Enhancing Human Capital Management through GPT-driven Questionnaire Generation

Lucrezia Laraspata[1,2], Fabio Cardilli[2], Giovanna Castellano[1] and Gennaro Vessio[1,*]

[1]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
[2]Talentia Software, Bari, Italy

Abstract

Survey questionnaires capture employee insights and guide strategic decision-making in Human Capital Management. This study explores the application of the GPT-3.5-Turbo and GPT-4-Turbo models for the automated generation of HR-related questionnaires, addressing a significant gap in the literature. We developed a novel dataset of HR survey questions and evaluated the models' performance using different task configurations, including zero-shot and one-shot prompting with various hyperparameter settings. The generated questionnaires were assessed for instruction alignment, syntactic and lexical diversity, semantic similarity to human-authored questions, and topic diversity, or serendipity. In collaboration with Talentia Software, we additionally examined the indistinguishability of AI-generated content from human-created counterparts. Results indicate that both models produce questionnaires with high serendipity and intra-questionnaire diversity. However, the indistinguishability test revealed that human evaluators could still distinguish AI-generated content, particularly noting differences in language style and answer variability. These findings underscore the potential of GPT-driven tools in automating questionnaire generation while highlighting the need for further refinement to achieve more human-like outputs. The source code, data, and samples of generated content are publicly available at: https://github.com/llaraspata/HRMQuestionnaireGenerationUsingLLM.

Keywords

Questionnaire generation, Human Capital Management, Generative AI, LLMs, Prompt engineering

1. Introduction

Artificial Intelligence (AI) has rapidly become a key driver of success in business organizations, mainly through the automation of critical processes and the reduced time required for task completion. Among AI advancements, Large Language Models (LLMs) have gained significant attention for their ability to generate text with remarkable fluency and coherence, making them valuable tools for content creation [2, 3, 4, 5]. One promising application of LLMs is the generation of survey questionnaires, essential decision-support tools for HR professionals and managers in modern organizations.

Survey questionnaires are instrumental in gathering continuous feedback and opinions from employees, enabling organizations to monitor and enhance various aspects such as employee satisfaction, value alignment, performance, engagement, and potential assessment [6, 7, 8]. Despite their importance, designing effective surveys that accurately capture employee insights is often time-consuming, requiring careful consideration of question structure, flow, and relevance.

Currently, *questionnaire generation* remains underexplored within the scientific community. Researchers often approach this task from a learning perspective, frequently overlooking different questionnaires' distinct types and characteristics. For example, unlike training questionnaires or skill assessments, which may include scored questions to evaluate soft skills, surveys typically lack right or wrong answers. This lack of differentiation has contributed to a shortage of appropriate datasets tailored specifically for survey generation. Furthermore, while LLMs have been employed to tackle this challenge [9, 10, 11, 12], the evaluation of generated questionnaires has primarily relied on metrics

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]
*Corresponding author.
llaraspata@talentia-software.com (L. Laraspata); fcardilli@talentia-software.com (F. Cardilli); giovanna.castellano@uniba.it (G. Castellano); gennaro.vessio@uniba.it (G. Vessio)
0009-0003-8136-9140 (L. Laraspata); 0009-0006-8292-0442 (F. Cardilli); 0000-0002-6489-8628 (G. Castellano); 0000-0002-0883-2691 (G. Vessio)

borrowed from related fields like text summarization and translation, such as BLEU and ROUGE for syntactic similarity and cosine similarity for semantic comparison. However, these metrics fail to capture critical aspects unique to questionnaires, such as engagement and the logical flow of questions.

This work contributes to the field of Human Capital Management (HCM) by providing a new dataset of HR surveys and a novel evaluation framework, both of which are currently absent in the literature. Specifically, this study investigates the effectiveness of using two models from the GPT family—GPT-3.5-Turbo and GPT-4-Turbo—to automatically generate tailored HR questionnaires that efficiently collect insightful feedback within organizations. By leveraging LLMs, the time required to create such surveys can be significantly reduced, allowing HR professionals to focus on more complex and strategic tasks for the companies.

Our research aims to analyze these LLMs' capabilities in generating high-quality surveys when provided with limited input, such as the topic and number of questions, varying prompting techniques, and hyperparameter values. Moreover, we propose a methodology to evaluate the generated content's quality that encompasses key characteristics of HR questionnaires, including engagement, variability, and diversity of topics, as well as the model's alignment with the given instructions. Recognizing the limitations of automated evaluations, we also conducted a human assessment in collaboration with Talentia Software, a company specializing in digital transformation solutions for HR and finance. This evaluation included an indistinguishability assessment, where participants were asked to identify AI-generated questionnaires and explain their reasoning.

The rest of this paper is structured as follows. Section 2 highlights key contributions in related fields. Section 3 outlines the research design. Section 4 presents our framework and the obtained results. Section 5 highlights key findings and remaining challenges.

2. Related work

While HR survey generation remains a relatively underexplored application, recent studies on questionnaire generation using LLMs have provided valuable insights and methodologies that broaden the scope of this research area.

Lei et al. [9] introduced a comprehensive approach to evaluating LLM-generated questionnaires automatically. Their methodology assessed the syntactic similarity using the ROUGE-L score [13] and the semantic similarity by employing BERT [14] for sentence embeddings. Additionally, they syntactically measured the repetition of generated questions through n-gram overlaps and semantically by calculating the cosine similarity between questions. Questions were flagged as duplicates if their similarity score exceeded a threshold of 0.95. Lei et al. also evaluated the alignment of generated questionnaires with the intended task by using BLEU-n to compute n-gram overlaps between the questions and the questionnaire's description, with higher scores indicating better alignment. Furthermore, they conducted human evaluations to explore more nuanced aspects of the questionnaires, such as ambiguity, logical flow, and coherence.

Similarly, Doughty et al. [10] developed a survey questionnaire to gather opinions on skill assessment. In their study, human evaluators were tasked with rating the completeness and correctness of the answer sets for each question, ensuring a clear and correct answer was available. In another related work, Rodriguez-Torrealba et al. [11] designed a questionnaire to evaluate the difficulty and quality of the generated questions, focusing on their clarity and well-formedness.

The findings from Lei et al. highlighted a significant disparity between human and automatic evaluations of questionnaires, regardless of the domain. Human-written questionnaires consistently received higher scores in human evaluations, while LLM-generated questionnaires often struggled to achieve similar quality levels. However, when evaluated using automatic metrics, LLM-generated questionnaires appeared comparable to those created by humans. Despite the focus on performance assessment, both Doughty et al. and Rodriguez-Torrealba et al. identified limitations in using LLMs for questionnaire generation. The complexity of questions was often reduced, with instances of literal repetition from source materials within the correct answers. Additionally, generated questionnaires frequently con-

tained more than one correct answer or included incorrect options. These insights underscore the need for further research to address these limitations.

3. Materials and methods

This research aimed to integrate a GPT-driven questionnaire generation feature into the HCM system developed by Talentia Software, which already incorporates several automation mechanisms for dynamic data collection across different entities. To achieve this integration, the system's "interoperability skill" was utilized. This mechanism accepts a JSON string as an input parameter with a predefined structure, mapping each entry to specific fields in the HCM database. Consequently, it became necessary to instruct the model to generate output in JSON format, creating a seamless and transparent pipeline for the end-user.

Given the limitations of existing datasets used in previous studies [9, 10, 11], which predominantly focus on learning assessment questionnaires, we recognized the need to develop a new dataset specifically tailored to HR survey questionnaires. Different types of questionnaires come with distinct needs, constraints, and characteristics, necessitating a dataset that reflects these nuances in the HR domain. Therefore, a new data collection strategy was implemented.

3.1. Dataset

The dataset was created by choosing 14 HR questionnaires from Talentia HCM data. These questionnaires formed the basis for creating the entire set, including its entities and attributes. To expand the dataset and ensure thorough analysis, a data augmentation process was used, as described below:

1. *Topic identification*: The Talentia HCM R&D department identified 40 topics relevant to HR survey questionnaires, focusing on areas such as employee satisfaction, work experiences, and growth opportunities.
2. *Survey generation*: The content for the questionnaires was generated using the ChatGPT web application.[1] For each identified topic, one or more questionnaires were generated with the following prompt:

   > *I'm working on surveys for gathering feedbacks from Human Resources in a company. Can you please generate me a survey about* '`topic`'*?*

   No further constraint concerning the surveys' structure was imposed during this generation process to allow for greater flexibility and creativity.
3. *Human correction and validation*: To ensure high-quality data, the Talentia HCM R&D team reviewed, corrected, and validated 65 generated questionnaires. This step was crucial for addressing potential issues, such as hallucinations, and maintaining the unstructured text format.
4. *Conversion to JSON*: To streamline the process, an automatic data import mechanism in Talentia HCM was used to ingest the questionnaires as JSON objects, mapping them to the appropriate database tables. For this purpose, GPT-3.5-Turbo was utilized to convert the unstructured text into JSON format. The one-shot prompting technique was employed with a fixed example. Such a survey was selected as it comprehends different question types so that the model could better learn how to convert them. We used a temperature and frequency penalty set to 0 and a max token limit of 6,000. The system prompt was tailored to include only four question types, as the model showed difficulty managing a more extensive variety, leading to incorrect assignments in preliminary trials.
5. *Final human validation*: After the JSON conversion, a final human validation step was conducted to correct any remaining error, such as misaligned question types or missing answers, ensuring the accuracy and reliability of the dataset.

[1] https://chatgpt.com/ (accessed on September 2024)

Table 1
Statistics of the proposed HR questionnaire dataset.

	Total	Talentia HCM	Augmented
Questionnaires	79	14	65
Questions	603	113	490
Question types	8	8	8
Answers	2170	424	1746
Questionnaire subtopics	434	434	434
Average questions per questionnaire	8	8	8
Average answers per question	5	5	5
Average subtopics per main topic	11	11	11
Average question length (words)	12	9	13
Average answer length (words)	1	1	1

The resulting collection of generated questionnaires, now in JSON format, was stored in the local Talentia HCM database and later extracted in CSV format for analysis. Key statistics about the dataset are presented in Table 1.

3.2. Task definition

This study explores the capabilities of GPT models in generating HR questionnaires, focusing on two task variants:

1. The user requests the model to generate a questionnaire by specifying the questionnaire topic and the number of questions.
2. The user requests the model to generate a questionnaire by specifying only the questionnaire topic.

Differently from the data augmentation step, here we defined an additional restriction represented by the number of questions in task (1). These task definitions impose minimal constraints on the content to be generated, providing the model with a significant degree of freedom to demonstrate its creativity. However, the challenge lies in the limited information provided, which requires the model to rely heavily on its internal knowledge, increasing the risk of generating irrelevant or inaccurate content.

3.3. GPT models

This study focused on two advanced models from the GPT family: GPT-3.5-Turbo [15] and GPT-4-Turbo [16]. These models are well-known for their versatility and high-quality text generation capabilities, making them popular choices across various disciplines. The experiments utilized Azure OpenAI APIs to deploy these models. The specific configurations for GPT-3.5-Turbo and GPT-4-Turbo on Azure are summarized in Table 2.

The decision to use Azure AI services aligns with the strategic deployment of Talentia HCM on Azure, especially for new customers. GPT-3.5-Turbo was chosen for its cost-effectiveness and speed, while GPT-4-Turbo was primarily used to explore the JSON mode feature and evaluate the performance improvements associated with a larger model in terms of questionnaire quality and adherence to instructions.

Table 2
Configurations of GPT-3.5-Turbo and GPT-4-Turbo deployed on Azure.

Configuration	GPT-3.5-Turbo	GPT-4-Turbo
Version	0301	1106-Preview
Tokens per minute rate limit (thousands)	120	30
Rate limit (tokens per minute)	120,000	30,000
Rate limit (requests per minute)	720	180

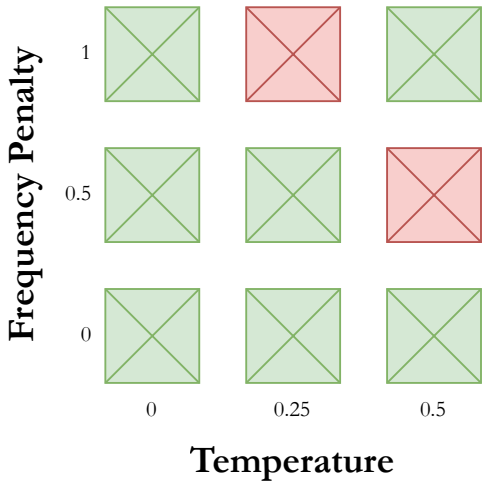

Figure 1: Graphical representation of the tested combinations of temperature (horizontal axis) and frequency penalty (vertical axis) values. Green entries indicate tested combinations, while red entries indicate untested combinations.

4. Experimental evaluation

4.1. Setting

4.1.1. Hyperparameter configuration

The experimental setup involved testing various hyperparameter configurations for the GPT models:

- *Temperature*: This parameter, ranging from 0 to 2, controls the randomness of the model's outputs. Lower values lead to more deterministic results, which means selecting the highest probability tokens. Altering this parameter can increase the variability in generated questions, helping avoid repetitive structures. The tested values were $T \in \{0, 0.25, 0.5\}$.
- *Frequency penalty*: With values between −2 and +2, this parameter adjusts the likelihood of token repetition. Higher values penalize repeated tokens, encouraging the model to generate more diverse responses. The tested values were $FP \in \{0, 0.5, 1\}$.

Figure 1 illustrates the combinations of temperature and frequency penalty values tested during the experiments.

Additionally, the following parameters were consistently configured across all experimental setups:

- *Max tokens*: This parameter sets the maximum number of tokens the model can generate. For GPT-3.5-Turbo, the limit was set at 6,000 tokens, and for GPT-4-Turbo, it was set at 4,000 tokens.
- *Response format*: For GPT-4-Turbo, the response format was configured to output a valid JSON object by setting the API call parameter to `{ "type": "json_object" }`. This ensures that the generated output adheres to a JSON structure, facilitating seamless integration with Talentia HCM database. This feature was available only for GPT-4-Turbo, as GPT-3.5-Turbo does not support JSON mode in the deployed version.

4.1.2. Prompt engineering

We employed both zero-shot and one-shot techniques to prompt the GPT models. Large-scale training enables LLMs to perform a wide range of tasks in a zero-shot manner, meaning they can generate responses without prior examples or demonstrations. This approach, introduced by Radford et al. [17], eliminates the need for additional training data and instead focuses on crafting specific prompts that guide the model's behavior for the given task. In zero-shot prompting, the model has a task description without labeled data or input-output mappings, relying on its pre-existing knowledge to generate responses. While LLMs demonstrate strong zero-shot capabilities, they may struggle with more complex tasks. In such cases, few-shot prompting [18] can be employed to enhance performance by providing a fixed number of high-quality examples. However, this approach increases token consumption, which can be a limitation for longer text inputs, and the selection of examples can significantly influence the model's output.

To further improve prompt effectiveness, we defined three key roles in the prompt structure to guide the conversation flow with the LLMs:

- *System*: The system prompt provides high-level instructions and describes the application context, guiding the model's behavior during the task. The designed system prompt, as detailed in Table 3, includes:
  - *Role definition*: Specifies the role the model should assume and constrains its behavior accordingly.
  - *User input format definition*: Defines how the user will interact with the model, specifying the information required to perform the task. This varies depending on the task variant being tested.
  - *Error message*: Instructs the model on responding when the user provides invalid input.
  - *Task definition*: Clarifies the task the model needs to accomplish.
  - *Model output format definition*: Details the output structure, such as the properties of the JSON response.
  - *Admitted question type definition*: Specifies the types of questions the model can generate, tailored to the requirements of this study.
  - *Style command*: Instructs the model to follow a specific syntactic and lexical style when generating text.
  - *Output format reinforcement*: Enhances the model's adherence to the specified output format, especially when multiple instructions are provided.
- *Assistant*: The assistant prompt is used only in few-shot scenarios to simulate the model's response. For the task defined in this study, the assistant prompt contains only the JSON of the questionnaire. It is designed similarly to the assistant prompt for JSON conversion in the data augmentation process, as detailed in Table 4.
- *User*: The user prompt represents the command the user gives to initiate the task the model is expected to perform. This prompt is critical as it directly influences the model's output. The specific user prompt varies depending on the task variant being tested, as detailed in Table 5.

4.2. Performance metrics

As introduced before, we propose a new evaluation framework that could automatically estimate the quality of the generated surveys. The framework is highly general and flexible, allowing for easy adaptation to domains beyond HCM with only minor modifications.

4.2.1. Intra-questionnaire similarity

To enhance the engagement of the generated questionnaires, the system prompt included the style command: "*Be creative and vary the syntax of your questions to enhance user engagement.*" An engaging

Table 3
System prompt design for questionnaire generation.

Prompt part	Content
Role definition	You are a Questionnaire Generator in the Human Resource Management field.
User input format definition (variant 1)	The user will ask you to generate a questionnaire specifying the topic and the number of questions.
User input format definition (variant 2)	The user will ask you to generate a questionnaire about a specified topic.
Error message	If the user does not specify a valid topic, reply with "Sorry, I can't help you."
Task definition	If the topic is valid, reply with only a JSON, which must respect the following format:
Model output format definition	`<JSON structure>`
Admitted question type definition	The admitted question types are: - ID: `<id>`, DESCRIPTION: `<description>` ...
Style command	Be creative and vary the syntax of your questions to enhance user engagement.
Output format reinforcement	Reply only with the JSON.

questionnaire typically features high lexical variability, which prevents it from becoming monotonous or tedious.

The effectiveness of this approach was measured by evaluating the intra-questionnaire lexical similarity of the generated questions. Lexical metrics provide valuable insights into this characteristic, where higher scores indicate that the questions share nearly identical syntactic and lexical structures, ultimately leading to a lower overall questionnaire quality.

Following preliminary trials with a subset of data, ROUGE-L [13] was selected as the primary metric for this analysis, as it provided more consistent and informative results than BLEU [19]. For each questionnaire generated under different experimental settings, ROUGE-L was calculated for all pairs of generated questions and then averaged.

4.2.2. Semantic similarity

A comprehensive semantic evaluation must consider more than the similarity between individual questions. It should also account for the following elements:

- *Question position*: One critical aspect of questionnaire design is the order of the questions, as highlighted by Taherdoost [20]. A common technique, the "funnel" approach, starts with general or broad questions and gradually narrows to specific topics. This method helps avoid biases and ambiguities, facilitating a smoother reasoning process and more effective questionnaires.
- *Generation task*: It is important to remember that the task involves generative models. As a result, the model may generate questions that are relevant to the questionnaire's main topic but do not closely match the ground-truth questions, especially if those sub-topics were not included in the original questionnaire.

To account for these factors, we specifically designed a score that evaluates the similarity between generated questions, ground-truth questions, and the overall questionnaire topic while penalizing deviations from the ideal question order. The defined score, SemSim, is formalized as follows:

$$\text{SemSim} = \frac{\alpha \cdot sim(G, H) + \beta \cdot sim(G, T)}{(\alpha + \beta) - dev(pos(G), pos(H))}, \tag{1}$$

Table 4
Assistant prompt design for JSON conversion.

Prompt part	Content

Converted JSON	<pre>{
	  "data": {
	  "TF_QUESTIONNAIRES": [
	  {
	  "CODE": "ACCESS_TECHNOLOGY_TOOLS",
	  "NAME": "Access to Technology and Tools",
	  "TYPE_ID": 3,
	  "_TF_QUESTIONS": [
	  {
	  "CODE": "Q1",
	  "NAME": "What is your role in the company?",
	  "TYPE_ID": 1,
	  "DISPLAY_ORDER": 1,
	  "_TF_ANSWERS": [
	  {"ANSWER": "Executive/Senior Management"},
	  {"ANSWER": "Manager"},
	  {"ANSWER": "Staff/Employee"},
	  {"ANSWER": "Intern"},
	  {"ANSWER": "Other"}
	  ]
	  },
	  ...</pre>

Table 5
User prompt design for questionnaire generation.

Prompt part	Content
Generation command (variant 1)	Generate me a questionnaire on `<topic>` with `<question_number>` questions.
Generation command (variant 2)	Generate me a questionnaire on `<topic>`.

where G indicates the generated question, H the human-written question, T the questionnaire topic, $sim(X, Y)$ indicates the semantic similarity between elements X and Y, calculated using cosine similarity on their embeddings, α is the weight assigned to the similarity between the generated question G and the human-written question H, β is the weight assigned to the similarity between the generated question G and the questionnaire topic T, $pos(X)$ indicates the position of question X in its respective questionnaire, and $dev(pos(G), pos(H))$ represents the normalized position deviation of the generated question G from the ideal position, given by the human-written question H. This deviation is computed as follows:

$$\frac{|pos(G) - pos(H)|}{\max(N, M)}, \tag{2}$$

where N is the number of questions in the generated questionnaire, and M is the number of questions in the ground-truth questionnaire. This deviation ranges from 0 to 1, with scores closer to 0 indicating that the model generated the question in the correct position and scores closer to 1 indicating significant deviation. SemSim ranges between 0 and 1. Lower scores suggest low weighted cosine similarity or high position deviation, while higher scores indicate substantial similarity and minimal deviation.

To compute SemSim, we first calculate the cosine similarity for every pair of generated and ground-truth questions using OpenAI's `text-embedding-3-large` for embeddings. Then, for each generated question, the SemSim score is computed based on the most similar human-written question, and the

results are averaged for each questionnaire.

4.2.3. Serendipity

As defined by Busch et al. [21], serendipity refers to the occurrence of surprising and valuable discoveries. Its importance spans various fields, including business and computer science. In particular, serendipity is crucial in recommendation systems, where it enhances diversity in users' recommendations, as described by Boldi et al. [22].

Serendipity can be interpreted as the thematic variability within a single questionnaire in the context of questionnaire generation. This variability enriches the content and increases engagement by avoiding repetitive or overly focused questions. Inspired by Boldi et al.'s definition, we adapted the concept of serendipity for our study as follows:

$$\text{Serendipity} = \frac{n}{\min(C, R)} \tag{3}$$

where n represents the number of generated questions relevant to the questionnaire topic, C is the number of possible subtopics generally relevant to the main topic, and R is the total number of generated questions. The serendipity score ranges from 0 to 1. A score closer to 1 indicates that almost every question addresses a different subtopic, contributing to high thematic variability. Conversely, a score closer to 0 suggests lower variability, increasing the risk of duplicate or redundant questions.

Before computing the serendipity scores, we defined relevant subtopics for each questionnaire topic in the HR survey dataset. On average, each topic was associated with 11 subtopics, resulting in 434 subtopics across 39 questionnaire topics, as identified by the Talentia HCM R&D team.

For each questionnaire, duplicate questions were removed based on their cosine similarity, using a threshold of 0.85, chosen empirically. Then, we extracted the subtopic for each generated question using GPT-3.5-Turbo (version `0301`). The zero-shot technique was employed, with both the temperature and frequency penalty parameters set to 0 and the max token value configured at 100.

Next, using `text-embedding-3-large`, we checked if any predefined subtopic (relevant to the current questionnaire topic) had a cosine similarity above 0.5 with the generated question. If so, the question was considered relevant.

4.2.4. Instruction alignment

The temperature and frequency penalty values variation influences the tokens sampled during the generation process. Increasing these values to encourage the model to be more variable and creative can degrade the quality of the generated JSON output. This degradation manifests in the model potentially omitting specified properties or generating text that does not adhere to JSON standards.

4.2.5. Indistinguishability assessment

The rapid advancement of LLMs has raised concerns about their potential, particularly their ability to generate content indistinguishable from that produced by humans. This capability has significant implications across various domains, including HCM.

The indistinguishability assessment was conducted on June 21, 2024, during a Talentia User Group initiative session. The session aimed to introduce new AI features available in the 13th release of Talentia HCM, including the questionnaire generation feature. The meeting, held on Microsoft Teams, involved a subset of proactive Talentia HCM customers.

The test design considered the online submission format and the fact that the participants were neither computer scientists nor familiar with such tests. The test consisted of three pairs of questionnaires, each pair containing one AI-generated questionnaire and one corresponding human-written questionnaire. The selection of these questionnaires was based on specific criteria: the first questionnaire was chosen for its high intra-questionnaire similarity; the second was selected for its strong semantic similarity; and the third was identified as one of the best based on its serendipity measures. Selecting the best

AI-generated questionnaires increased the complexity of the test. Additionally, this served as an initial assessment of the consistency of the designed metrics from a human perspective. The selection was made irrespective of the model, prompting technique, task variant, or hyperparameters.

The final part of the meeting was dedicated to the test. After a brief introduction, the selected pairs were shown to the customers one at a time. For each pair, participants were given 60 seconds to review the questionnaires and then asked to respond to the following questions:

1. `<Topic>` - Which questionnaire is AI Generated?
   a) Questionnaire A;
   b) Questionnaire B.
2. Why do you believe the questionnaire you chose was AI-generated?
   a) Variability of questions;
   b) Variability of answers;
   c) Variability of response types;
   d) Language style;
   e) Questions sequence/order;
   f) Consistency between questions and related answers;
   g) Relevance to topic.

The first question was single-choice, while the second was multi-choice. Although an open-ended question would have been preferable for deeper insights, the main goal was to maintain participants' interest and involvement without overwhelming them.

4.3. Results

With 56 different configurations generated by varying models, hyperparameters, and task variants, the following discussion focuses on aggregated data. Detailed results for each configuration can be found in the project repository.

4.3.1. Content quality

Table 6 presents the scores achieved across various experimental settings, grouped by task, prompting technique, and model:

- For the intra-questionnaire similarity (IQS) values, the mean (μ) and the variance (σ) are reported.
- For the semantic similarity values, the following information is shown:
  - $\mathcal{S}$: The average SemSim score as defined above.
  - $WSQT$ (Weighted Similarity of Questions and Topic): The weighted sum of cosine similarities between generated and ground-truth questions and between generated questions and the questionnaire topic.
  - δ: The average deviation from the ideal position of the generated questions.
  - Δ: The percentage variation between WSQT and the final SemSim, estimating the average influence of position deviation on WSQT.
- For the serendipity values, the mean (μ) and the variance (σ) are reported.

Upon examining the IQS' mean and variance values, it is evident that all tested configurations generally yielded low scores, indicating high variability in the generated questions. Notably, GPT-4-Turbo outperformed GPT-3.5-Turbo, consistently producing lower scores.

SemSim results suggest that the tested models demonstrated a relatively low level of semantic similarity with ground-truth questionnaires, even when considering WSQT scores alone. Moreover, on average, WSQT was penalized by 20.64% due to position deviation, indicating that the models may struggle to generate questions in the correct order.

Table 6
Performance metrics of the tested LLMs, grouped by prompting technique (PT) and task. The reported metrics include intra-questionnaire similarity (IQS), semantic similarity (SemSim), and serendipity (Sdp).

Model	PT	Task	ISQ		SemSim				Sdp	
			μ	σ	$\mathcal{S}$	$WSQT$	δ	Δ	μ	σ
GPT-3.5-Turbo	0S	1	0.34	0.0029	0.45	0.55	0.26	−20.65%	0.75	0.0032
GPT-3.5-Turbo	0S	2	0.32	0.0010	0.48	0.57	0.22	−18.14%	0.76	0.0052
GPT-3.5-Turbo	1S	1	0.27	0.0035	0.48	0.57	0.25	−18.81%	0.80	0.0006
GPT-3.5-Turbo	1S	2	0.27	0.0040	0.47	0.58	0.27	−20.87%	0.80	0.0008
GPT-4-Turbo	0S	1	0.18	0.0006	0.44	0.55	0.28	−21.97%	0.82	0.0005
GPT-4-Turbo	0S	2	0.19	0.0005	0.44	0.56	0.29	−22.49%	0.84	0.0005
GPT-4-Turbo	1S	1	0.21	0.0004	0.46	0.57	0.27	−20.68%	0.84	0.0005
GPT-4-Turbo	1S	2	0.22	0.0010	0.46	0.57	0.27	−21.51%	0.83	0.0006

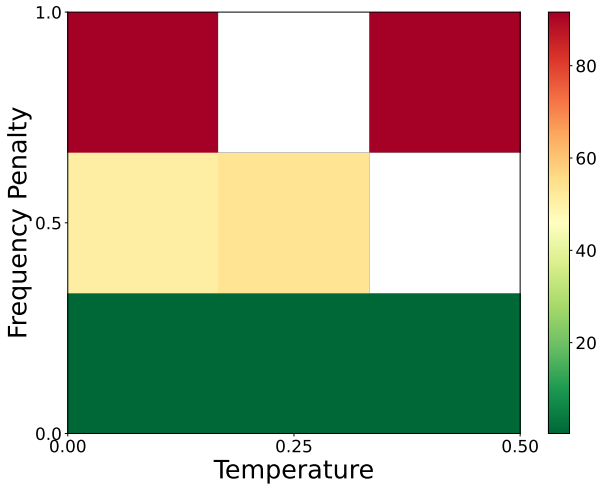

Figure 2: Conversion error rates as a function of temperature (horizontal axis) and frequency penalty (vertical axis). The color gradient indicates the rate of conversion errors, with red representing higher error rates and green representing lower error rates for each combination of values.

High serendipity scores were achieved across all tested configurations, reflecting a satisfactory level of creativity in the models. Notably, the one-shot approach improved serendipity scores for GPT-3.5-Turbo by 5.79%. Additionally, GPT-4-Turbo generally outperformed GPT-3.5-Turbo in generating serendipitous questionnaires, with an average score increase of 6.83%. Furthermore, GPT-4-Turbo demonstrated more consistent results with less variability compared to GPT-3.5-Turbo.

4.3.2. Instruction alignment

Figure 2 illustrates the conversion error rate across all experimented temperature and frequency penalty values combinations, regardless of the model, prompting technique, or task variant. The figure shows that varying the temperature value has minimal impact on the model's ability to maintain instruction alignment in terms of generating valid JSON. In contrast, increasing the frequency penalty significantly affects the model's adherence to the specified structure and the JSON standard, leading to higher conversion error rates.

4.3.3. Indistinguishability assessment

Thirteen customers participated in the assessment, with their responses collected anonymously. In the presented pairs, the majority of customers successfully identified the AI-generated questionnaire.

Table 7
Details of selected AI-generated questionnaires used in the Indistinguishability assessment. Each questionnaire's reference metric score is reported along with the settings used to generate it. Note: T stands for temperature, and FP for frequency penalty.

No.	Topic	Score	Model	Technique	T	FP	Task
1	Kick-off meeting	0.87	GPT-3.5-Turbo	One-shot	0.25	0	2
2	Employee feedback	0.16	GPT-3.5-Turbo	Zero-shot	0.25	0.5	1
3	Stress tolerance	1.00	GPT-4-Turbo	One-shot	0.5	0	2

On average, 8 customers correctly recognized the AI-generated surveys, while 5 mistakenly identified them. Thus, the tested models are still far from deeply imitating human behavior. The details of the selected AI-generated questionnaires are provided in Table 7, with the corresponding human-written questionnaires sourced from the HR survey collection.

Participants who correctly identified the AI-generated questionnaires often pointed to greater variability in the types of questions and highlighted the importance of language style. Conversely, those who misidentified the source focused on the perceived variability of responses and the sequence of questions and answers. Consistency and relevance were crucial across the different questionnaire pairs for those who accurately recognized the AI-generated questionnaires. At the same time, response variability was common among those who misclassified them.

5. Conclusion and future work

Our research focused on the underexplored area of questionnaire generation in Human Resource Management. Due to the scarcity of relevant data, we developed a new collection of HR surveys comprising 79 questionnaires, many of which were enhanced through an augmentation process involving the Talentia HCM R&D team. Using GPT-3.5-Turbo and GPT-4-Turbo, we generated and assessed the quality of these questionnaires from multiple perspectives. Our experiments aimed to identify factors that contribute to higher-quality content, testing various prompting techniques, hyperparameter settings, and task variations to ensure seamless integration into existing HR systems.

One key finding is that increasing the frequency penalty adversely affects the model's ability to adhere to the specified structure, thereby reducing instruction alignment. Based on our results, we recommend keeping the frequency penalty at 0 while slightly increasing the temperature to encourage creativity without compromising structure.

GPT-4-Turbo demonstrated particularly robust results, not only in maintaining engagement but also in generating diverse questions, as reflected in the higher serendipity scores. Additionally, one-shot prompting further enhanced the thematic diversity of the generated surveys. However, a significant issue arose regarding the semantic similarity between generated and ground-truth questionnaires, particularly in ordering questions. We found that changing configurations did not significantly improve the semantic similarity scores.

Finally, the results of the indistinguishability assessment highlighted that variability in answers and linguistic style were key factors distinguishing AI-generated content from human-created questionnaires. Therefore, future work should focus on improving these aspects to increase the performance of LLMs in this context, also involving RAG-based techniques [23] to mitigate hallucinations.

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. Bamber, J. Beltran, P. Boselie, F. Cooke, S. Decker, A. DeNisi, P. Dey, D. Guest, A. Knoblich, A. Malik, J. Paauwe, S. Papagiannidis, C. Patel, V. Pereira, S. Ren, A. Varma, Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT, Human Resource Management Journal 33 (2023) n/a–n/a. doi:10.1111/1748-8583.12524.

[3] H. Jin, Y. Zhang, D. Meng, J. Wang, J. Tan, A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods, CoRR abs/2403.02901 (2024). URL: https://doi.org/10.48550/arXiv.2403.02901. doi:10.48550/ARXIV.2403.02901. arXiv:2403.02901.

[4] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can Large Language Models Transform Computational Social Science?, Comput. Linguistics 50 (2024) 237–291. URL: https://doi.org/10.1162/coli_a_00502. doi:10.1162/COLI_A_00502.

[5] M. E. Spotnitz, B. R. Idnay, E. R. Gordon, R. Shyu, G. Zhang, C. Liu, J. J. Cimino, C. Weng, A Survey of Clinicians' Views of the Utility of Large Language Models, Applied Clinical Informatics 15 (2023) 306–312. URL: https://api.semanticscholar.org/CorpusID:268250530.

[6] A. H. Church, J. Waclawski, Designing and Using Organizational Surveys, Routledge, 1998. URL: https://api.semanticscholar.org/CorpusID:169505746.

[7] T. M. Welbourne, The Potential of Pulse Surveys: Transforming Surveys into Leadership Tools, Employment Relations Today 43 (2016) 33–39. URL: https://api.semanticscholar.org/CorpusID:112257748.

[8] J. Hartley, Employee surveys-Strategic aid or hand-grenade for organisational and cultural change?, International Journal of Public Sector Management 14 (2001) 184–204.

[9] Y. Lei, L. Pang, Y. Wang, H. Shen, X. Cheng, Qsnail: A Questionnaire Dataset for Sequential Question Generation, in: N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 13407–13418.

[10] J. Doughty, Z. Wan, A. Bompelli, J. Qayum, T. Wang, J. Zhang, Y. Zheng, A. Doyle, P. Sridhar, A. Agarwal, C. Bogart, E. Keylor, C. Kültür, J. Savelka, M. Sakr, A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education, in: N. Herbert, C. Seton (Eds.), Proceedings of the 26th Australasian Computing Education Conference, ACE 2024, Sydney, NSW, Australia, 29 January 2024- 2 February 2024, ACM, 2024, pp. 114–123.

[11] R. Rodriguez-Torrealba, E. García-Lopez, A. García-Cabot, End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models, Expert Syst. Appl. 208 (2022) 118258.

[12] H. S. Yun, M. Arjmand, P. R. Sherlock, M. K. Paasche-Orlow, J. W. Griffith, T. W. Bickmore, Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions, CoRR abs/2311.12707 (2023). URL: https://doi.org/10.48550/arXiv.2311.12707. doi:10.48550/ARXIV.2311.12707. arXiv:2311.12707.

[13] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of summaries, 2004, p. 10.

[14] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[15] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, X. Huang, A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models, CoRR abs/2303.10420 (2023). arXiv:2303.10420.

[16] OpenAI, GPT-4 Technical Report, CoRR abs/2303.08774 (2023). arXiv:2303.08774.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot

Learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[19] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, ACL, 2002, pp. 311–318.

[20] H. Taherdoost, Designing a Questionnaire for a Research Paper: A Comprehensive Guide to Design and Develop an Effective Questionnaire, Asian Journal of Managerial Science 11 (2022) 8–16.

[21] C. Busch, Towards a Theory of Serendipity: A Systematic Review and Conceptualization, Journal of Management Studies 61 (2022).

[22] R. Boldi, A. Lokhandwala, E. Annatone, Y. Schechter, A. Lavrenenko, C. Sigrist, Improving Recommendation System Serendipity Through Lexicase Selection, CoRR abs/2305.11044 (2023). arXiv:2305.11044.

[23] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, CoRR abs/2312.10997 (2023). URL: https://doi.org/10.48550/arXiv.2312.10997. doi:10.48550/ARXIV.2312.10997. arXiv:2312.10997.