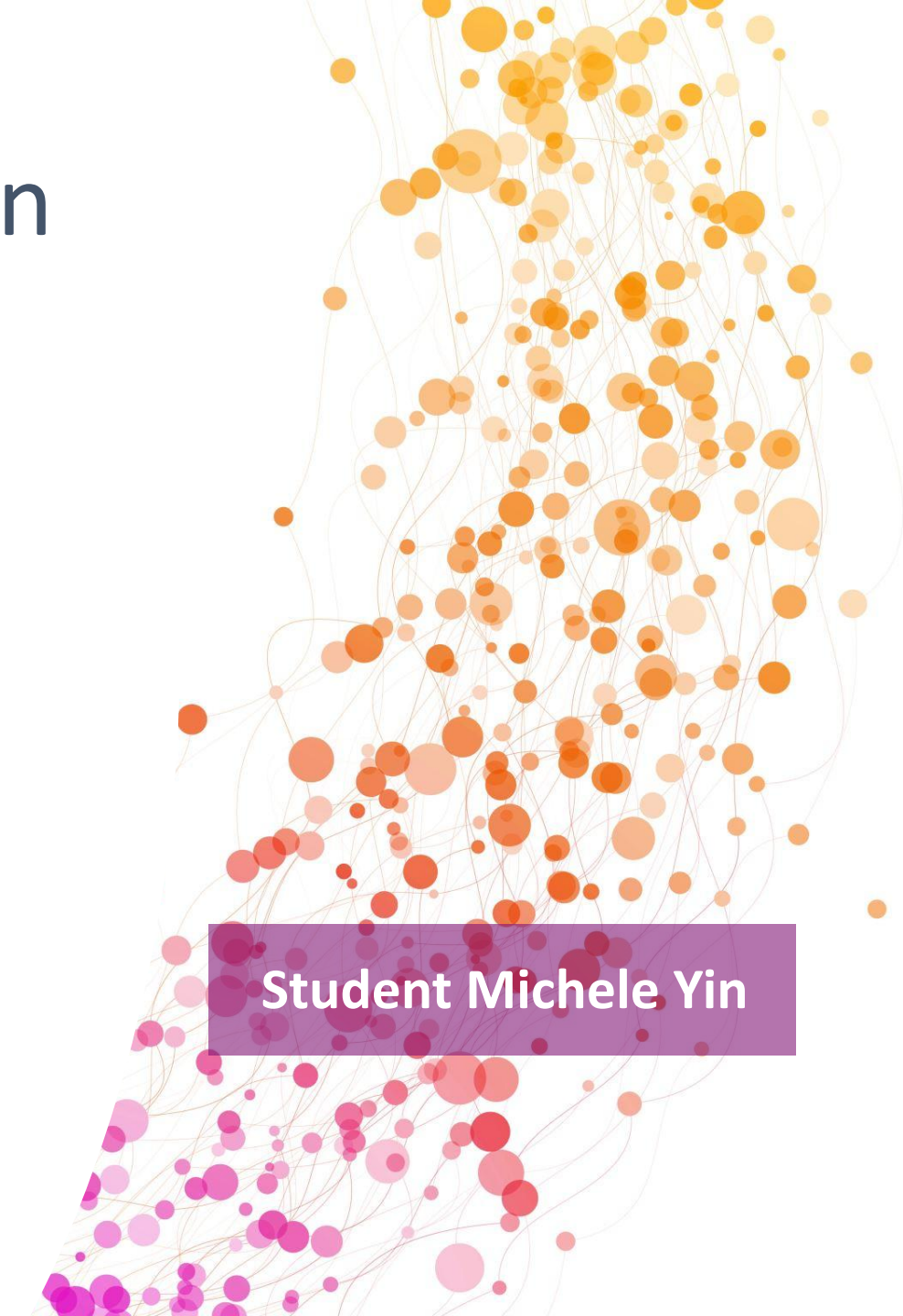
Automatic Narrative Elicitation with Large Language Models



Supervisor Prof. Giuseppe Riccardi

Co-Supervisor Gabriel Roccabruna

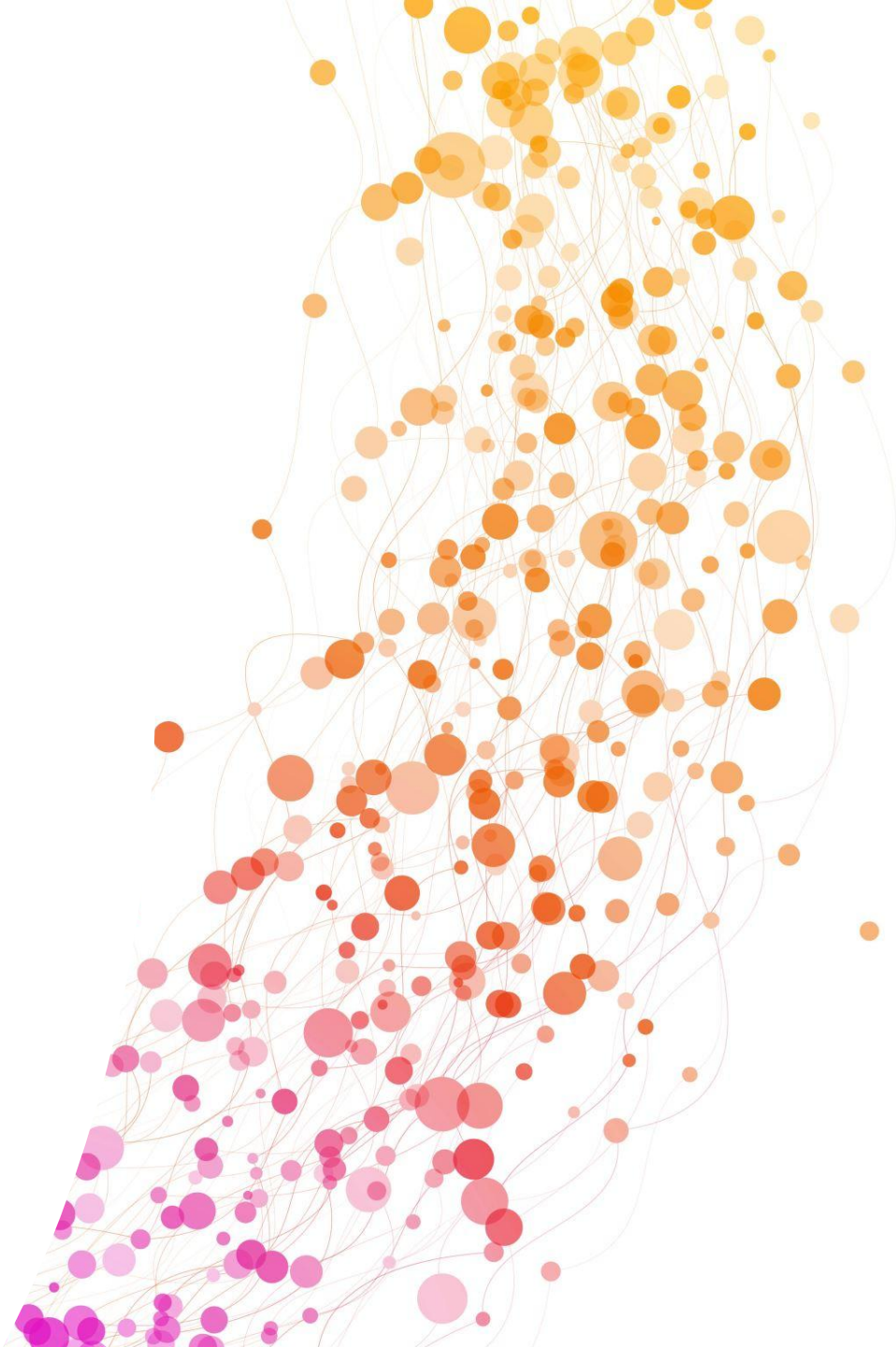
Student Michele Yin

UNIVERSITÀ DI TRENTO

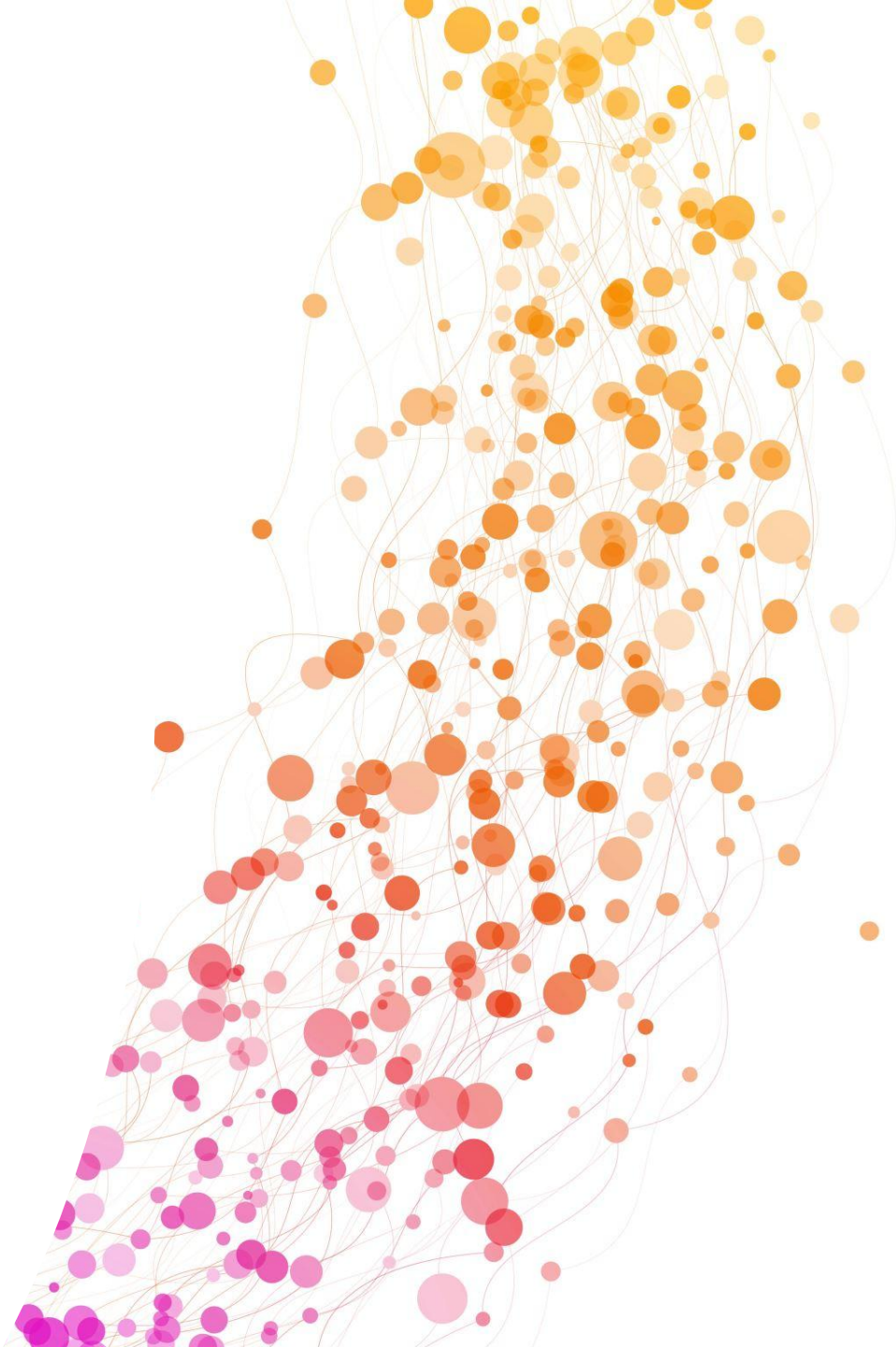
SiS LAB Signals & Interactive Systems

Outline

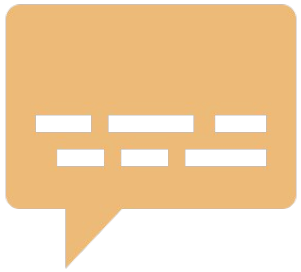
- Introduction
- Our contributions
- Methodology
- Results
- Conclusion



Introduction



Personal narrative



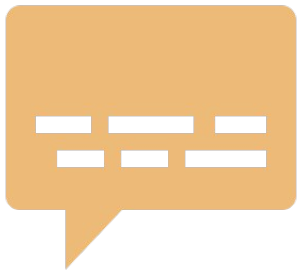
I feel happy. I’m going to visit my daughter.

- Qualitative research to gain insights on the human mind
- Narration act may stop

Narrative Interview

Convey the feeling that the interviewer is ***actively listening*** to the narrative

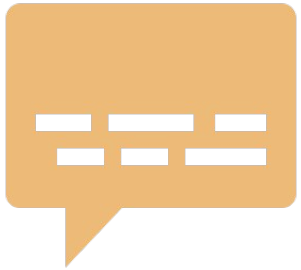
Personal narrative - active listening



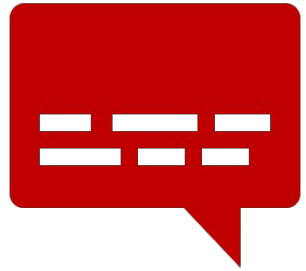
I feel happy. I’m going to visit my daughter.

I’m happy to hear that. Where does she live?

Personal narrative - not active listening

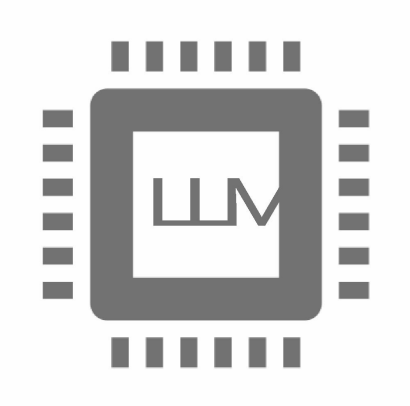


I feel happy. I'm going to visit my daughter.



I don't care.

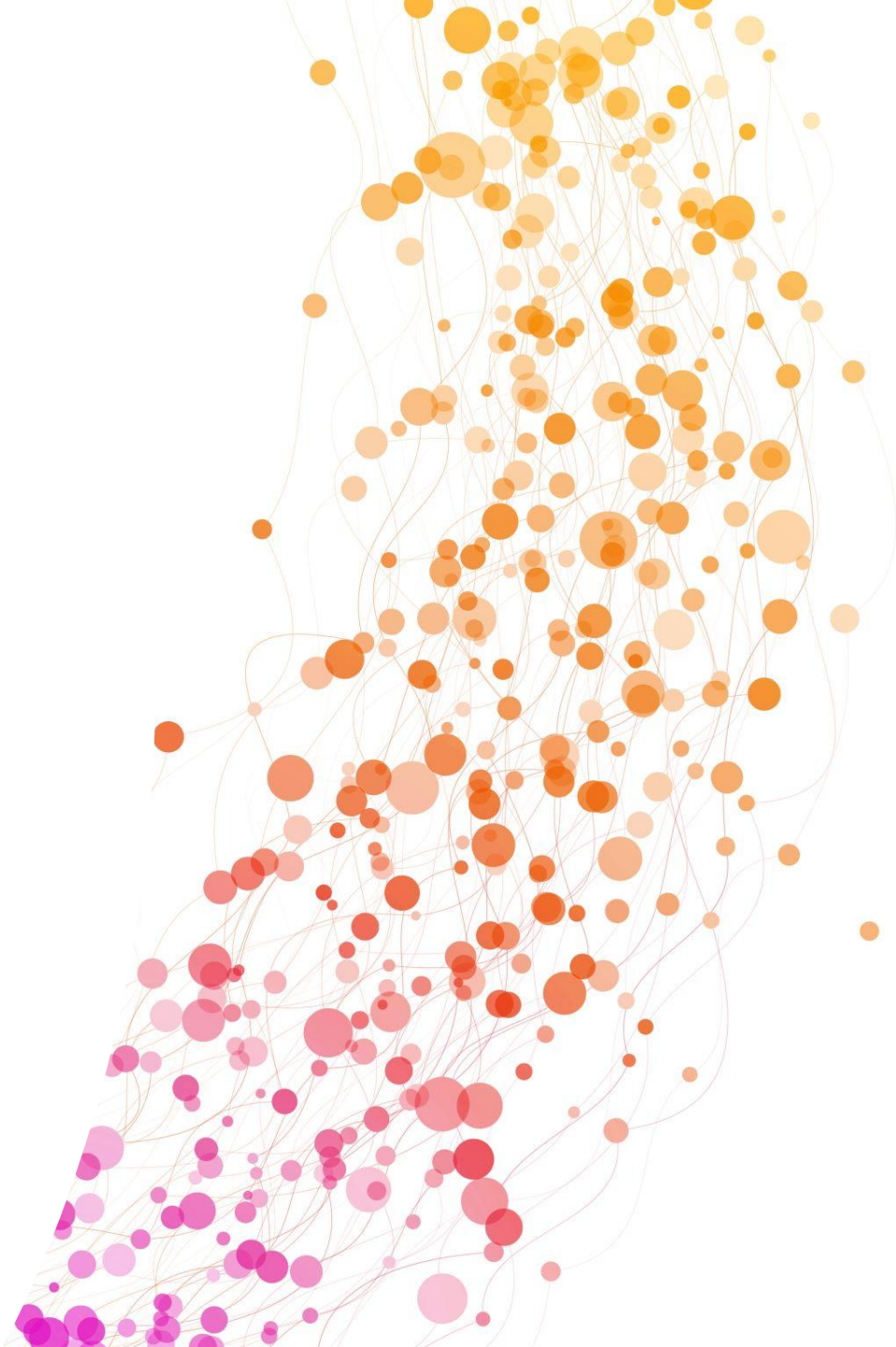
Large Language Models



Large Language Models (LLMs) have achieved remarkable success in a wide range of applications and domains

Their capacity to comprehend and generate human-like text has us posit whether LLMs can effectively used for this task

Our Contributions

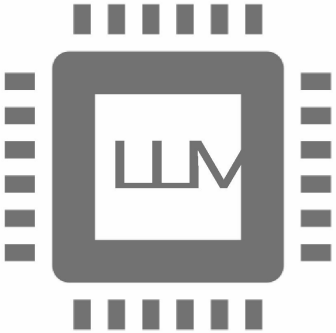


Our Contributions

- This thesis investigates **the performance of LLMs** in the novel task of ***Automatic Narrative Elicitation*** (ANE).

- The aim is to investigate whether **LLMs can** be prompted to **engage in** interactions that lead to the **correct elicitation of** continuation of **personal narratives**.

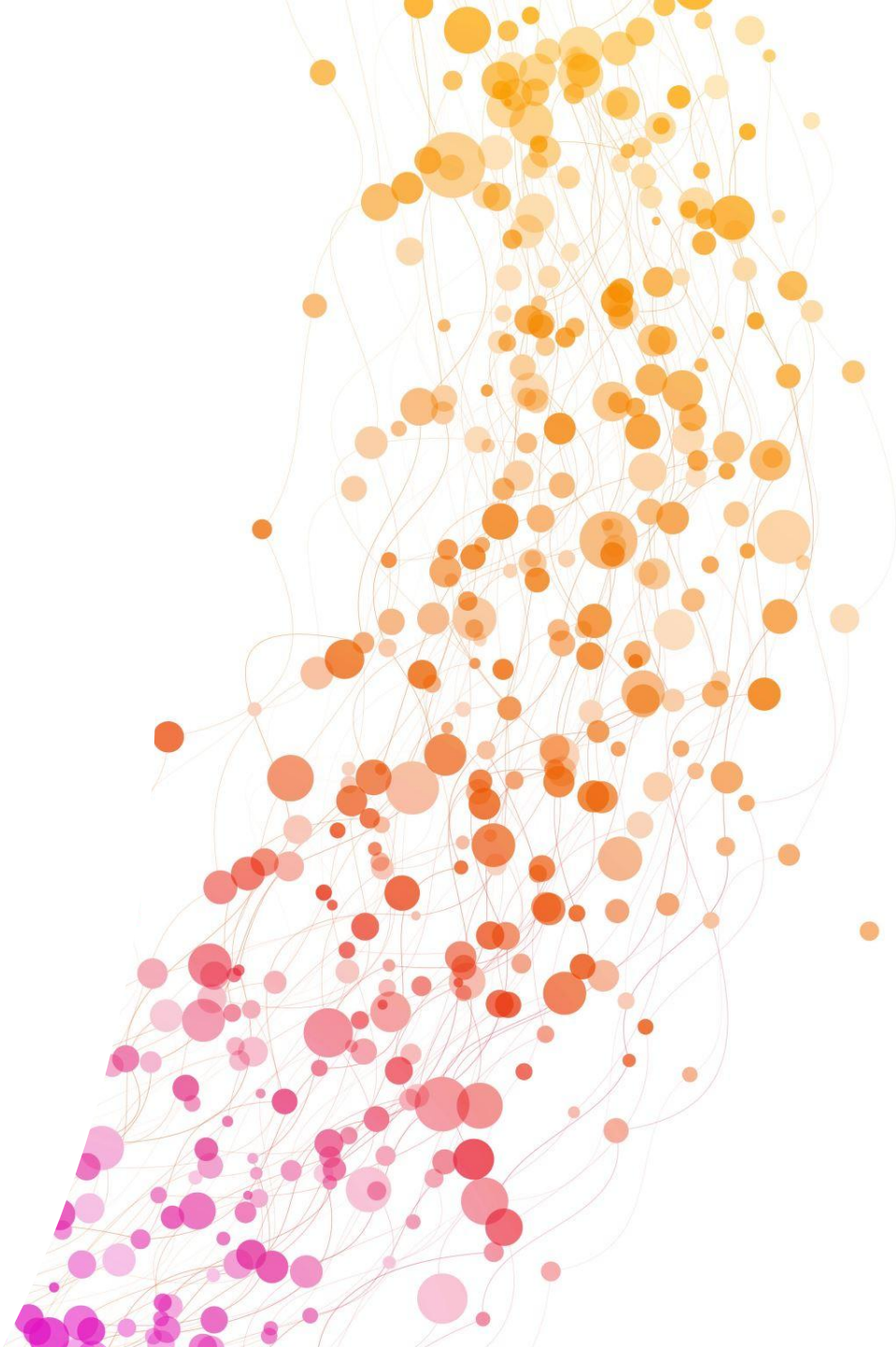
Automatic Narrative Elicitation



In the ANE task, the model acts as interviewer

The goal is to generate **active listening questions** to continue the narration

Methodology



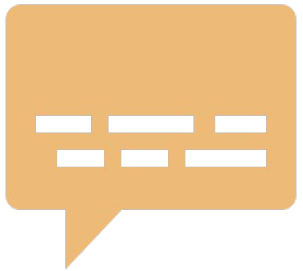
Corpus

481 personal narratives in the CoAdapt corpus, in the Italian language, annotated with Valence.

Stress and sorrowful narratives are a good challenge as they require **empathetic responses**

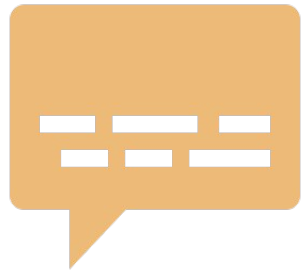
Valence

Negative Valence



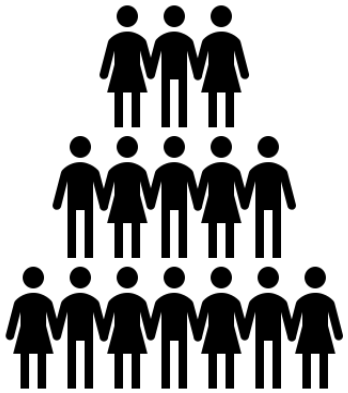
This evening I felt sad and alone.

Positive Valence



Hi, I’m all good today

Crowdsourcing



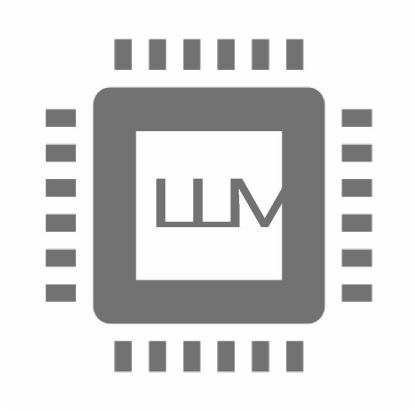
Crowdsourced eliciting questions

Corpus

Stress and sorrowful narratives are key as they require empathetic responses

Data collection

Large Language Models



Prompt design

LLMs - Prompt components

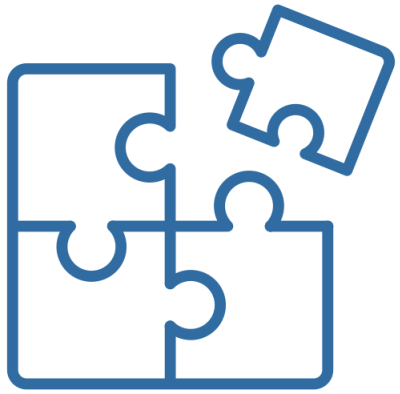
Number of shots (examples)
Personal narratives

Guidelines
Use of annotation task instructions

Valence

LLMs selection challenges

No Italian language



Coherent listening behaviour

LLMs selection solution

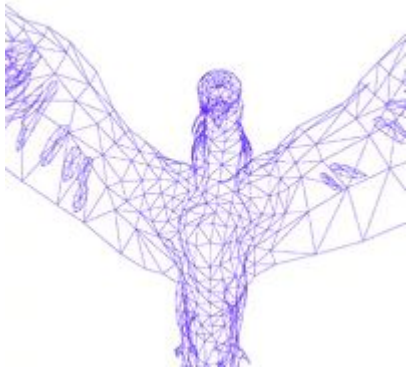
Italian language Test

- Ciao, come stai?
- Correggi questa frase:
  me: Ogi o litiagto coll cappo.

Selected LLMs

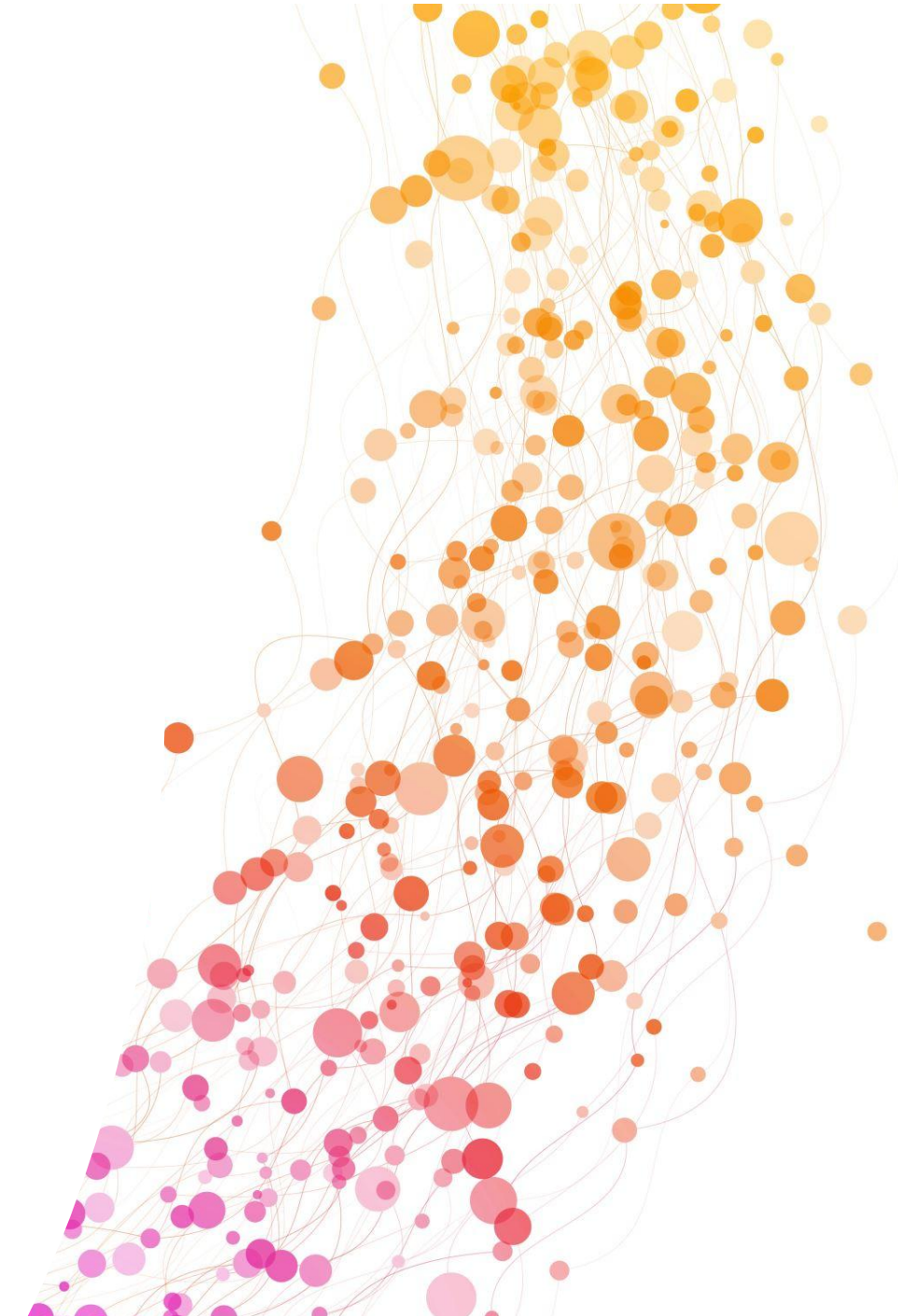
From the top 22 models from the HuggingFace 🤗 Leaderboard*, only 9 were selected

Among which there are Falcon, Vicuna and ChatGPT models

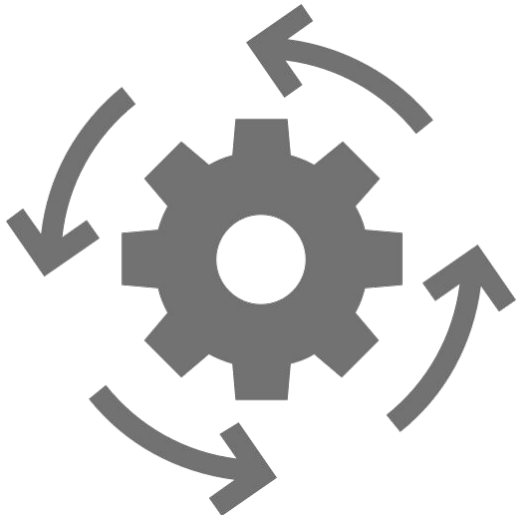


https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboardce.co/spaces/HuggingFaceH4/open_llm_leaderboard

Results



Results



Automatic metrics
(BLEU)

Human Evaluation

Results BLEU

	Without Valence								With Valence							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
ChatGPT	0.10	0.11	0.14	0.12	0.14	0.15	0.15	0.15	0.10	0.10	0.15	0.15	0.16	0.16	0.15	0.16
W. V. 13B	0.07	0.05	0.07	0.07	0.03	0.03	0.07	0.04	0.10	0.05	0.05	0.07	0.05	0.03	0.04	0.03
Vic 33B	0.06	0.04	0.07	0.08	0.05	0.06	0.08	0.05	0.07	0.06	0.05	0.06	0.03	0.03	0.02	0.01

Results BLEU

	0 - shot	1 - shot	3 - shot	5 - shot
ChatGPT	0.10	0.11	0.14	0.12
Wizard Vicuna 13B	0.07	0.05	0.07	0.07
Vicuna 33B	0.06	0.05	0.07	0.07

Number of examples increases performances for ChatGPT

Results BLEU

	Without Guidelines	With Guidelines
ChatGPT	0.12	0.15

The presence of guidelines increases performances for ChatGPT

Results BLEU

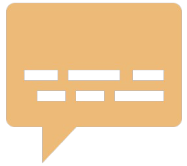
	Without Valence	With Valence
ChatGPT	0.15	0.16

The presence of valence increases performances for ChatGPT

Results

Automatic evaluation metrics do not really correlate with human evaluation judgments

Example



I feel happy. I’m going to visit my daughter.

That’s nice. Where are you going precisely?

I’m happy to hear that. Where does she live?

Example

Mi sento leggero. Sono in viaggio per andare a trovare mia figlia.

Sono molto felice per te, dove abita tua figlia?

Results

Human Evaluation

Human Evaluation Protocol*

Correctness

Appropriateness

Contextualisation

Listening

Mousavi, S. M., Roccabruna, G., Lorandi, M., Caldarella, S., & Riccardi, G. (2022, December). Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 136-147).

Results

Narrative Coming back from work, my daughter told me she felt a headache and called the doctor, she lost her sense of smell and taste and we are worried.

ChatGPT I’m sorry your daughter is experiencing this worrying condition. What did the doctor say about it?

Correctness	Appropriateness	Contextualisation	Listening
✓	✓	✓	✓

Results

Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa, non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.

Wizard Vicuna 13B Mi dispiace molto, come ti senti ora? Hai avuto altri sintomi? Hai già fatto una visita al medico?

Correctness	Appropriateness	Contextualisation	Listening
			

Results

Narrative Coming back from work, my **daughter** told me she felt a headache and called the doctor, she lost her sense of smell and taste and we are worried.

Wiz. Vic. 13B I'm very sorry, how do you feel now? Did you have other symptoms? Did you already have a visit with the doctor?

Correctness	Appropriateness	Contextualisation	Listening
✓	✓	✗	✗

Results

ChatGpt				

Results

	Correct.	Appropr.	Context.	Liste.
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

Results

	Correct.	Appropr.	Context.	Liste.
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

Human responses are not perfect

Results

Narrative Coming back from work, my daughter told me she felt a headache and called the doctor, she lost her sense of smell and taste and we are worried.

Human I’m sorry to hear that, how long did she have these symptom for ?

Correctness	Appropriateness	Contextualisation	Listening
✓	✓	✓	?

Results

	Correct.	Appropr.	Context.	Liste.
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

ChatGPT performs similarly to crowdworkers

Results

	Correct.	Appropr.	Context.	Liste.
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

The other models cannot achieve the same

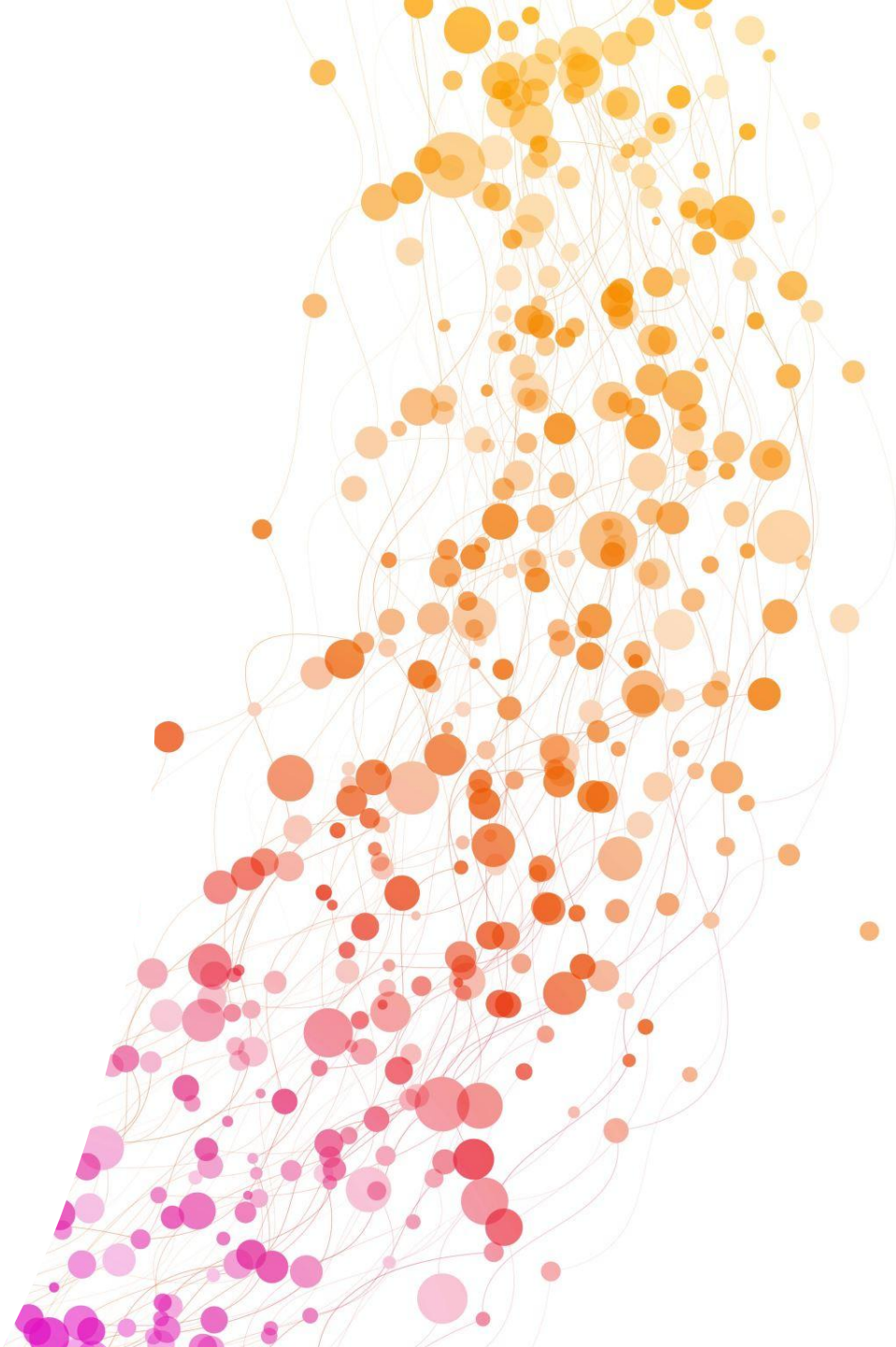
Results

	Correctness	Appropriateness	Contextualisation	Listening
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

Results

	Correctness	Appropriateness	Contextualisation	Listening
Human	90%	90%	86%	81%
ChatGPT	74%	86%	100%	88%
Wizard Vicuna 13B	48%	53%	66%	43%
Vicuna 33B	50%	22%	28%	16%

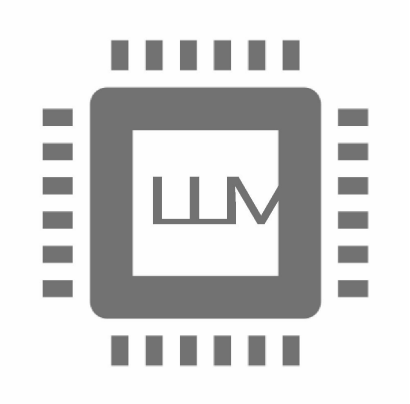
Summary



Summary

Defined a novel task of *Automatic Narrative Elicitation*

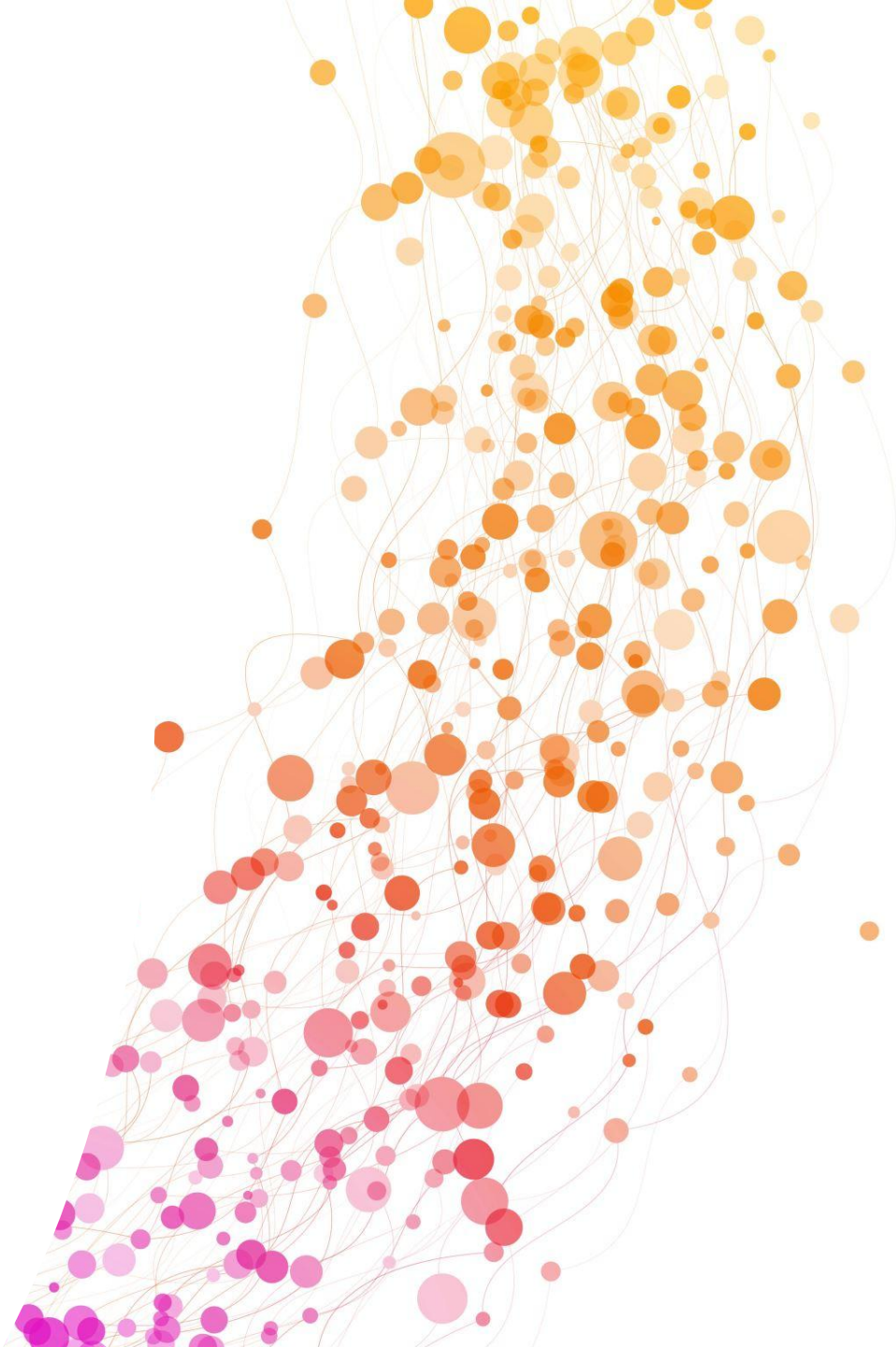
Crowdsourced a new corpus



Applied LLMs on the new task

Human Evaluation

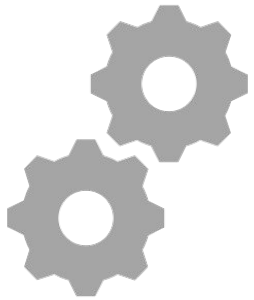
Conclusion



Conclusions

Our findings suggests that **LLMs**, in particular ChatGPT, **manage** the grounding context provided **to generate** adequate **correct and contextualised questions** based on the narrative topic.

Future Works



Fine-tuning LLMs focusing on the valence information.

Deeper **focus on their Italian language** abilities.

Thank you

Results

Narrative Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa, non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.

ChatGPT Mi dispiace sentire che tua figlia sta avendo questi sintomi preoccupanti. Cosa ha detto il medico riguardo ai suoi sintomi?

Correctness	Appropriateness	Contextualisation	Listening
			

Results

Narrative Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa, non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.

Wiz. Vic. 13B Mi dispiace molto, come ti senti ora? Hai avuto altri sintomi? Hai già fatto una visita al medico?

Correctness	Appropriateness	Contextualisation	Listening
			

Results BLEU

	Without Valence								With Valence							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
ChatGPT	0.10	0.11	0.14	0.12	0.14	0.15	0.15	0.15	0.10	0.10	0.15	0.15	0.16	0.16	0.15	0.16
W. V. 13B	0.07	0.05	0.07	0.07	0.03	0.03	0.07	0.04	0.10	0.05	0.05	0.07	0.05	0.03	0.04	0.03
Vic 33B	0.06	0.04	0.07	0.08	0.05	0.06	0.08	0.05	0.07	0.06	0.05	0.06	0.03	0.03	0.02	0.01