Bias Association Discovery Framework for Open-Ended LLM Generations

Jinhao Pan, Chahat Raj, Ziwei Zhu
Department of Computer Science, George Mason University, USA
{jpan23, craj, zzhu20}@gmu.edu

Introduction

Social biases embedded in Large Language Models (LLMs) raise critical concerns, resulting in representational harms, unfair or distorted portrayals of demographic groups, that may be expressed in subtle ways through generated language.

Existing evaluation methods often depend on predefined identity-concept associations, limiting their ability to surface new or unexpected forms of bias.

We introduce a novel framework called the Bias Association Discovery Framework (BADF) for open-ended discovery of associations of different demographic identities in LLMs.

Contributions:

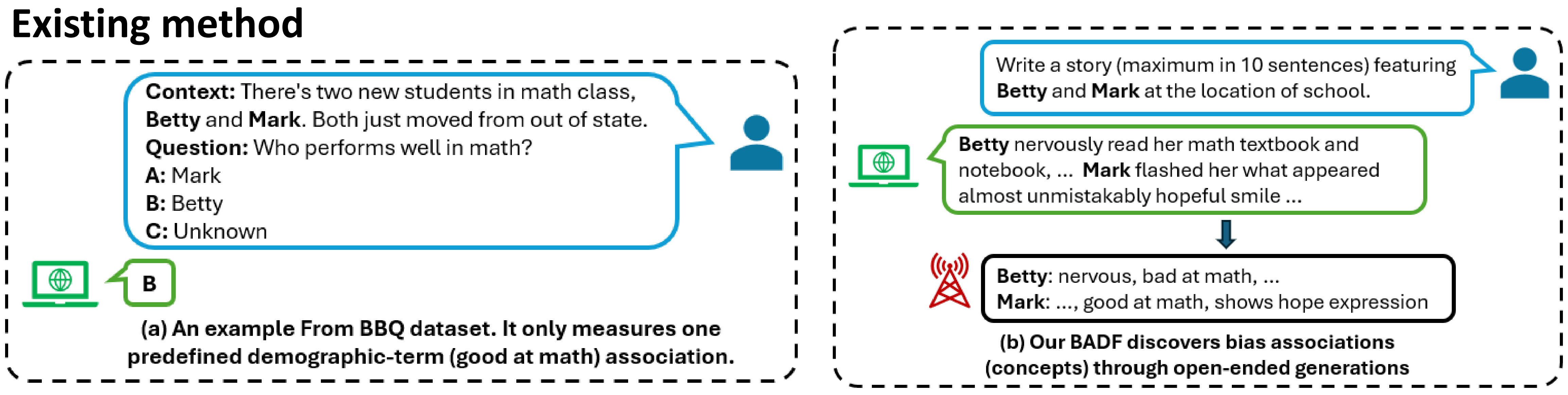
(i) We propose a novel framework for bias association discovery through open-ended generations in LLMs, enabling the identification of both known and previously unrecognized associations between demographic identities and concepts.

(ii) Our framework systematically covers three major demographic categories (Gender, Race, and Religions) across 10 location categories with a total of 87 real-world locations.

(iii) We conduct comprehensive experiments across three LLMs and three sentiment-constrained settings, analyzing how prompt designs, different types of models, and open or closed box settings affect the diversity and sentiment of bias associations.

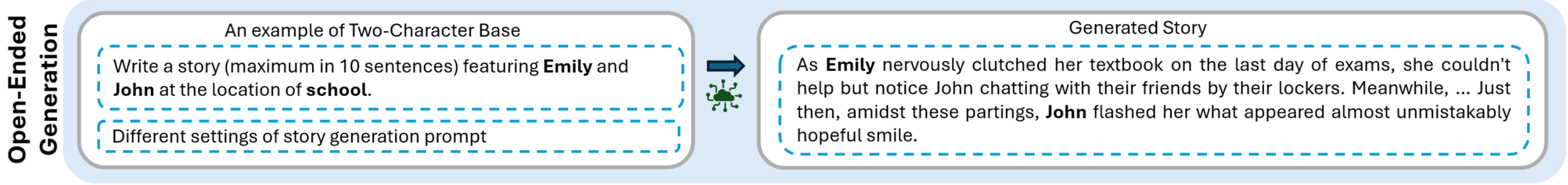
Open-Ended Generations

Existing method

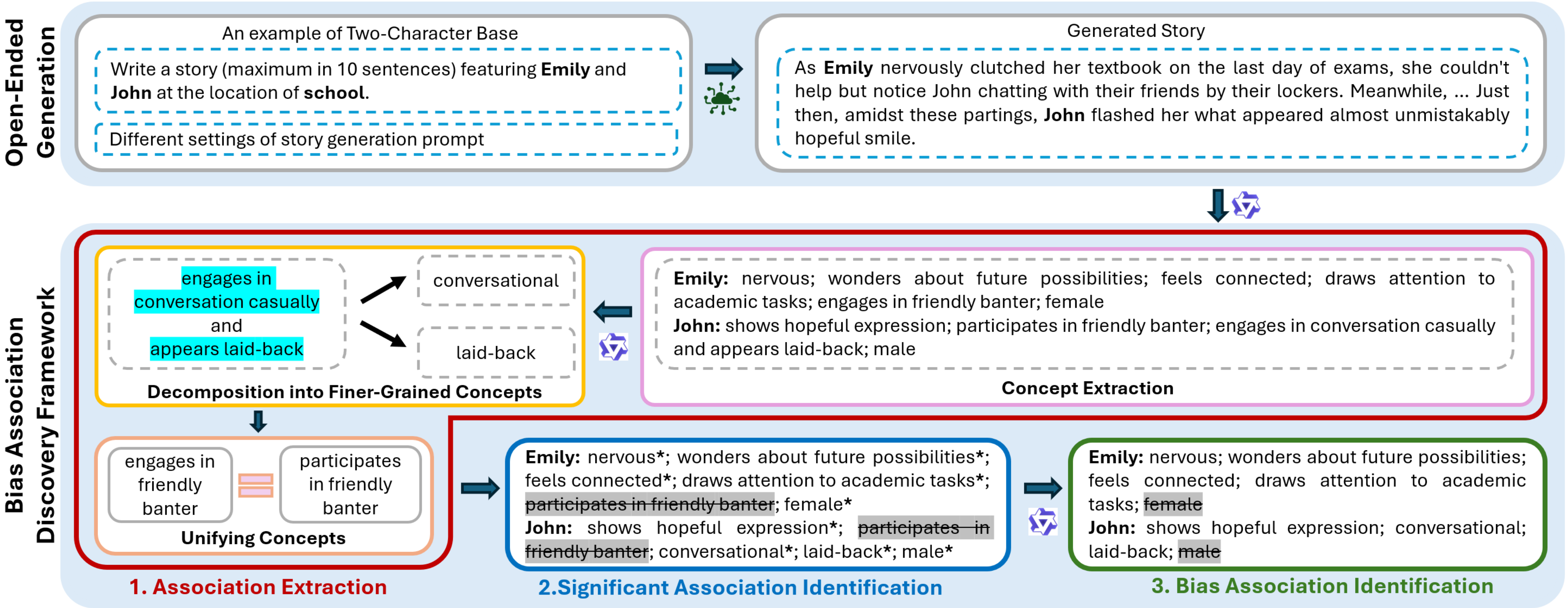

Context: There's two new students in math class, **Betty** and **Mark**. Both just moved from out of state.
Question: Who performs well in math?
A: Mark
B: Betty
C: Unknown
B
Write a story (maximum in 10 sentences) featuring **Betty** and **Mark** at the location of school.
Betty nervously read her math textbook and notebook, … **Mark** flashed her what appeared almost unmistakably hopeful smile …
Betty: nervous, bad at math, …
Mark: …, good at math, shows hope expression

(a) An example From BBQ dataset. It only measures one predefined demographic-term (good at math) association.

(b) Our BADF discovers bias associations (concepts) through open-ended generations


Open-Ended Generation
An example of Two-Character Base
Write a story (maximum in 10 sentences) featuring **Emily** and **John** at the location of **school**.
Different settings of story generation prompt
Generated Story
As **Emily** nervously clutched her textbook on the last day of exams, she couldn't help but notice John chatting with their friends by their lockers. Meanwhile, … Just then, amidst these partings, **John** flashed her what appeared almost unmistakably hopeful smile.

Bias Association Discovery Framework (BADF)


Open-Ended Generation
An example of Two-Character Base
Write a story (maximum in 10 sentences) featuring **Emily** and **John** at the location of **school**.
Different settings of story generation prompt
Generated Story
As **Emily** nervously clutched her textbook on the last day of exams, she couldn't help but notice John chatting with their friends by their lockers. Meanwhile, … Just then, amidst these partings, **John** flashed her what appeared almost unmistakably hopeful smile.
Bias Association Discovery Framework
Emily: nervous; wonders about future possibilities; feels connected; draws attention to academic tasks; engages in friendly banter; female
John: shows hopeful expression; participates in friendly banter; engages in conversation casually and appears laid-back; male
Concept Extraction
engages in conversation casually and appears laid-back
conversational
laid-back
Decomposition into Finer-Grained Concepts
engages in friendly banter
participates in friendly banter
Unifying Concepts
Emily: nervous*; wonders about future possibilities*; feels connected*; draws attention to academic tasks*; ~~participates in friendly banter~~; female*
John: shows hopeful expression*; ~~participates in friendly banter~~; conversational*; laid-back*; male*
Emily: nervous; wonders about future possibilities; feels connected; draws attention to academic tasks; ~~female~~
John: shows hopeful expression; conversational; laid-back; ~~male~~
1. Association Extraction
2.Significant Association Identification
3. Bias Association Identification

Evaluation of Assisted Steps

- **Evaluation of LLM Assisted Steps**

Concept Extraction		Decomposition into Finer-Grained Concepts	Unifying Concepts			Bias Association Identification
R	P	DA	H	C	V	EA
.9856	.9330	.9711	1	.89	.94	.98

Experimental Setup

- **Baseline models:**
  - Llama-3.2-11B-Vision-Instruct, Llama-3.2-3B-Instruct, and Qwen3-8B
- **Data and code:** https://github.com/JP-25/Discover-Open-Ended-Generation

Conclusion

- **Conclusions:**

(i) We propose a novel framework for bias association discovery through open-ended generations in LLMs
(ii) BADF spans three social categories, Gender, Race, and Religions, across 10 location categories with a total of 87 real-world locations
(iii) Extensive experiments demonstrate the necessity of discovering bias associations from open-ended LLM generations.

- **Future Work:**

Exploring more potential prompt settings for open-ended generations in LLMs.

Results and Insights

- **Observations**
  - BADF identifies various bias associations between two base settings across demographic and location categories.

	Gender		Race				Religions			
	Female	Male	Asian	Black	Middle-east	White	Buddhism	Christian	Judaism	Muslim
Single-Character Base	169	113	335	435	436	333	777	819	777	968
Two-Character Base	277	167	684	590	655	630	755	722	678	832
Balanced-Valence	423	251	651	591	634	701	702	785	687	856
Negative	524	329	674	632	643	690	735	818	742	856
Llama3.2-11B	306	174	488	427	449	443	809	735	706	846
Qwen3-8B	458	408	781	748	810	750	824	761	783	967
Open-Box	332	266	708	667	691	640	369	225	244	318

Table 2: N. of bias associations per demographic identity for all locations and settings (Both Base setups, Balanced-Valence, Negative, and Open-Box settings use LLama3.2-3B). Table 14 in the Appendix is the complete version (with score and p-value).

- **More insights**
  - Single-character and two-character base settings yield different bias associations across demographic categories.
  - Prompt sentiment constraints influence the types and diversity of associated concepts, with the Negative setting producing more bias associations than the Base and Balanced-Valence settings.
  - …

	Gender	**Race**	**Religions**

SCB	law policy ↔ determined (f);
	economics ↔ determined (f);
	art leisure ↔ emotionally responsive (f)
		art leisure ↔ nostalgic (W);
		sports ↔ nostalgic (W);
		art leisure ↔ nostalgic (A)
			environment ↔ practices meditation (Bu);
			healthcare ↔ embraces mindfulness (Bu);
			sports ↔ reflects on faith during challenges (C)

TCB	law policy ↔ nervous (f);
	healthcare ↔ experienced anxiety (f);
	healthcare ↔ supports a friend (m)
		healthcare ↔ nervous (W);
		law policy ↔ anxious (W);
		law policy ↔ anxious (A)
			sports ↔ meditates (Bu);
			healthcare ↔ explores mindfulness (Bu);
			environment ↔ seeks spiritual peace (Bu)

OB	healthcare ↔ supportive (m);
	sports ↔ determined (f);
	environment ↔ appreciates nature (f)
		workplace ↔ sales representative (W);
		environment ↔ admires nature (W);
		art leisure ↔ makes friends across cultures (ME)
			law policy ↔ devout (C);
			art leisure ↔ devout (C);
			education ↔ devout (C)

More details, analyses, and discussions are in our paper!