Research Group Social Computing
TUM School of Computation, Information and Technology
Technical University of Munich

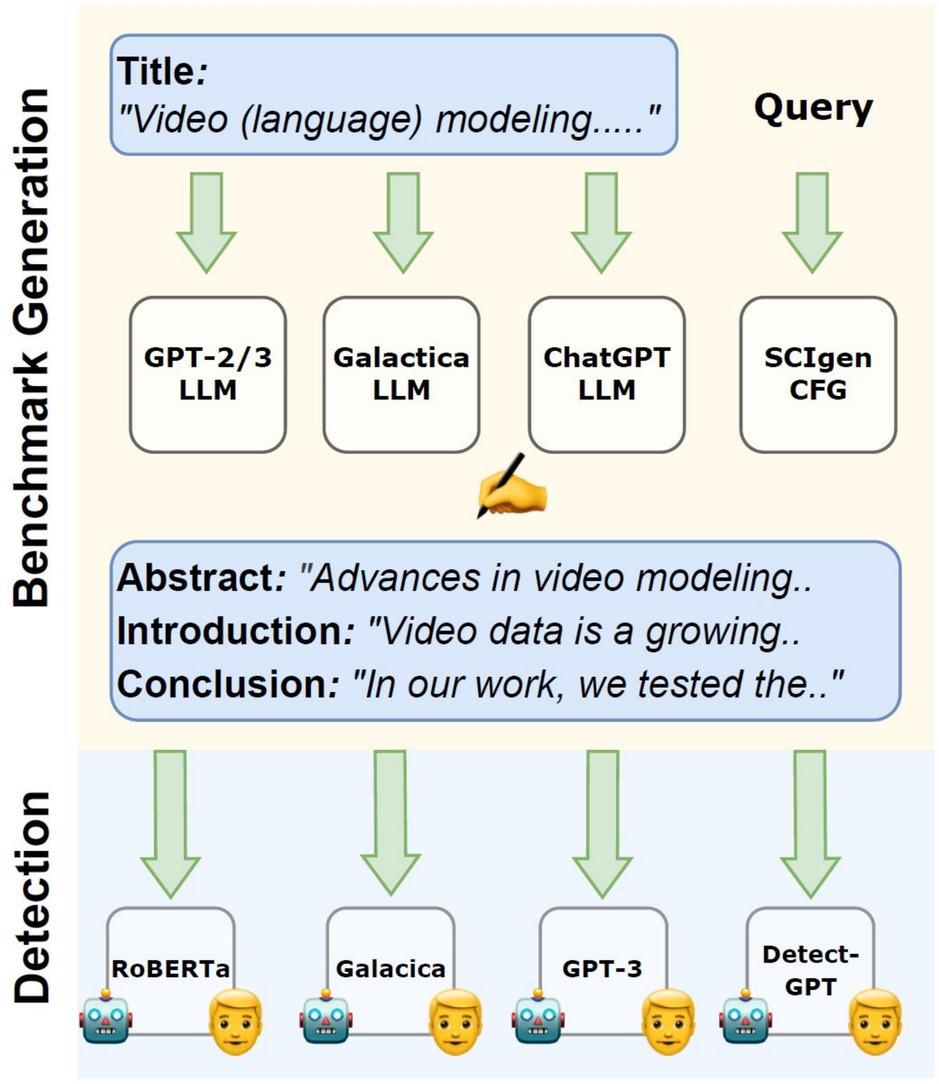
A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era.

Edoardo Mosca, Mohamed Hesham I. Abdalla, Paolo Basso, Margherita Musumeci, Georg Groh

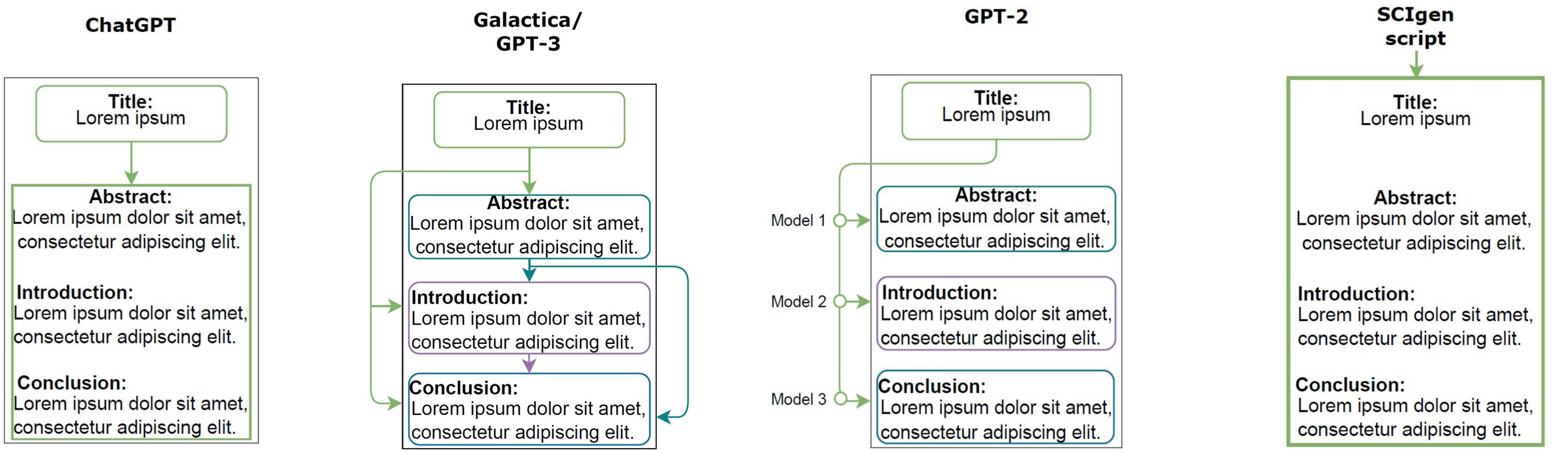
Contributions

- **Benchmark dataset** comprising **real (human-written)** and **fake (machine-generated) scientific documents**. Each contains an **abstract, introduction,** and **conclusion**.
- Evaluation of **four different classifiers** to **determine the authorship** (real or fake) of the documentst.
- Analysis and expla of **classifiers' generalization abilities** by evaluating their performance on both **in-domain** and **out-of-domain settings**.

Can we distinguish **facts** from **fiction**?



Generation Sequence



Overview of the Dataset

Source	Quantity	Tokens
arXiv parsing 1 (real)	12k	13.40M
arXiv parsing 2 (real)	4k	3.20M
SCIgen (fake)	3k	1.80M
GPT-2 (fake)	3k	2.90M
Galactica (fake)	3k	2.00M
ChatGPT (fake)	3k	1.20M
GPT-3 (fake)	1k	0.50M
Total real (extraction)	16k	16.60M
Total fake (generators)	13k	8.40M
Total	29k	25M

Experimental Results

Indicates out-of-domain experiments

Model	Train Dataset	TEST	OOD-GPT3	OOD-REAL	TECG
GPT-3 (our)	TRAIN-SUB	**99.96%**	25.9%	99.07%	**100%**
Galactica (our)	TRAIN	98.3%	24.6%	95.8%	83%
Galactica (our)	TRAIN+GPT3	98.5%	70%	92.1%	87.2%
Galactica (our)	TRAIN-CG	95%	11.1%	96.9%	42%
RoBERTa (our)	TRAIN	86%	23%	76%	**100%**
RoBERTa (our)	TRAIN+GPT3	68%	**100%**	36%	63%
RoBERTa (our)	TRAIN-CG	75%	32%	58%	88%
DetectGPT	-	61.5%	0%	**99.92%**	68.7%

Data
Models
Code

Takeaways

- Detection baselines can be really good, but sometimes **struggle with out-of-domain data**.
- No good **open-source detectors** available to test against.
- Future research should **include more paper sections.**
- **Human-Machine hybrids** are also a must for future research.