I want a horror – comedy – movie: Slips-of-the-Tongue Impact Conversational Recommender System Performance

Maria Teleki, Lingfeng Shi, Chengkai Liu, James Caverlee
Texas A&M University

1 We synthetically inject slip-of-the-tongue speech errors via our psycholinguistically-grounded Syn-WSSE Framework.

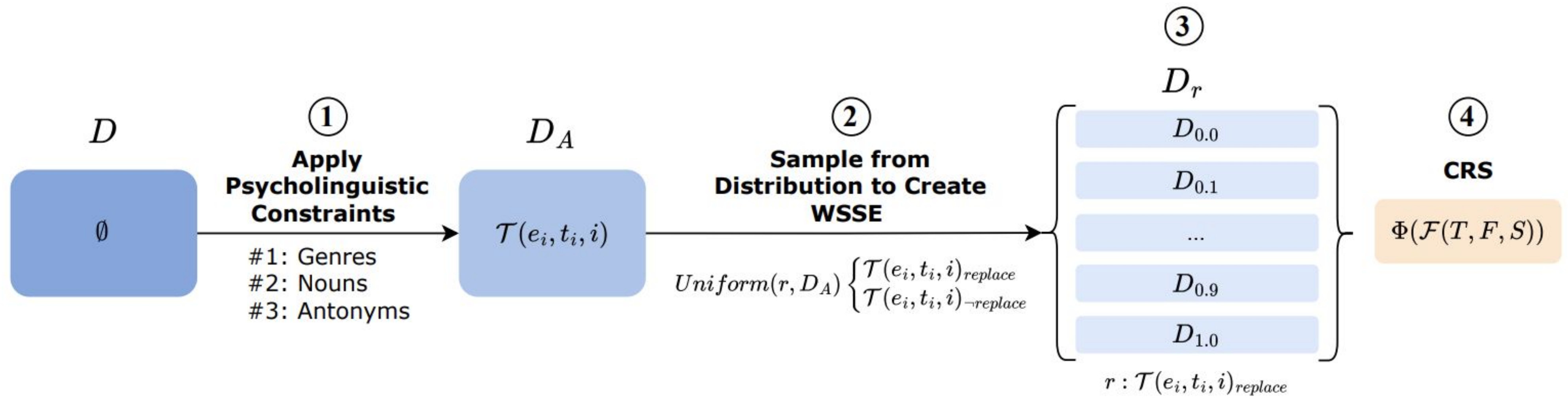

Figure 1: *Syn-WSSE Framework:* ① *We apply psycholinguistic constraints to create the Candidate-WSSEs – i.e.,* $\mathcal{T}(e_i, t_i, i)$ *– in* D_A*.* ② *We draw r percent of samples uniformly from the list of Candidate-WSSEs tokens to create WSSE.* ③ D_r *are created and Syn-WSSE is complete.* ④ *We evaluate the performance of the CRS (§3.2) on each* D_r*, shown in Tables 1, 2, and 3.*

2 We study the impact of these errors on the LLM-based Conversational Recommender System task.

We use a prompt to elicit recommendations from a set of backbone LLMs, and score the recommendations using these metrics:

Mean NDCG@k $= \mathbb{E}_c \left[\mathbb{I}\{g_c \in (L_c)_0^k\} \cdot \frac{1}{\log_2(r(g_C)+1)} \right]$

This metric aggregates the ranked presence of g_c in $(L_c)_0^k$, discounting lower-ranked g_c occurrences.

NDCG@5							
	NDCG@$5_{0.0}$	$\Delta_{0.1}$	$\Delta_{0.2}$	$\Delta_{0.3}$	$\Delta_{0.5}$	$\Delta_{0.7}$	$\Delta_{1.0}$
llama	0.076	17.58	16.77	7.04	13.65	7.43	12.57
mixtral	0.051	6.81	7.63	2.63	6.40	2.92	8.95
gemini	0.056	-8.21	-13.00	-11.90	-17.14	-1.94	-6.48
gpt-4o	**0.105**	-13.20	-13.71	-6.87	-12.36	-16.80	-14.02
gpt-4o-mini	0.083	3.85	-4.86	-4.71	-20.74	-12.99	-25.06
NDCG@10							
	NDCG@$10_{0.0}$	$\Delta_{0.1}$	$\Delta_{0.2}$	$\Delta_{0.3}$	$\Delta_{0.5}$	$\Delta_{0.7}$	$\Delta_{1.0}$
llama	0.091	9.70	17.00	9.31	13.21	1.34	12.07
mixtral	0.068	0.19	-1.56	1.75	2.86	8.11	15.40
gemini	0.068	-2.27	-8.80	-11.90	-3.84	-1.50	-5.72
gpt-4o	**0.126**	-9.73	-10.44	-6.53	-15.10	-18.38	-16.37
gpt-4o-mini	0.105	1.29	-6.82	-8.43	-24.80	-16.95	-28.20

Table 2: ***Best***, *Worst; Most Resilient , Least Resilient;* Δ_r *is percent change in NDCG@k for* $D = r$ *and* $D = 0.0$*.*

Mean Recall@k $= \mathbb{E}_c \left[\mathbb{I}\{g_c \in (L_c)_0^k\} \right]$

This metric aggregates the presence of g_c in $(L_c)_0^k$.

RECALL@5							
	RECALL@$5_{0.0}$	$\Delta_{0.1}$	$\Delta_{0.2}$	$\Delta_{0.3}$	$\Delta_{0.5}$	$\Delta_{0.7}$	$\Delta_{1.0}$
llama	0.104	18.18	18.18	4.55	13.64	9.09	13.64
mixtral	0.081	11.76	11.76	0.00	5.88	5.88	5.88
gemini	0.090	-10.53	-10.53	-10.53	-21.05	-10.53	-5.26
gpt-4o	**0.133**	-14.29	-10.71	0.00	-7.14	-17.86	-7.14
gpt-4o-mini	0.114	0.00	-4.17	-4.17	-16.67	-16.67	-25.00
RECALL@10							
	RECALL@$10_{0.0}$	$\Delta_{0.1}$	$\Delta_{0.2}$	$\Delta_{0.3}$	$\Delta_{0.5}$	$\Delta_{0.7}$	$\Delta_{1.0}$
llama	0.152	3.13	18.75	9.38	12.50	-3.13	12.50
mixtral	0.133	0.00	-3.57	0.00	0.00	0.00	0.00
gemini	0.128	0.00	-3.70	-11.11	3.70	-7.41	-3.70
gpt-4o	**0.199**	-7.14	-4.76	-2.38	-14.29	-21.43	-14.29
gpt-4o-mini	0.185	-2.56	-7.69	-10.26	-25.64	-23.08	-30.77

Table 1: ***Best***, *Worst; Most Resilient , Least Resilient;* Δ_r *is percent change in Recall@k for* $D = r$ *and* $D = 0.0$*.*

3 We find that LLMs respond very differently to these speech errors, and hypothesize it may be due to differences in their synthetic pretraining processes.

These findings indicate that the choice of backbone LLM is a critical design decision for real-world applications.