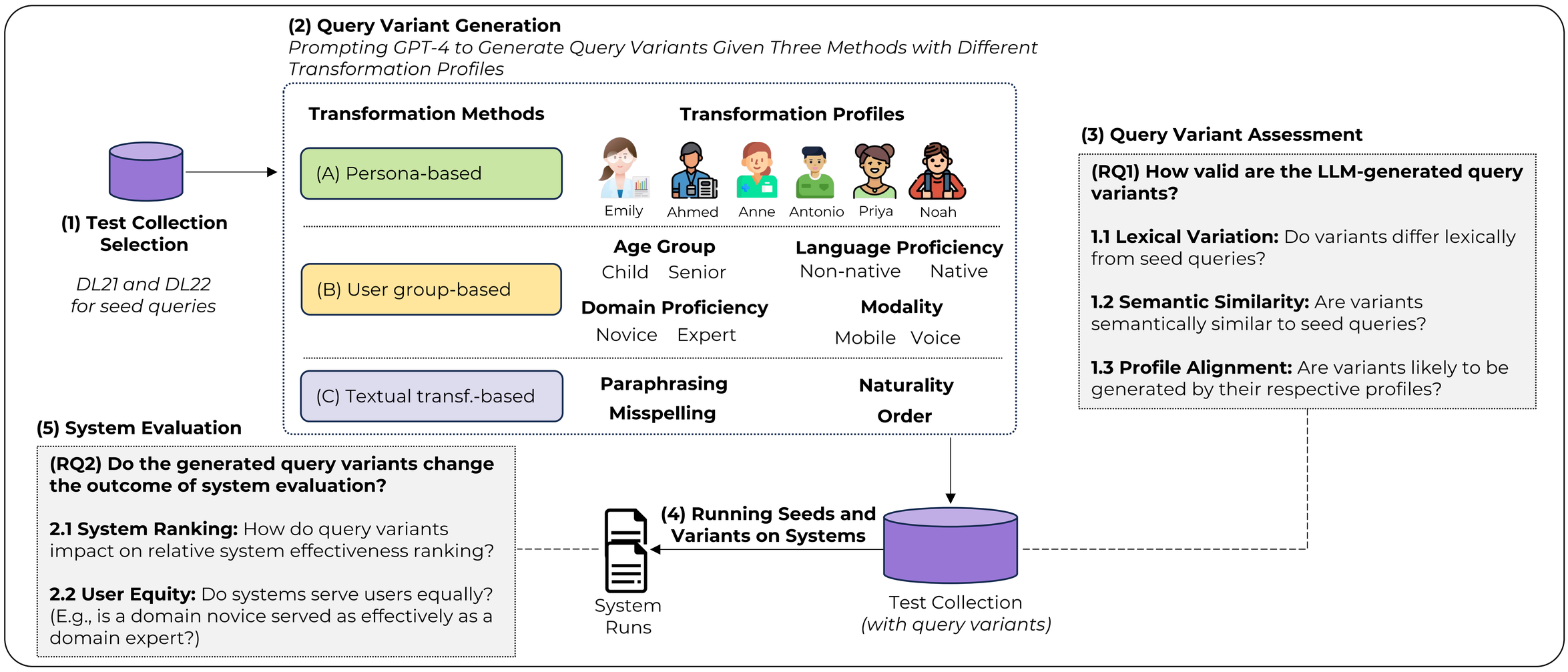
Demographically-Inspired Query Variants Using an LLM

Marwah Alaofi [1, 4], Nicola Ferro[2], Paul Thomas[3], Falk Scholer[1, 4], Mark Sanderson[1, 4]

[1]RMIT University, [2]University of Padua, [3]Microsoft, [4]ARC Centre of Excellence for Automated Decision Making + Society

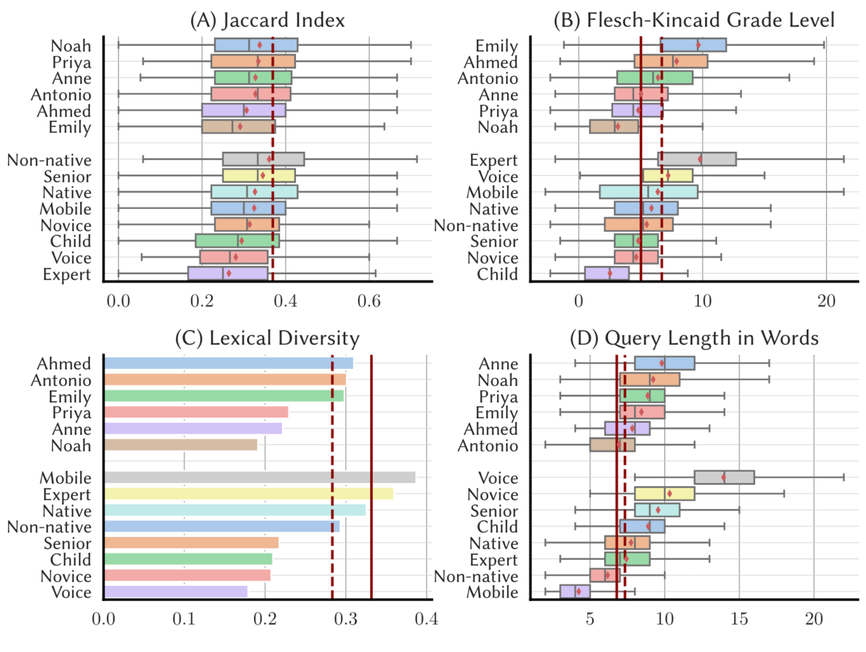
(1) Experiment Pipeline and Research Questions



(2) Results

(RQ1) Query Variant Assessment

(1.1) Automatically generated features for *lexical similarity* and *profile alignment*



The generated variants are different and seem to reflect their demographics.

(1.2) Human assessment for *semantic similarity* and *profile alignment*

	Profile	Similarity	Profile Alignment
Persona	Emily	0.97	0.87
	Ahmed	0.95	0.68
	Anne	0.92	0.70
	Antonio	1.00	0.62
	Priya	0.97	0.55
	Noah	0.92	0.73
Group	Child	0.67	0.74
	Senior	0.82	0.85
	Non-native	0.92	0.62
	Native	0.84	0.84
	Novice	0.84	0.70
	Expert	0.72	0.79
	Mobile	0.92	0.95
	Voice	0.97	1.00
Textual	Naturality	0.97	0.97
	Paraphrasing	0.76	0.76
	Order	-	0.69
	Misspelling	-	0.80

Variants are topical and reasonably aligned with demographics.

(RQ2) System Evaluation

(2.1) System Ranking

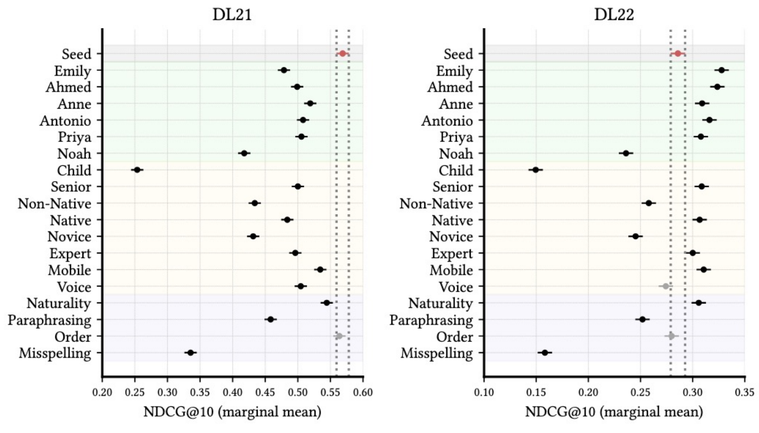
		τ	AA	AD	MA	MD	PA	PD
DL21	(A) Persona	0.74	0.48	0.00	0.01	0.00	0.38	0.13
	(B) Group	0.71	0.48	0.00	0.02	0.01	0.36	0.14
	(C) Textual	0.76	0.48	0.00	0.11	0.02	0.30	0.10
DL22	(A) Persona	0.77	0.50	0.00	0.02	0.00	0.36	0.12
	(B) Group	0.69	0.50	0.00	0.03	0.00	0.32	0.15
	(C) Textual	0.69	0.50	0.00	0.03	0.00	0.32	0.15

Kendall's τ and **A**ctive, **P**assive, and **M**ixed Agreement (AA, PA, MA) and **D**isagreement (AD, PD, MD)

Source	DL21	DL22
Topic	0.7338	0.7741
System	0.4439	0.3926
Profile	0.3300	0.2189
Topic*System	0.3149	0.3470
Topic*Profile	0.6713	0.6566
System*Profile	0.0104	0.0156

ANOVA effect size: Dark = large (≥ 0.14), Light = small (≤ 0.06)

(2.2) User Equity



Variants provide new evaluation perspectives.

(3) Opportunities and Challenges

- An opportunity to expand queries in existing test collections to capture user variability.
- However, query validation is a challenge:
  - Do LLMs generate realistic queries?
  - Do they enforce some stereotypes?
- Query variants can be difficult to evaluate due to relevance judgment holes.
- As queries vary, so does relevance.

Acknowledgment: *Marwah Alaofi is supported by a scholarship from Taibah University, Saudi Arabia. This research was also supported in part by the CAMEO PRIN 2022 Project, funded by the Italian Ministry of Education and Research (2022ZLL7MW) and the Australian Research Council (DP190101113). We thank the anonymous reviewers for their helpful feedback.*

ICTIR'25