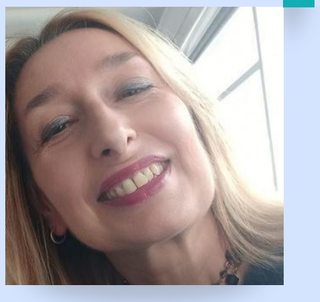
Enhancing HCM through GPT-driven questionnaire generation

Lucrezia Laraspata

Fabio Cardilli



Giovanna Castellano

Gennaro Vessio

Introduction

Context

HR Questionnaires

Decision-support tool used by HR and Managers to **investigate phenomena** within the company by **collecting feedback and opinions** from employees.

Idea and Objective

Apply **LLMs** to generate *tailored* and *engaging* surveys to exploit their **text generation** capabilities.

Reduce the needed **time** to prepare the content.

Introduction

Problems

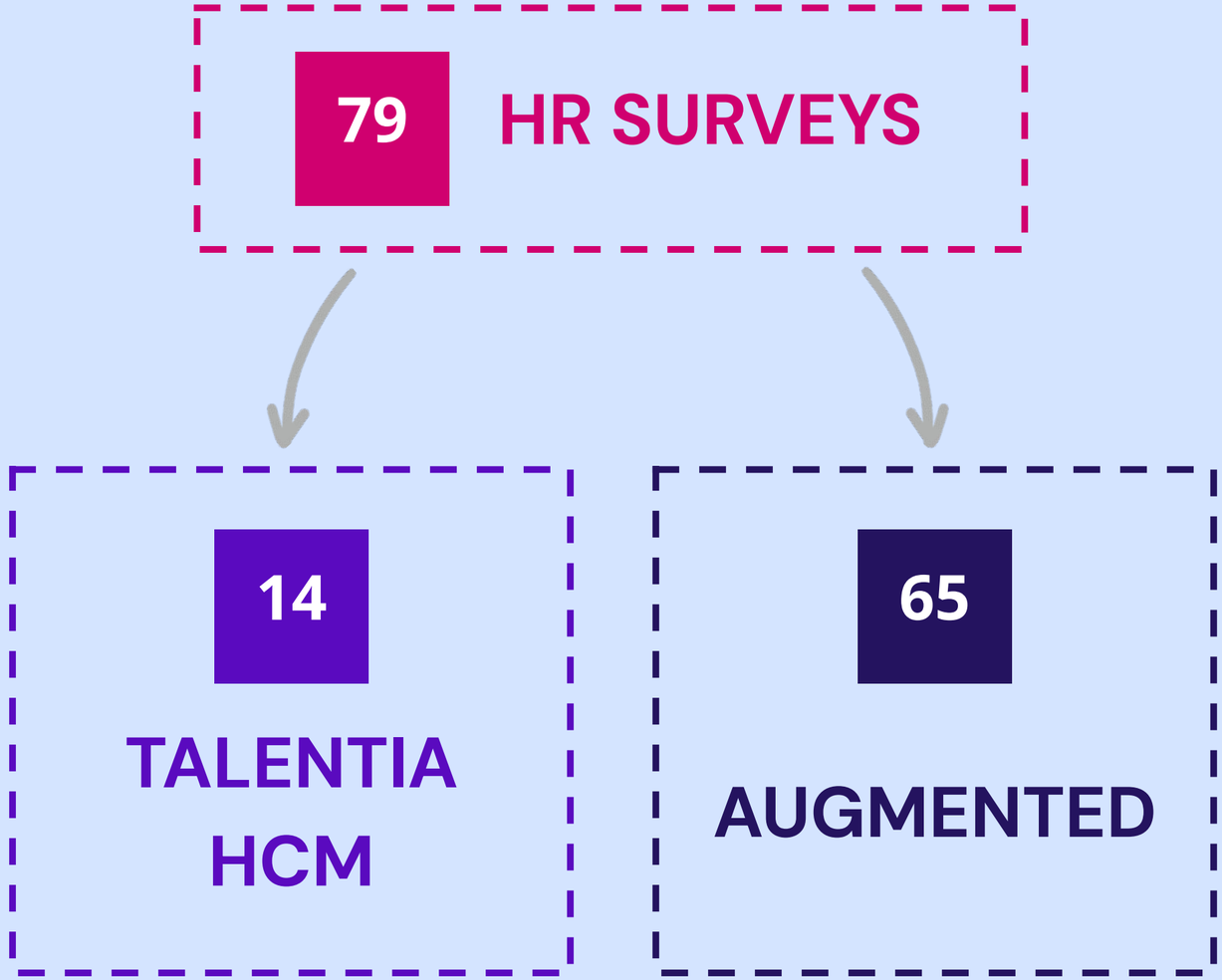
01 No dataset available

02 No proper distinction among types

03 Poor content quality evaluation

Materials

Dataset



Materials

Dataset

01 **Topic identification**
by Talentia R&D Team

02 **Survey Generation**
by ChatGPT

03 **Human Validation**
by Talentia R&D Team

Experiments

Workflow

Questionnaire topic
Number of questions

Generated JSON
Token consumption
Response time

01

Prompt generation

Techniques:

- zero-shot
- one-shot

Roles:

- system
- assistant
- user

02

Ask LLM

GPT Models:

- 3.5-Turbo
- 4-Turbo

Hyper-Params:

- Temperature
- Frequency Penalty

Experiments

System prompt

01

“You are a **Questionnaire Generator** in the **HRM field.**”

02 A

“The user will ask you to generate a **questionnaire** specifying the **topic** and the **number of questions.**”

02 B

“The user will ask you to generate a **questionnaire** about a specified **topic.**”

Experiments

System prompt

03

“If the user does **not** specify a **valid topic**, reply with "Sorry I can’t help you".”

04

“If the topic is valid, **reply with only a JSON**, which must respect the following format:”

05

<**JSON format** description>

Experiments

System prompt

06

“The admitted **question's types** are the following:
- ID: <id>, <description>
[...]”

07

“Be **creative** and **vary** the **syntax** of your questions to enhance **user engagement**.”

08

“Reply **only** with the **JSON**.”

Experiments

User prompt

01 **A**

“Generate me a questionnaire on **Career development** with **10 questions**”

01 **B**

“Generate me a questionnaire on **Career development**”

Experiments

Assistant prompt

01

```
“{
  "data": {
    "QUESTIONNAIRE": [
      {
        "NAME": "Access to Technology and Tools",
        "QUESTIONS": [
          {
            [ ... ]
}”
```

Results

Proposed metrics

01

Generated content

- **Syntactic** and **lexical** similarity
- **Serendipity**
- **Instruction following** capability

02

Ground-truth comparison

- **Semantic** similarity
- **Indistinguishability** assessment

Q1 - Did the kick-off meeting meet your expectations?

- Yes, it exceeded my expectations.
- Yes, it met my expectations.
- No, it did not meet my expectations.
- No, it fell short of my expectations.

Q2 - What was the most valuable aspect of the kick-off meeting?

Q3 - What was the least valuable aspect of the kick-off meeting?

Q4 - Were the goals and objectives of the project clearly communicated during the kick-off meeting?

- Yes
- No

Q5 - Did you feel engaged and involved during the kick-off meeting?

- Yes, I felt very engaged and involved.
- Yes, I felt somewhat engaged and involved.
- No, I did not feel engaged or involved.

Q6 - What could have been done differently to improve the kick-off meeting?

Q7 - Overall, how would you rate the kick-off meeting?

Results

Generated content

Metric	Model	Technique	Task	Score	Variance
IQS	**GPT-4-Turbo**	Zero-shot	A	**0.18**	**0.0006**
	GPT-3.5-Turbo	Zero-Shot	A	**0.34**	**0.0029**
SDP	**GPT-4-Turbo**	Zero-shot One-Shot	B A	**0.84**	**0.0005**
	GPT-3.5-Turbo	Zero-Shot	A	**0.75**	**0.0032**

Best and worst experiments

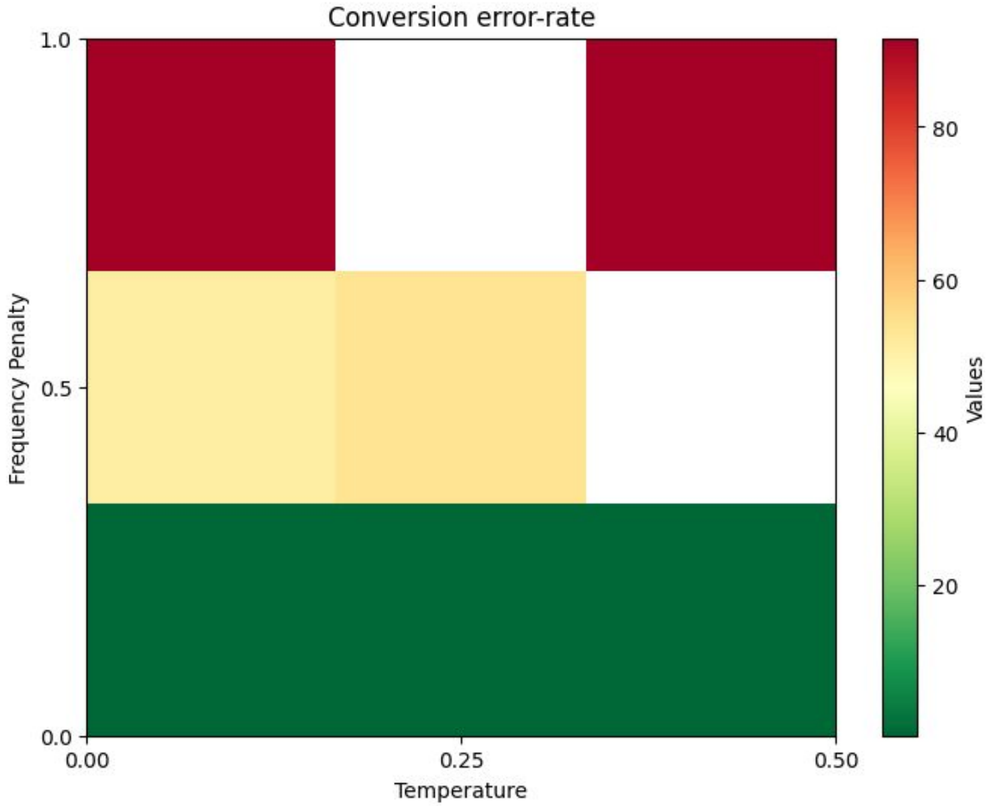
according to intra-questionnaire similarity (ISQ) and serendipity (SDP)

Results - Generated content

Instruction following

Check that the specified JSON structured was followed by the model.

Varying the temperature level does not affect the JSON structure generation.



Increasing the Frequency Penalty values may lead to critical integration issues.

Results - Ground-truth comparison

Semantic similarity

Weigh the **similarity** to the **ground-truth questions** and the **topic**, **penalizing** the final score according to the **deviation** from the **ideal position**.

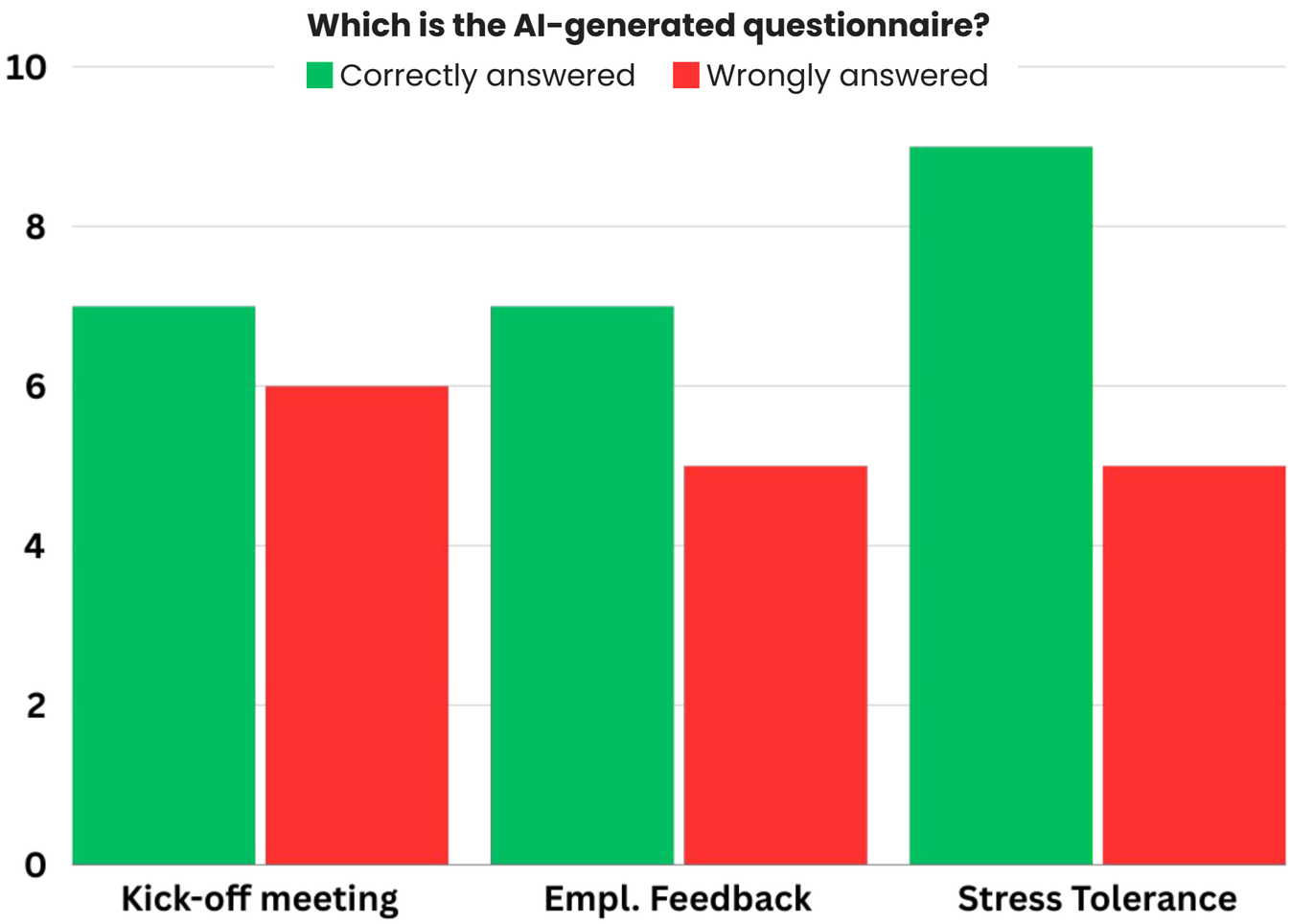
Model	Technique	Task	Score	Delta
GPT-3.5-Turbo	Zero-shot	B	**0.48**	**-18.14%**
GPT-4-Turbo	Zero-Shot	B	**0.44**	**-22.49%**

Best and worst experiments

Scores based on a metric designed for this study which uses Cosine similarity among OpenAI embeddings

Results - Ground-truth comparison

Indistinguishability assessment



Language style and **variability of questions** and **answers** resulted to be the most considered characteristics.

Future works

01 Enhance question ordering

02 Improve instruction following

03 Mitigate hallucination

Thank you!

Any question?