragraph).



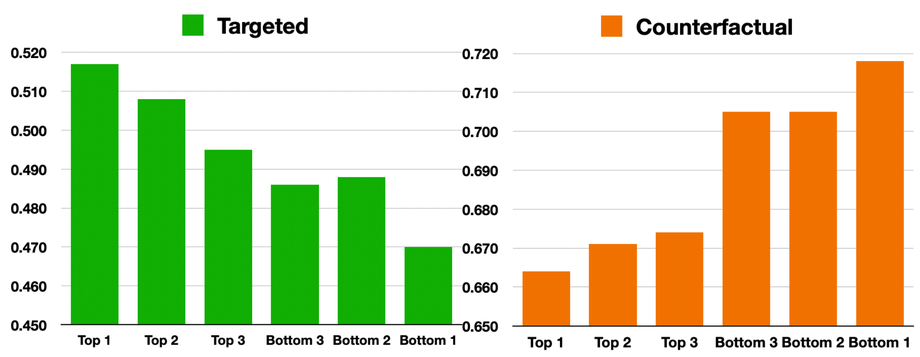
Selection 1: Rule-based. Choose left / right / first / last paragraph in WSJ articles.

Selection 2: Similarity-based. Using Sentence-BERT, rank paragraphs by the embedding similarity between <(1) arg1+arg2, (2) a given paragraph>.

	DiSQ Score(s)				Overall DiSQ Scores per Discourse Relations										
	Overall	Targeted	CF	Consist-ency	Comp. Conc.	Comp. Contrast	Cont. Reason	Cont. Result	Exp. Conj.	Exp. Equiv.	Exp. Instan.	Exp. Detail	Exp. Subst.	Temp. Async.	Temp. Sync.
Vanilla	0.253	**0.591**	0.544	0.786	**0.193**	**0.487**	0.129	0.172	0.29	0.155	0.328	0.372	**0.291**	**0.194**	**0.028**
Most Similar (MS)	0.268	0.507	0.670	0.789	**0.111**	0.350	**0.157**	0.199	0.320	0.188	0.374	0.433	0.217	0.137	0.005
Least Similar (LS)	**0.274**	0.481	0.709	**0.800**	0.083	0.355	0.145	0.201	**0.342**	**0.194**	0.373	**0.465**	0.226	0.107	0.012
Left Context (LEC)	0.273	**0.516**	0.669	0.789	**0.100**	0.345	**0.157**	**0.206**	0.333	0.180	0.372	0.443	0.243	0.140	0.008
Right Context (RC)	0.270	0.498	**0.680**	**0.796**	0.095	0.381	0.153	0.193	0.326	0.188	**0.376**	0.453	**0.249**	0.120	0.010
First Paragraph (FP)	**0.276**	0.513	**0.680**	0.791	0.068	0.343	**0.157**	**0.219**	0.331	**0.201**	**0.393**	0.455	0.172	**0.149**	**0.015**
Last Paragraph (LP)	0.272	0.458	**0.739**	**0.804**	0.035	**0.386**	0.154	0.192	**0.345**	0.174	0.375	**0.47**	0.214	0.079	**0.015**
Local Context (LC)	**0.31**	**0.606**	0.647	0.791	0.088	**0.407**	**0.202**	**0.274**	**0.383**	**0.227**	**0.394**	**0.459**	**0.252**	**0.217**	**0.031**



Analysis 1: Rule-based. Local context is still the best. The other contexts are performing similarly, while the last paragraph produces the best CF score.



Analysis 2: Similarity-based. It's like a seesaw. If Targeted score decreases, then CF score increases. The more similar, the higher the targeted score.

Reference: [1] Discursive Socratic Questioning: Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations. ACL '24.