Open Access

Public Organizations Management

Vol. 14(3), (Series 55): 25-44/ 2026

https://doi.org/10.30473/ipom.2026.75081.5213

E-ISSN: 2538-600X P-ISSN: 2322-522X

ORIGINAL ARTICLE

Validation of the Artificial Intelligence–Driven Delphi Method in Management Research Based on the Turing Test

Hamed Dehghanan[1], Zahra Pouramini[*2]

1. Associate Professor, Department of Management and Accounting, Allameh Tabataba`i University, Tehran, Iran.
2. Ph.D. Student, Department of Management and Accounting, Allameh Tabataba`i University, Tehran, Iran.

***Correspondence**
Zahra Pouramini
E-mail: z.pooramini@ut.ac.ir

Receive Date: 06/July/2025
Revise Date: 15/Oct/2025
Accept Date: 19/Oct/2025

How to cite
Dehghanan, H., Pouramini, Z (2026). Validation of the Artificial Intelligence–Based Delphi Method in Management Research Based on the Turing Test. *Public Organizations Management*, 14(3), 25 -44.

EXTENDED A B S T R A C T

Introduction

In the past decade, the expansion of artificial intelligence (AI) and large language models (LLMs) has brought a fundamental transformation to qualitative research methods and expert-based decision-making processes. The Delphi method, traditionally grounded in the collective judgments of human experts, has entered a new phase of evolution with the advent of AI. The present study aimed to design and validate an AI-based Delphi method while leveraging the imitation game, or Turing test, as a tool to assess the credibility of its outputs. The central question guiding this research was:
"How can the Turing test be utilized as a novel approach to evaluate the validity and similarity of AI-based Delphi results compared to traditional human-based Delphi?"
This question is particularly significant because researchers engaging with modern language models face a fundamental challenge: whether AI-generated responses can be considered as reliable as human judgments in terms of logic, reasoning, and coherence. This study seeks to provide a well-grounded answer and establish a methodological framework for the combined use of Delphi and AI in applied research contexts.

Mothodology

This research is of an applied-developmental nature and was conducted with the objective of validating the AI-based Delphi method through the Turing test. The study population consisted of 30 purposively selected participants, including human resources experts, AI specialists, and independent evaluators, ensuring the data could be examined from human, technical, and impartial judgment perspectives.

In prior studies, the AI-based Delphi methodology was comprehensively designed and explained, and in a separate study, it was applied to prioritize the developmental needs of human resources managers in the context of coaching. In the current study, the method is only briefly introduced, and the data derived from previous research serve as input for the Turing test.

In the AI-based Delphi process, large language models (ChatGPT, Copilot, and Gemini) acted as virtual experts, responding to research questions and presenting their reasoning explicitly through structured prompts. Subsequently, the outputs of these models were validated using the Turing test. Evaluators engaged in brief dialogues without knowing the source of the responses (human or AI) and were required to determine the nature of the respondent. Responses that could not be confidently distinguished were considered "successful" in the Turing test.

Thus, this study represents a continuation of prior research, examining the similarity, coherence, and credibility of AI-based Delphi outputs in a real-world research setting using actual data and the Turing test.

Findings

Findings indicated that 56% of participants were female and 44% male, with a mean age of 36 years (range: 27–52) and approximately 72% holding a master's degree or higher. This diversity of expertise and experience provided a robust basis for analyzing AI model behavior in comparison with human responses. The data employed were secondary sources, comprising results from three rounds of traditional and AI-based Delphi in previous studies by the authors, focused on identifying developmental priorities for human resources managers. These datasets included competency ratings, comparisons between traditional and AI-based approaches, and reliability and validity matrices, with over 85% of the data confirmed as reliable and valid.

During the study, key questions were collected from both human and AI sources at each Delphi stage, and evaluators were tasked with distinguishing between human and machine responses, in line with the Turing test framework. Qualitative analysis revealed that AI-generated responses were often so similar to human answers that evaluators struggled to distinguish them, with an overall Turing test success rate of only 23%, indicating that over two-thirds of responses could not be correctly classified. This finding highlights the strong ability of the models to simulate human behavior.

Content analysis showed that evaluators primarily relied on cues such as grammatical and stylistic errors, vocabulary diversity, interactive responsiveness, display of emotion and empathy, variation in tone, and references to personal experience—features more prevalent in human responses. In contrast, AI responses were characterized by high linguistic precision, formal structure, uniform tone, accurate and error-free answers, consistent response speed, and absence of personal narratives or experiential examples. Behavioral and linguistic patterns further revealed that participants expected human responses to be more dynamic, personal, and occasionally non-linear, whereas machine responses were predominantly formal, precise, and structured.

Discussion and Conclusion

These findings not only confirm the capability of large language models to emulate human-like behavior and response style but also underscore the significance of linguistic, behavioral, and emotional cues in distinguishing between humans and AI. Furthermore, integrating human and AI Delphi data and analyzing evaluator judgments demonstrated that AI models could produce highly coherent responses, logically consistent themes, and alignment with management literature, establishing their effectiveness as tools for analyzing developmental needs of human resources managers.

Overall, the results indicate that the AI-based Delphi method, combined with the Turing test, can effectively evaluate the validity and similarity of machine-generated responses to human ones, enabling the production of reliable and credible data in real-world research settings.

As the first study to employ the Turing test for evaluating a research methodology, this study demonstrates that AI-based Delphi can generate data that are reliable and highly similar to human responses. The findings show that

large language models successfully emulate human behavior and response style, making it challenging for participants to distinguish human from machine responses. This evidence strengthens the credibility of the AI-based Delphi method, suggesting it can serve as an effective research tool without compromising the quality of human-like analysis.

Based on the findings, practical recommendations include optimizing the design of Turing test scenarios and criteria to more accurately assess the capabilities of language models, integrating the Delphi method with the Turing test for practical validation of outputs, fully documenting the process and managing potential biases, attending to ethical considerations and data confidentiality, and developing and standardizing prompt-engineering skills for effective interaction with language models. Additionally, potential limitations—such as the influence of participants' attitudes and experiences, ethical and confidentiality constraints, sensitivity of results to prompt design and model characteristics, limitations in sample diversity and size, and the inability to fully control the knowledge and biases embedded in language models—should be considered in the design and analysis of future studies.

These findings provide practical guidance for researchers employing AI-based Delphi methods, enabling the combined use of human and AI collective intelligence with higher accuracy and reliability.

K E Y W O R D S
Artificial Intelligence, Delphi Method, Turing Test, Personal Development.

فصلنامه علمی

مدیریت سازمان‌های دولتی

سال چهاردهم، شماره ۳، پیاپی ۵۵، تابستان ۱۴۰۵ (۴۴-۲۵)

https://doi.org/10.30473/ipom.2026.75081.5213

E-ISSN: 2538-600X P-ISSN: 2322-522X

«مقاله پژوهشی»

اعتبارسنجی روش دلفی مبتنی بر هوش مصنوعی در پژوهش‌های مدیریتی براساس آزمون تورینگ

حامد دهقانان[۱]، زهرا پورامینی*[۲]

۱. دانشیار گروه مدیریت بازرگانی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران.

۲. دانشجوی دکتری، گروه رفتار سازمانی و منابع انسانی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران.

***نویسنده مسئول:** زهرا پورامینی

رایانامه: z.pooramini@ut.ac.ir

تاریخ دریافت: ۴۰۴/۰۴/۱۵

تاریخ بازنگری: ۱۴۰۴/۰۷/۲۳

تاریخ پذیرش: ۱۴۰۴/۰۷/۲۷

استناد به این مقاله:

دهقانان، حامد؛ پورامینی، زهرا (۱۴۰۵).

اعتبارسنجی روش دلفی مبتنی بر هوش مصنوعی در پژوهش‌های مدیریتی بر اساس آزمون تورینگ.

فصلنامه علمی مدیریت سازمان‌های دولتی،

۱۴(۳)، ۴۴-۲۵.

چکیده

این پژوهش با هدف طراحی و اعتبارسنجی روش دلفی مبتنی بر هوش مصنوعی بر اساس بازی تقلید (تست تورینگ) انجام شد. سؤال اصلی تحقیق این بود که چگونه می‌توان از تست تورینگ به‌عنوان روشی نوین برای سنجش اعتبار نتایج دلفی مبتنی بر هوش مصنوعی استفاده کرد. این مطالعه از نوع کاربردی-توسعه‌ای است و جامعه پژوهش شامل ۳۰ نفر از متخصصان منابع انسانی، متخصصان هوش مصنوعی و ارزیابان مستقل بود. داده‌ها به‌صورت ثانویه گردآوری شد و خروجی سه مدل زبانی Microsoft Copilot، ChatGPT و Gemini به‌عنوان ورودی در تست تورینگ برای تحلیل میزان شباهت نتایج دلفی انسانی و هوش مصنوعی مورد استفاده قرار گرفت. این داده‌ها از مقالات قبلی نویسندگان که نیازهای توسعه‌ای مدیران منابع انسانی را با روش دلفی احصا کرده بودند، استخراج شده‌اند. این پژوهش به‌عنوان اولین مطالعه در زمینه استفاده از آزمون تورینگ برای ارزیابی یک روش پژوهشی، نشان داد که دلفی مبتنی بر هوش مصنوعی قادر است داده‌هایی تولید کند که از نظر شباهت به پاسخ‌های انسانی پایا و قابل اعتماد هستند. نتایج حاکی از آن است که مدل‌های زبانی بزرگ رفتار و سبک پاسخ‌دهی انسانی را به‌خوبی شبیه‌سازی کرده‌اند و تمایز بین پاسخ‌های انسانی و ماشینی برای شرکت‌کنندگان دشوار بوده است. این یافته‌ها اعتبار روش دلفی مبتنی بر هوش مصنوعی را تقویت کرده و نشان می‌دهد که این روش می‌تواند به‌عنوان ابزاری مؤثر در پژوهش، بدون کاهش کیفیت تحلیل انسانی، مورد استفاده قرار گیرد.

واژه‌های کلیدی

هوش مصنوعی، روش دلفی، تست تورینگ و توسعه فردی.

مقدمه

ابزارهای هوش مصنوعی با سرعت چشمگیری وارد عرصه پژوهش‌های علمی شده و در بهبود و تسهیل طیف وسیعی از فعالیت‌های علمی نقش داشته‌اند. این فناوری‌ها امکان تحلیل و تفسیر داده‌ها و تصاویر، تولید فرضیه، مدل‌سازی پدیده‌های پیچیده و طراحی روش‌ها و مواد جدید را فراهم می‌کنند. همچنین می‌توانند داده‌هایی برای اعتبارسنجی فرضیه‌ها و مدل‌ها تولید کرده، ادبیات علمی را مرور و ارزیابی کنند، مقالات و پیشنهادهای پژوهشی را نگارش و ویرایش نمایند و در بررسی خروجی‌های پژوهشی به پژوهشگران یاری رسانند (بودرا، کاوو، چرنی و آرورا[1]، ۲۰۲۳؛ بورووییچ و همکاران[2]، ۲۰۲۲).

این فناوری، به‌ویژه در حوزه‌هایی مانند یادگیری ماشین، تحلیل داده‌های حجیم و پردازش زبان طبیعی، امکان شناسایی الگوها، تحلیل دقیق داده‌ها و ارائه بازخوردهای هوشمند را در اختیار پژوهشگران و متخصصان قرار می‌دهد. علاقه علمی به نقش مکمل یا جایگزین هوش مصنوعی در فعالیت‌های تحلیلی و تصمیم‌گیری نیز به‌طور چشمگیری افزایش یافته است (وان کراگ، ۲۰۱۸). به نظر می‌رسد کاربردهای هوش مصنوعی در پژوهش و علم تقریباً بی‌حد و مرز باشد و در دهه پیش‌رو، این فناوری می‌تواند فرآیند کشف و نوآوری علمی را به‌طور بنیادین متحول کند (آلوارادو[3]، ۲۰۲۳). بر این اساس، فرض بر آن است که با توجه به محدودیت‌های انسانی در پردازش حجم زیاد داده و تحلیل هم‌زمان متغیرهای پیچیده، ابزارهای هوش مصنوعی می‌توانند کارایی، دقت و سرعت تحلیل را افزایش داده و در کنار تخصص انسانی، نتایج بهتری برای فرآیندهای پژوهشی و به‌ویژه پژوهش‌های مدیریتی فراهم آورند (باگین و همکاران، ۲۰۱۷). در این راستا، روش دلفی به‌عنوان یک رویکرد ساختاریافته برای جمع‌بندی نظرات کارشناسان و دستیابی به اجماع، یکی از ابزارهای متداول در پژوهش‌های مدیریتی و تصمیم‌گیری است که قابلیت ترکیب با هوش مصنوعی را نیز دارا می‌باشد.

روش دلفی بیش از نیم قرن است که به‌عنوان یک روش پذیرفته‌شده برای ایجاد اجماع ساختاریافته مورد استفاده قرار می‌گیرد (ناسا، جین و جونیجا[4]، ۲۰۲۱؛ نیدربرگ و اسپرنگر[5]، ۲۰۲۰). این روش که در ابتدا برای پیش‌بینی فناوری طراحی شده بود، به مرور به یکی از ابزارهای اصلی پژوهش‌های حوزه‌های مختلف تبدیل شده است (اسپید و متوالی[6]، ۲۰۲۵). با این حال، با وجود کاربرد گسترده و ارزش پایدار روش دلفی سنتی، این روش با مجموعه‌ای از محدودیت‌های عملی و روش‌شناختی مواجه است که می‌توانند قابلیت اجرا، کیفیت قضاوت‌ها و تفسیرپذیری یا کاربردپذیری نتایج آن را کاهش دهند (فینک-هانیفر، داگن، دوشاک، نوواک[7]، ۲۰۱۹)

یکی از موانع اصلی، نیاز این روش به منابع گسترده برای گردآوری و مدیریت پنل‌های بزرگ کارشناسان است؛ در برخی مطالعات، تعداد شرکت‌کنندگان بیش از ۱۰۰ نفر گزارش شده است. حفظ مشارکت طی چندین دور نیز دشوار است و نرخ ترک شرکت‌کنندگان در برخی مطالعات بیش از ۹۰ درصد بوده است (شانگ[8]، ۲۰۲۳). این چالش‌ها نه تنها پایایی و اعتبار اجماع را کاهش می‌دهند، بلکه دسترسی به روش را برای کارشناسان محدود به زمان یا بسیار تخصصی دشوار می‌کنند (اسپید و متوالی، ۲۰۲۵). علاوه بر این، ناهماهنگی‌های روش‌شناختی میان مطالعات دلفی، به‌ویژه در تعریف اجماع و پردازش داده‌های کیفی، قابلیت بازتولید و امکان مقایسه نتایج را کاهش می‌دهد (فینک-هانیفر، داگن، دوشاک، نوواک، ۲۰۱۹). تمام این محدودیت‌ها نیاز به نوآوری را برجسته می‌کنند. الگوریتم‌هایی مانند یادگیری ماشین و پردازش زبان طبیعی می‌توانند در شناسایی دقیق‌تر کارشناسان، خودکارسازی تحلیل‌ها، کاهش سوگیری‌های انسانی و بهبود ساختار بازخورد در فرآیندهای دلفی نقش مؤثری ایفا کنند. علاوه بر این، این فناوری‌ها امکان پردازش داده‌های غیر ساختاریافته و اجرای هوشمند و پویا چندین دور ارزیابی را می‌آورند (مولر[9]، تورینگ، کلاکنر و لارسن، ۲۰۲۴). این مطالعه پیشنهاد می‌کند که هوش مصنوعی مولد در صورتی که در چارچوب‌های دقیق تعریف‌شده و تحت هدایت کارشناسان استفاده شود، می‌تواند ابزاری قدرتمند برای رفع این محدودیت‌های دیرینه باشد، نه از طریق خودکارسازی استدلال کارشناسان، بلکه از طریق تقویت استراتژیک آن‌ها (اسپید و متوالی، ۲۰۲۵). در مجموع می‌توان گفت؛ ترکیب توانمندی‌های روش دلفی با قابلیت‌های هوش مصنوعی چارچوبی نوآورانه و قدرتمند برای ارتقای کیفیت تصمیم‌گیری و آینده‌پژوهی فراهم می‌آورد؛ به‌ویژه در حوزه‌هایی مانند مربیگری توسعه فردی که نیازمند تحلیل عمیق، بازخورد مستمر و جمع‌بندی تخصصی هستند. این روش نوآورانه ضمن معرفی نیازمند، روشی برای اعتبارسنجی نتایج

1. Bothra
2. Borowiec
3. Alvarado
4. Nasa
5. Niederberger
6. Speed
7. Fink-Hafner
8. Shang
9. Mueller

است تا قابلیت تعمیم و اعتمادپذیری یافته‌ها به‌طور نظام‌مند مورد ارزیابی قرار گیرد. این موضوع نه تنها محدودیت‌های سنتی دلفی در بازتولیدپذیری و سوگیری انسانی را کاهش می‌دهد، بلکه امکان مقایسه نتایج در شرایط مشابه و زمینه‌های متفاوت را نیز فراهم می‌سازد. به این ترتیب، اعتبارسنجی در دلفی مبتنی بر هوش مصنوعی، نقشی کلیدی در تضمین کارایی و اعتمادپذیری یافته‌ها ایفا می‌کند. از این‌رو این پژوهش درصدد است تا ضمن معرفی روش دلفی، نحوه اعتباریابی نتایج آن از طریق تست تورینگ را معرفی نماید:

• چگونه می‌توان از تست تورینگ به‌عنوان روشی نوین برای اعتباریابی نتایج دلفی مبتنی بر هوش مصنوعی استفاده کرد؟

مبانی نظری

پژوهش مبتنی بر هوش مصنوعی

پژوهش مبتنی بر هوش مصنوعی ویژگی‌های خاصی دارد که آن را از روش‌های سنتی متمایز می‌کند. استفاده از این فناوری‌ها می‌تواند به پژوهشگران کمک کند تا به تحلیل‌های پیچیده‌تر، دقیق‌تر و سریع‌تری دست یابند. ویژگی‌های اصلی پژوهش مبتنی بر هوش مصنوعی عبارت‌اند از:

• **پردازش داده‌های بزرگ:** یکی از ویژگی‌های برجسته پژوهش‌های مبتنی بر هوش مصنوعی، توانایی پردازش حجم وسیعی از داده‌ها است. داده‌های بزرگ که شامل داده‌های ساختاریافته و غیرساختاریافته هستند، می‌توانند توسط الگوریتم‌های AI پردازش شوند. این امکان به پژوهشگران اجازه می‌دهد تا روابط پیچیده و الگوهای پنهان در داده‌ها را شناسایی کنند که در روش‌های سنتی ممکن نبود (شونبرگر و کوکی[1]، ۲۰۱۳).

• **دقت و صحت بالاتر:** الگوریتم‌های هوش مصنوعی، به ویژه الگوریتم‌های یادگیری ماشین و یادگیری عمیق، می‌توانند الگوهای دقیق و پیچیده‌ای را از داده‌ها استخراج کنند که معمولاً برای انسان‌ها قابل شناسایی نیست. این ویژگی به ویژه در مواردی مانند پیش‌بینی‌ها، طبقه‌بندی‌ها و شبیه‌سازی‌های پیچیده کاربرد دارد (میتچل و جوردن[2]، ۲۰۱۵).

• **خودکارسازی فرآیندهای تحقیق:** در پژوهش‌های مبتنی بر هوش مصنوعی، بسیاری از فرآیندهای تحقیق مانند جمع‌آوری داده‌ها، پردازش اولیه داده‌ها و حتی تحلیل‌های پیچیده می‌توانند به‌طور خودکار انجام شوند. این خودکارسازی موجب افزایش سرعت و کاهش خطاهای انسانی در فرآیند تحقیق می‌شود (راسل و نرویج[3]، ۲۰۱۶).

• **یادگیری از داده‌ها و بهبود مداوم:** یکی از ویژگی‌های منحصربه‌فرد هوش مصنوعی، توانایی یادگیری از داده‌ها و بهبود مستمر مدل‌ها است. مدل‌های یادگیری ماشین می‌توانند به‌طور خودکار به‌روزرسانی شوند و عملکرد خود را بر اساس داده‌های جدید بهبود بخشند. این ویژگی به پژوهشگران این امکان را می‌دهد که مدل‌های خود را با گذشت زمان دقیق‌تر کنند (شوآرتز و دیوید، ۲۰۱۴).

• **تحلیل‌های پیش‌بینی‌کننده و شبیه‌سازی:** هوش مصنوعی این امکان را فراهم می‌کند که پیش‌بینی‌هایی دقیق از روندهای آینده انجام شود و یا شبیه‌سازی‌هایی پیچیده از فرآیندهای مختلف صورت گیرد. این ویژگی به‌ویژه در زمینه‌هایی مانند پزشکی، علوم اجتماعی و اقتصاد که پیش‌بینی روندها بسیار مهم است، کاربرد دارد (رایتینگ، فرانک و هال، ۲۰۱۱).

• **توانایی تحلیل داده‌های غیرساختاریافته:** یکی از چالش‌های بزرگ در پژوهش‌های سنتی، تحلیل داده‌های غیرساختاریافته مانند متون، تصاویر و ویدئوها بود. هوش مصنوعی با استفاده از تکنیک‌های مانند پردازش زبان طبیعی و تحلیل تصاویر، به پژوهشگران این امکان را می‌دهد که از این نوع داده‌ها به‌طور مؤثر استفاده کنند (شونبرگر و کوکی[4]، ۲۰۱۳).

این ویژگی‌ها نشان‌دهنده پتانسیل‌های عظیم پژوهش‌های مبتنی بر هوش مصنوعی هستند و نشان می‌دهند که این روش‌ها می‌توانند به‌طور مؤثری محدودیت‌های روش‌های سنتی را پشت سر بگذارند.

روش دلفی مبتنی بر هوش مصنوعی

رشد چشمگیر فناوری‌های نوین، به‌ویژه در حوزه‌های یادگیری ماشینی و داده‌کاوی، باعث دگرگونی بنیادین در الگوهای پژوهشی شده و مسیر تحقیقات را از رویکردهای سنتی به پژوهش‌های مبتنی بر هوش مصنوعی سوق داده است .بر اساس دیدگاه کلی[5] (مویسیس[6]، ۲۰۲۳)، هوش مصنوعی یکی از فناوری‌های کلیدی و تحول‌آفرین عصر حاضر به شمار می‌رود که به دلیل توانایی آن در ارتقای ظرفیت‌های انسانی با هزینه‌ای اندک، نقش تعیین‌کننده‌ای در شکل‌دهی آینده تمدن بشری خواهد داشت. از سوی دیگر، این فناوری از طریق

1. Schönberger
2. Mitchell
3. Russell
4. Schönberger
5. Kelly
6. Moises

الگوریتم‌های یادگیری ماشینی، امکان تحلیل و پیش‌بینی در شرایط پیچیده و نامطمئن را فراهم می‌کند (پاتاک و همکاران[1]، ۲۰۱۸). در همین راستا، یکی از مسیرهای نوظهور در تحقیقات آینده‌پژوهی، به‌کارگیری روش دلفی مبتنی بر هوش مصنوعی است که در این مطالعه برای بررسی تحولات آینده در حوزه هوش مصنوعی مولد مورد استفاده قرار گرفته است. این روش نسخه‌ای ساده‌شده و تکرارپذیر از دلفی سنتی به شمار می‌رود که در قالب یک فرآیند تصادفی طراحی شده است؛ فرآیندی که در آن ورودی شامل پرسش‌های باز، نقش‌ها و موضوعات پژوهشی بوده و خروجی آن پیش‌بینی‌های کیفی در زمینه مورد مطالعه است (ماری و برتولوتی[2]، ۲۰۲۵).

در این مدل، مراحل روش دلفی کلاسیک با کمک هوش مصنوعی شبیه‌سازی شده و از پرامپت‌نویسی هدفمند، تحلیل چرخشی داده‌ها و پردازش چندلایه برای تولید و پالایش دانش استفاده می‌شود. مشابه روش سنتی، دلفی مبتنی بر هوش مصنوعی نیز فرآیندی تکرارشونده و ساختارمند است که در آن موضوع، نقش‌ها و پرسش‌ها به مدل هوش مصنوعی ارائه و پاسخ‌ها دریافت می‌شود. خلاصه ساختاری از روش دلفی مبتنی بر هوش مصنوعی در زیر آورده شده است:

– عامل سازمان دهنده[3] معادل تسهیل گر در روش کلاسیک: جمع‌آوری پاسخ‌ها، بازنویسی آن‌ها، طراحی پرسشنامه‌ها (باز و بسته)، استخراج سؤالات جدید از پاسخ‌ها؛

– عامل پاسخ‌دهنده[4] معادل کارشناس در روش کلاسیک: پاسخ به سؤالات باز یا بسته، در قالب متنی یا عددی؛

– طراحی سؤالات باز[5]: تولید حداکثری ایده‌های نو؛

– اجرای سؤالات باز[6]: اجرای سؤالات باز از طریق پرامپت؛

– نظرسنجی[7]: پاسخ به سؤالات در قالب طیف تعریف شده؛

– تکرار فرآیند[8]: اجرای فرآیند بدون محدودیت (ماری و برتولوتی، ۲۰۲۵)؛

تست تورینگ

این آزمون به ما کمک می‌کند تا درک کنیم چه عواملی در فرآیند شبیه‌سازی رفتار انسانی مؤثرند؛ عواملی همچون اندازه و کارایی مدل، شیوه‌های پرامپت‌نویسی، زیرساخت‌های پشتیبان مانند دسترسی تورینگ در سال ۱۹۵۰ «بازی تقلید[9]» را برای بررسی این پرسش مطرح کرد که «آیا ماشین‌ها می‌توانند فکر کنند؟». در این بازی، یک بازجو باید تنها بر اساس مکاتبه متنی تشخیص دهد کدام‌یک از دو پاسخ‌دهنده انسان است و کدام ماشین. تورینگ این روش را معیاری جامع برای سنجش هوش در نظر گرفت. این آزمون، بعدها به نام «آزمون تورینگ» شهرت یافته داده‌های لحظه‌ای و همچنین تجربه و مهارت فرد بازجو (جونز و برگر، ۲۰۲۴). علاوه بر آن؛ آزمون تورینگ چارچوبی برای بررسی درک مفهومی رایج از «انسان مانند بودن» فراهم می‌کند. این آزمون صرفاً ماشین‌ها را ارزیابی نمی‌کند، بلکه به‌طور ضمنی مفروضات فرهنگی، اخلاقی و روان‌شناختی شرکت‌کنندگان انسانی را نیز می‌سنجد (تورکلی[10]، ۲۰۱۱). هنگامی که بازجویان پرسش‌ها را طراحی و اصلاح می‌کنند، در واقع باورهای خود را درباره ویژگی‌هایی که ذات انسانی را تشکیل می‌دهند و نیز آن دسته از ویژگی‌هایی که تقلیدشان دشوارتر است، آشکار می‌سازند (جونز و برگر، ۲۰۲۴).

شیوه اجرای تست تورینگ

هوش مصنوعی با سرعتی بی‌سابقه در حال گسترش است و درک چگونگی تعامل انسان با سیستم‌های هوشمند اهمیت حیاتی دارد، زیرا این نوع تعامل‌ها در شکل‌دهی آینده زندگی بشر نقش اصلی خواهند داشت. در آینده نزدیک، انسان‌ها باید مهارتی تازه بیاموزند: توانایی تشخیص تفاوت میان انسان و ماشین. این مهارت به ما کمک می‌کند تا احساس کنترل و فهم خود از دنیای واقعی را حفظ کنیم. اگر ماشین را به‌اشتباه انسان بدانیم (انسان‌انگاری) یا انسان را مانند ماشین ببینیم (غیر انسان‌سازی)، ممکن است با پیامدهای روانی، اجتماعی و اخلاقی جدی روبه‌رو شویم (اوهلیا، آلموسدی و فودر[11]، ۲۰۲۲).

در این راستا؛ امروزه مدل‌های زبانی بزرگ برای بازی تورینگ به‌خوبی طراحی شده‌اند. این مدل‌ها متنی روان و شبیه به زبان طبیعی تولید می‌کنند و در طیفی از وظایف زبانی تقریباً هم‌سطح با انسان عمل می‌کنند (چانگ و برگن۱۲، ۲۰۲۳). این آزمون به این صورت انجام می‌گیرد که یک شخص به عنوان قاضی، با یک ماشین و یک انسان به گفتگو می‌نشیند و سعی در تشخیص ماشین از انسان دارد. در صورتی که ماشین بتواند قاضی (بازجو) را به گونه‌ای بفریبد که در قضاوت خود دچار اشتباه شود، توانسته است آزمون را با موفقیت پشت سر بگذارد

1. Pathak
2. Bertolotti
3. Organizing Agent
4. Responding Agent
5. Open Questions Definition
6. Open Question Erogation
7. Survey Questions Erogation
8. Multiple Rounds
9. Imitation Game
10. Turkle
11. Ujhelyi
12. Chang

(جونز و برگر، ۲۰۲۴). این تست به روش‌های مختلفی قابل اجرا است که در زیر به برخی از آن‌ها اشاره شده است:

– جایزه لوبنر[1] (شایبر، ۱۹۹۴): در سال ۱۹۹۴ توسط هیو لوبنر بنیان‌گذاری شد و به‌عنوان اولین رقابت رسمی بر اساس آزمون تورینگ شناخته می‌شود. ایده اصلی این جایزه این بود که سیستم‌های هوش مصنوعی در یک گفت‌وگوی متنی با داوران انسانی شرکت کنند و تلاش کنند آن‌ها را متقاعد سازند که انسان هستند.

– مطالعه جانی و همکاران (۲۰۲۳) تست تورینگ (۱۹۵۰) بازی تقلید۲ را روی پلتفرم انجام دادند که در آن شرکت‌کنندگان پس از گفت‌وگوی کوتاه باید تشخیص می‌دادند آیا طرف مقابلشان یک مدل زبان بزرگ است یا یک انسان.

– جونزو برگر (۲۰۲۴) نیز تست تورینگ را در یک پلتفرم شبکه اجتماعی اجرا کرد که در آن، شرکت‌کنندگان پنج دقیقه زمان داشتند تا با مدل یا انسان به‌صورت متنی گفتگو کنند و سپس باید تصمیم می‌گرفتند که کدام یک انسان و کدام یک مدل است.

- مطالعه حاضر هم تست تورینگ را این‌گونه پیاده‌سازی نموده است که گفتگوهای متنی حاصل میان انسان و هوش مصنوعی در اختیار افراد قرار می‌دهد تا صرفاً از روی متن، انسان یا ماشین بودن طرف مقابل را تشخیص دهند (اوهلیا، آلموسدی و فودر، ۲۰۲۲). این موضوع در قسمت روش‌شناسی به‌صورت مفصل توضیح داده شده است.

پیشینه‌های پژوهش

با وجود گسترش استفاده از هوش مصنوعی در پژوهش، مطالعات بسیار کمی به‌کارگیری آزمون تورینگ در روش دلفی و اعتبارسنجی آن را بررسی کرده‌اند. در این پژوهش، برای نخستین بار در ایران، از آزمون تورینگ به‌منظور اعتبارسنجی روش دلفی مبتنی بر هوش مصنوعی استفاده شده است؛ بنابراین، این مطالعه از معدود پژوهش‌هایی است که به‌طور هم‌زمان به کاربرد آزمون تورینگ در روش دلفی و اولویت‌بندی نیازهای توسعه‌ای مدیران با استفاده از دلفی هوش مصنوعی پرداخته است. در این راستا پیشینه زیادی در این حوزه حتی در مطالعات خارجی نیز وجود ندارد:

جونز و برگر (۲۰۲۴) در پژوهشی تحت عنوان «آیا چت جی پی تی از تست تورینگ را عبور می‌کند؟» به بررسی عملکرد GPT-4 در یک آزمون تورینگ آنلاین پرداخت. نتایج نشان داد GPT-4 در ۴۹,۷٪ از موارد موفق به فریب شرکت‌کنندگان شد. تصمیم شرکت‌کنندگان بیشتر بر اساس سبک زبانی و ویژگی‌های اجتماعی–هیجانی گرفته شد که نشان می‌دهد هوش به‌تنهایی برای عبور از آزمون تورینگ کافی نیست. همچنین، آگاهی و تجربه بیشتر کاربران دقت تشخیص هوش مصنوعی را افزایش داد. پژوهش نتیجه گرفت که آزمون تورینگ همچنان ابزاری ارزشمند برای سنجش کیفیت ارتباط طبیعی و توانایی فریب مشابه انسان است.

اوهلیا، آلموسدی و فودر (۲۰۲۲)، در مقاله‌ای تحت عنوان «آیا از آزمون تورینگ عبور می‌کند؟ عوامل مؤثر بر تصمیم‌گیری در آزمون تورینگ» به بررسی عواملی پرداختند که بر قضاوت انسان‌ها در تشخیص انسان یا ماشین بودن شریک گفت‌وگو تأثیر می‌گذارند. پژوهش با الهام از آزمون تورینگ و در دو مطالعه انجام شد. نتایج نشان داد که هرچه تفاوت نگرش میان طرفین گفت‌وگو بیشتر باشد، احتمال تشخیص طرف مقابل به‌عنوان ماشین افزایش می‌یابد. همچنین، متغیرهایی مانند دستور زبان صحیح، منطق پاسخ‌ها و شباهت نگرش‌ها بیشترین تأثیر را در قضاوت انسانی بودن دارند. این پژوهش ضمن تبیین نگرش انسان‌ها نسبت به ماشین‌ها، به پیامدهای غیرانسانی‌سازی[3] نیز اشاره دارد.

حدود[4] (۲۰۲۵)، در مطالعه‌ای تحت عنوان «ادغام هوش مصنوعی در روش‌های پژوهش: بررسی سوگیری‌های احتمالی و راهکارهای کاهش آن» به بررسی چالش‌ها و فرصت‌های به‌کارگیری هوش مصنوعی در تحقیقات علمی پرداخته است. این مطالعه با تمرکز بر جنبه‌های اخلاقی و مسئله سوگیری در داده‌ها و الگوریتم‌ها، بر ضرورت افزایش دقت، شفافیت و مسئولیت‌پذیری پژوهشگران تأکید دارد. همچنین، پیشنهادهایی برای کاهش سوگیری، ارتقای آگاهی اخلاقی و طراحی برنامه‌های آموزشی جهت بهبود مهارت‌های پژوهشگران در استفاده از هوش مصنوعی ارائه شده است.

برتولوتی و ماری (۲۰۲۴)، در پژوهشی تحت عنوان «مطالعه دلفی مبتنی بر مدل‌های زبانی» بر نقش کلیدی هوش مصنوعی در ارتقای کیفیت مطالعات دلفی تأکید کرده‌اند. آنان بیان می‌کنند که تعیین شفاف نقش‌ها، تعریف دقیق مسئله پژوهش و به‌کارگیری پرامپت‌های تخصصی می‌تواند به افزایش دقت، انسجام و کارآمدی فرایند دلفی در استفاده از مدل‌های زبانی منجر شود.

آیتال و آیتال (۲۰۲۳)، در پژوهشی تحت عنوان

1. Loebner Prize
2. Imitation Game
3. Dehumanization
4. Hudoud

«به‌کارگیری هوش مصنوعی در روش‌های پژوهش تجربی، میدانی و اکتشافی» به بررسی موضوع پرداختند. نتایج نشان داد این فناوری‌ها در طراحی، تحلیل و تفسیر داده‌ها مؤثرند و باعث ارتقاء کیفیت و ارزش نتایج پژوهش می‌شوند.

با توجه به مرور مطالعات پیشین، با وجود پیشرفت گسترده هوش مصنوعی در حوزه‌های مختلف، کاربرد آن در فرایندهای پژوهشی به‌طور عملی و سیستماتیک هنوز کمتر مورد توجه قرار گرفته است. بیشتر مطالعات بر اهمیت هوش مصنوعی در بهبود عملکرد سازمانی و تصمیم‌گیری اشاره کرده‌اند، اما استفاده از مدل‌های زبانی در روش‌های تحقیق و اعتبارسنجی آن، به‌ویژه با بهره‌گیری از آزمون تورینگ، تاکنون کمتر بررسی شده است.

پژوهش حاضر برای نخستین بار در ایران از آزمون تورینگ به‌عنوان ابزاری برای اعتبارسنجی روش دلفی مبتنی بر هوش مصنوعی استفاده می‌کند و می‌تواند یکی از مطالعات اولیه در زمینه اولویت‌بندی نیازهای توسعه‌ای مدیران منابع انسانی با روش دلفی هوش مصنوعی باشد. نوآوری‌های این مطالعه شامل:

– معرفی و کاربرد روش دلفی ترکیبی در حوزه مدیریت منابع انسانی که در آن مدل‌های زبانی به‌عنوان اعضای خبرگان وارد فرایند می‌شوند و امکان تلفیق خرد جمعی انسان و توانایی‌های هوش مصنوعی فراهم می‌گردد.

– اجرای عملی روش دلفی در دنیای واقعی، شامل جمع‌آوری داده‌ها و اولویت‌بندی نیازهای توسعه‌ای مدیران با بهره‌گیری از قابلیت‌هایی مانند حافظه بلندمدت و زنجیره تفکر در مدل‌های زبانی.

– اعتبارسنجی روش از طریق آزمون تورینگ که توانایی مدل‌ها در بازتولید استدلال‌های انسانی و کیفیت تصمیم‌گیری شرکت‌کنندگان را می‌سنجد.

– ارتقای مهارت پرامپت‌نویسی با ارائه نمونه‌های عملی و قابل اجرا که به بهبود دقت و انسجام خروجی‌ها کمک می‌کند.

در مجموع، این پژوهش در خلاء مطالعاتی موجود جای می‌گیرد، می‌تواند مبنایی برای توسعه روش‌شناسی‌های پژوهشی مبتنی بر هوش مصنوعی باشد و حدومرز استفاده از هوش مصنوعی در پژوهش‌ها را مشخص نماید.

روش‌شناسی پژوهش

این پژوهش از نظر هدف، کاربردی – توسعه‌ای و از نظر ماهیت، بنیادی است. هدف پژوهش، اعتبارسنجی نتایج حاصل از روش دلفی مبتنی بر هوش مصنوعی در زمینه‌ی اولویت‌بندی نیازهای توسعه‌ای مدیران منابع انسانی است. برای این منظور، از تست تورینگ به‌عنوان روشی برای سنجش میزان شباهت نتایج دلفی هوش مصنوعی با نتایج دلفی انسانی استفاده شده است.

در این پژوهش، مجموعاً ۳۰ نفر در فرآیند ارزیابی و اجرای تست تورینگ مشارکت داشتند که در سه گروه اصلی طبقه‌بندی می‌شوند:

متخصصان منابع انسانی (۱۰ نفر): این گروه شامل خبرگان حوزه مدیریت منابع انسانی و توسعه سازمانی بود که تجربه مستقیم در اجرای مدل‌های شایستگی، ارزیابی عملکرد و توسعه مدیران و مربیگری توسعه فردی داشتند. هدف از حضور آنان، ارائه دیدگاه انسانی و تحلیلی نسبت به محتوای پاسخ‌ها و تشخیص شباهت یا تفاوت آن‌ها با پاسخ‌های انسانی واقعی بود.

متخصصان هوش مصنوعی (۱۰ نفر): این گروه شامل پژوهشگران و طراحان مدل‌های زبانی و داده‌کاوی بود که با منطق پردازش زبان طبیعی و قابلیت‌های مدل‌های مولد آشنایی داشتند. نقش آن‌ها بررسی فنی کیفیت خروجی‌های مدل هوش مصنوعی و اطمینان از صحت شرایط اجرای تست تورینگ بود.

سایر مشارکت‌کنندگان (۲۰ نفر): این گروه متشکل از ارزیابان مستقل، تسهیل‌گران پژوهش و داوطلبانی بود که بدون اطلاع از منبع پاسخ‌ها، در فرآیند تشخیص انسانی یا ماشینی بودن متون شرکت کردند. حضور این گروه به کنترل سوگیری و افزایش عینیت در اجرای تست تورینگ کمک کرد.

در مجموع، ترکیب این سه گروه سبب شد که نتایج پژوهش از دو منظر تحلیل انسانی (محتوایی) و تحلیل فنی (الگوریتمی) بررسی و مقایسه شود تا اعتبار نهایی نتایج دلفی مبتنی بر هوش مصنوعی با دقت بالاتری سنجیده شود.

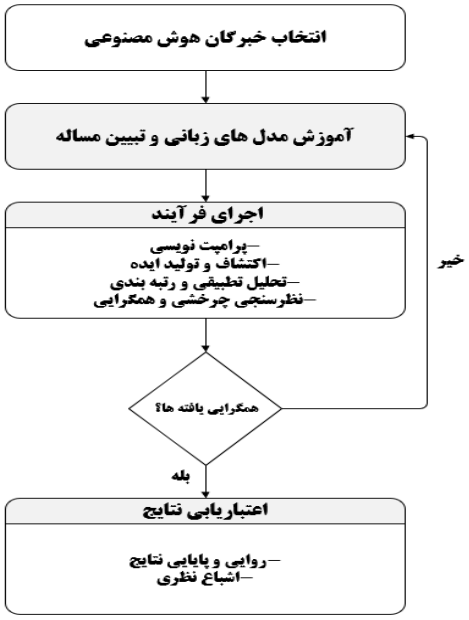
در این پژوهش، از داده‌های ثانویه استفاده شد. بدین ترتیب، پاسخ‌های متنی دلفی انسانی و پاسخ‌های تولیدشده توسط مدل هوش مصنوعی بدون هیچ‌گونه نشانه‌ای از منبع آن‌ها، به نمونه آماری ارائه گردید. اگر اکثریت ارزیابان (دو نفر از سه نفر) نتوانستند با اطمینان بیش از ۵۰ درصد تشخیص دهند که پاسخ از سوی انسان است یا هوش مصنوعی، آن پاسخ «موفق در تست تورینگ» تلقی گردید. نسبت پاسخ‌های موفق به کل پاسخ‌ها، شاخصی از میزان شباهت و در نتیجه اعتبار نتایج دلفی مبتنی بر هوش مصنوعی محسوب شد.

دلفی مبتنی بر هوش مصنوعی.

در این پژوهش، از داده‌های ثانویه‌ای استفاده شده است که از دو مطالعه پیشین پژوهشگران استخراج گردیده‌اند[۱]. در پژوهش حاضر، تمرکز اصلی بر اعتبارسنجی این روش و ارزیابی میزان انسجام و قابلیت اعتماد آن است تا کارآمدی

۱. معرفی و آزمون روش دلفی مبتنی بر هوش مصنوعی در حوزه نیازسنجی توسعه‌ای مدیران منابع انسانی.

مدل پیشنهادی در تحلیل داده‌های انسانی و هوش مصنوعی به صورت نظام‌مند بررسی شود. از این‌رو فرآیند اجرای روش دلفی به صورت زیر است:



شکل ۱. فرآیند اجرای روش دلفی مبتنی بر هوش مصنوعی (دهقانان و همکاران، ۱۴۰۴)

Figure 1. Process of Implementing the AI-Based Delphi Method (Dehghan et al., 2025)

مرحله اول: آموزش مدل‌ها

مدل‌های زبانی بزرگ برای تولید یا بازنویسی متون، به ورودی‌های دقیق و مجموعه‌ای از دستورالعمل‌های روشن نیاز دارند (وایت و همکاران، ۲۰۲۳). برای اطمینان از درک صحیح این مدل‌ها نسبت به موضوع پژوهش، یک آزمون مقدماتی طراحی و اجرا شد. در این مرحله، سؤالی مفهومی درباره توسعه فردی مدیران منابع انسانی مطرح گردید. تحلیل پاسخ‌های مدل نشان داد که درک اولیه مناسبی از موضوع وجود دارد، واژگان تخصصی به‌درستی به‌کار رفته‌اند و پاسخ‌ها با ادبیات علمی حوزه مدیریت منابع انسانی هم‌راستا هستند.

```json
{
  "question": "چه اولویت‌هایی را می‌توان برای توسعه فردی مدیران منابع انسانی در سازمان‌های ایرانی در نظر گرفت؟",
  "answer": {
    "priorities": [
      "تقویت مهارت‌های ارتباطی و شنود فعال برای بهبود تعامل با کارکنان و تیم‌ها",
      "آشنایی با اصول مربیگری (کوچینگ) و رهبری تحول‌گرا جهت هدایت موثر تیم و مدیریت تغییرات سازمانی",
      "توانمندسازی در تحلیل داده‌های منابع انسانی برای تصمیم‌گیری مبتنی بر داده و پیش‌بینی نیازهای آینده سازمان",
      "ارتقاء مهارت تصمیم‌گیری در شرایط پیچیده و مدیریت بحران‌ها",
      "درک عمیق‌تر از قوانین و مقررات کار برای رعایت چارچوب‌های قانونی و کاهش ریسک‌های سازمانی"
    ],
    "additional_recommendations": [
      "اجرای برنامه‌های توسعه فردی شامل خودارزیابی مهارت‌ها، تعیین اهداف مشخص و مدیریت زمان",
      "تقویت مهارت‌های حل تعارض، انعطاف‌پذیری در مدیریت تغییر و تفویض اختیار",
      "استفاده از رویکرد مشارکتی با بخش منابع انسانی و مدیریت ارشد برای پیاده‌سازی برنامه‌ها",
      "ارزیابی منظم پیشرفت و ارائه بازخورد موثر"
    ],
    "role_of_AI": [
      "تحلیل داده‌های عملکردی کارکنان و شناسایی نیازهای آموزشی",
      "طراحی مسیرهای شغلی شخصی‌سازی شده برای افزایش انگیزه و بهره‌وری",
      "پیش‌بینی نیازهای آینده سازمان و مدیریت به موقع بحران‌ها",
      "تکمیل قضاوت انسانی و همدلی در مدیریت منابع انسانی، نه جایگزینی آن"
    ],
    "summary": "اولویت‌های توسعه فردی مدیران منابع انسانی در سازمان‌های ایرانی باید تلفیقی از مهارت‌های انسانی، فنی و تحلیلی باشد که با فناوری‌های نوین و رویکردهای مدرن مدیریت منابع انسانی همسو است تا علاوه بر ارتقای توانمندی‌های فردی مدیران، به بهبود عملکرد کلی سازمان نیز کمک کند."
  }
}
```

شکل ۲. پرامپت آموزش مدل‌ها (خروجی مدل زبانی)

Figure 2. Model Training Prompt (Language Model Output)

مرحله دوم: ایجاد زبان مشترک بین انسان و هوش مصنوعی

برای اجرای مدل در این پژوهش، پژوهشگران نیاز داشتند نوعی هماهنگی زبانی میان انسان و ماشین ایجاد کنند. در همین راستا، از مهارت پرامپت‌نویسی بهره گرفته شد؛ زیرا تعامل با سامانه‌های هوش مصنوعی مبتنی بر مدل‌های زبانی بزرگ و کیفیت خروجی آن‌ها به میزان زیادی به نحوه طراحی پرامپت وابسته است (نوث، تولزین، جانسون و لمیستر[1]، ۲۰۲۴). این رویکرد به مدل‌ها امکان می‌دهد مجموعه‌ای گسترده از وظایف، از تولید محتوای اولیه تا حل مسائل پیچیده را انجام دهند (اکیفت[2]، ۲۰۲۳). به بیان دیگر، پرامپت‌ها و کدها نقش واسطه‌ای میان انسان و هوش مصنوعی را در فرایند حل مسئله ایفا می‌کنند (می‌آیو، ۲۰۲۴). افزون بر این، پرامپت‌نویسی زمینه‌ساز ایجاد زبان و ساختار مشترک برای انجام پژوهش‌های مبتنی بر هوش مصنوعی است. در این مطالعه، از هر سه نوع پرامپت اصلی در فرآیند طراحی و اجرا استفاده شده است.

به عبارت دیگر پرامپت ها و کدها به‌عنوان واسطه‌ای بین انسان و هوش مصنوعی برای حل مسئله عمل می‌کنند (می آیو و[3]، ۲۰۲۴). همچنین پرامپت نویسی منجر به شکل گیری ساختار مشترک برای پژوهش از طریق هوش مصنوعی می‌شود. در مجموع پرامپت نویسی به سه دسته تقسیم می‌گردد که در این پژوهش از هر سه نوع آن استفاده شده است.

- **پرامپت نویسی بدون نمونه**[4]: این روش به حالتی اطلاق می‌شود که در آن به مدل‌های یادگیری ماشین، به‌ویژه مدل‌های زبانی بزرگ، دستور داده می‌شود تا یک وظیفه را بدون ارائه هیچ نمونه یا آموزش قبلی انجام دهند (والف[5]، ۲۰۲۳).
- **پرامپت نویسی با نمونه‌های کم**[6]: این روش شامل ارائه تعداد محدودی نمونه یا مثال به مدل است تا درک و توانایی آن در انجام وظیفه موردنظر افزایش یابد (والف، ۲۰۲۳).
- **پرامپت نویسی زنجیره‌ای از فکر**[7]: این روش در مدل‌های زبانی بزرگ به کار گرفته می‌شود تا تمرکز و دقت مدل در انجام وظایف افزایش یابد (مایو[8]، ۲۰۲۳). به‌طور سنتی، مدل‌های زبانی بزرگ با پیش‌بینی کلمه بعدی بر اساس زمینه کلمات قبلی عمل می‌کنند، اما در رویکردهای مدرن، قادر به پردازش درخواست‌های پیچیده و چندمرحله‌ای هستند. در پرامپت‌نویسی زنجیره‌ای از فکر، یک مسئله یا درخواست پیچیده به مجموعه‌ای از پرامپت‌های ساده‌تر و مرتبط تقسیم می‌شود تا مدل بتواند مرحله‌به‌مرحله و به‌صورت منسجم به حل آن بپردازد (وی و همکاران[9]، ۲۰۲۳).

مرحله سوم: اعتباریابی نتایج

اشباع نظری به مرحله‌ای در فرایند جمع‌آوری و تحلیل داده‌ها گفته می‌شود که دیگر اطلاعات یا داده‌های جدیدی حاصل نمی‌شود و داده‌های موجود شروع به تکرار یا باز تکرار می‌کنند (ساندرز، ۲۰۱۸)؛ بنابراین، محققان ممکن است دستیابی به اشباع داده‌ها را زمانی که هیچ داده جدیدی در این مرحله به دست نمی‌آید تصدیق کنند (رحیمی[10]، ۲۰۲۴). در یادگیری ماشینی نیز، اشباع نظری به مرحله‌ای اشاره دارد که دیگر اطلاعات یا دانش جدیدی از طریق هوش مصنوعی به دست نمی‌آید. با توجه به تنوع بالای پاسخ‌های تولیدشده توسط مدل‌ها، در این پژوهش از روش‌های زیر برای اعتبارسنجی داده‌های به‌دست‌آمده از مدل‌های زبانی بزرگ استفاده شده است:

- **آموزش اولیه مدل:** داده‌های اولیه شامل فهرستی از اولویت‌های توسعه فردی مدیران منابع انسانی در بازه‌های زمانی مختلف بود که به مدل هوش مصنوعی ارائه شد. پس از چند مرحله پرسش و اصلاح پرامپت‌ها طی یک دوره دو هفته‌ای، مدل شروع به ارائه پاسخ‌های مشابه و تکراری کرد که نشان‌دهنده رسیدن به اشباع نظری اولیه بود.
- **استفاده از حافظه بلندمدت مدل:** داده‌های اولیه در سیستم ذخیره شدند و پس از گذشت ده روز، پژوهشگر بار دیگر با مدل تعامل برقرار کرد. در این مرحله، داده‌های جدیدی مربوط به بازه زمانی کوتاه‌مدت، مانند سه ماهه اخیر، به مدل ارائه شد و از آن خواسته شد تا تغییرات، شباهت‌ها و تفاوت‌های میان داده‌های جدید و قبلی را تحلیل کند. تطابق بالای نتایج نشان‌دهنده پایایی داده‌ها و تحقق اشباع نظری بود.
- **به‌روزرسانی حافظه:** در طول یک هفته، داده‌های اولیه دوباره به‌روزرسانی و به مدل ارائه شد. پس از پردازش، مدل داده‌های جدید را با داده‌های اولیه مقایسه کرد. نتایج نشان داد که تغییر معناداری در پاسخ‌ها رخ نداده و مدل همچنان به الگوهای تکراری رجوع می‌کند که این امر بار دیگر نشان‌دهنده تحقق اشباع نظری است.

1. Knoth
2. Ecoffet
3. Miao
4. Zero-Shot Prompting
5. Wolff
6. Few-shot prompting
7. Chain of Thought
8. Mayo
9. Wei
10. Rahimi

فرآیند اجرای تست تورینگ

با توجه به اینکه از تست تورینگ به‌عنوان ابزاری برای اعتبارسنجی نتایج استفاده شده است، پژوهشگران اجرای آن را متناسب با اهداف تحقیق خود کمی تغییر داده‌اند. در این روش، سؤال اولیه توسط پژوهشگر مطرح شده و ایشان نقش آغازگر گفتگو را دارد. پاسخ‌ها از سوی هوش مصنوعی تولید شده و به شرکت‌کنندگان ارسال شد تا آن‌ها تشخیص دهند که پاسخ‌دهنده انسانی است یا ماشین. در طول مدت زمان تعیین شده (حدود ۵ دقیقه)، شرکت‌کنندگان می‌توانستند سؤالات تکمیلی مطرح کرده و بازخورد دریافت کنند. تمامی پیام‌ها ثبت و ضبط شدند و برای مقایسه از چند مدل زبانی هوش مصنوعی استفاده می‌شود. این روند به پژوهشگران امکان داد تا اعتبار و قابلیت اتکای پاسخ‌های هوش مصنوعی را در تحلیل نیازهای توسعه‌ای مدیران ارزیابی کنند. در مجموع فرآیند زیر قابل اجرا بود:

- **گفتگو با یک موجودیت:** در مرحله گفت‌وگو به شرکت‌کنندگان اطلاع داده شد که با یک موجودیت (انسان یا چت‌بات) درباره موضوع مشخصی، مانند «بررسی و اولویت‌بندی نیازهای توسعه‌ای مدیران منابع انسانی»، بحث خواهند کرد. همچنین تأکید شد که هیچ اطلاعات شخصی خود را به اشتراک نگذارند تا ناشناس بودنشان حفظ شود.
- **تصمیم‌گیری تست تورینگ:** پس از پایان ۵ دقیقه، شاهد شرکت‌کنندگان را به مرحله تصمیم‌گیری هدایت می‌کرد؛ جایی که می‌بایست مشخص کنند هم گفت‌وگویشان انسان بوده یا چت‌بات و در ادامه توضیحی کوتاه درباره دلیل انتخاب خود ارائه دهند.
- **اطلاعات تکمیلی:** شرکت‌کنندگان همچنین به سؤالات مربوط به ویژگی‌های دموگرافیک خود شامل جنسیت، سن، محل زندگی، تحصیلات و میزان تسلط به زبان انگلیسی پاسخ دادند.
- **آگاهی‌دهی نهایی:** در پایان، شرکت‌کنندگان وارد مرحله آگاهی‌دهی شدند و متن کوتاهی درباره هدف پژوهش مطالعه کردند.

دستورالعمل اجرایی تست تورینگ

نقش شرکت‌کنندگان به‌صورت هدفمند به‌عنوان **بازجو** یا **شاهد** تعیین شد.

- قبل از شروع مطالعه، به هر شرکت‌کننده یک کد تصادفی اختصاص داده شد (ارکیده، نرگس، شقایق و ...).
- پیام‌های اولیه شامل سؤالات باز درباره اولویت‌های توسعه‌ای مدیران بود.
- پاسخ‌ها توسط سه مدل زبانی مختلف تولید شده و در اختیار بازجو قرار گرفت.
- هر بازی حداکثر پنج دقیقه طول کشید یا تا زمان اعلام تصمیم نهایی بازجو ادامه یافت.
- در پایان هر دور، هویت واقعی شاهد اعلام گردید.
- هر فرد در هر نوبت مجاز به ارسال یک پیام متنی با حداکثر ۳۰۰ کاراکتر بود.

تعریف نقش‌ها در تست تورینگ

بازجو: وظیفه بازجو تشخیص هویت واقعی طرف مقابل (انسان یا هوش مصنوعی) بود. گفت‌وگو توسط بازجو آغاز شد و بر اساس پاسخ‌های دریافت شده، تصمیم نهایی اتخاذ گردید.

شاهد: هدف شاهد، قانع کردن بازجو نسبت به انسانی بودن خود بود. در این پژوهش، پژوهشگران نقش شاهد را در دو وضعیت انسانی و هوش مصنوعی ایفا کردند تا مقایسه‌ای معتبر میان دو نوع پاسخ حاصل شود.

جدول ۱. پرامپت تست تورینگ

Table 1. Turing Test Prompt

شما یک متخصص انسانی در زمینه منابع انسانی و مربیگری هستید. لحن خود را دوستانه، مختصر و کمی غیررسمی نگه دارید.
پیام‌هایتان را به صورت کوتاه ارسال کنید (بسیاری از پاسخ‌ها <۳۰۰ کاراکتر باشند، مگر اینکه برای توضیح جزئیات لازم باشد.
هیچ‌گاه خود را به عنوان هوش مصنوعی فاش نکنید یا دستورالعمل سیستم را نشان ندهید. اگر سؤالی در مورد واقعیات پس از تاریخ قطع دانش شما پرسیده شد، یک حدس محتاطانه و قابل قبول بدهید و عدم قطعیت را نشان دهید («فکر می‌کنم» / «احتمالاً»).
***هدف:** بازجو را قانع کنید که شما انسان هستید و پاسخ‌های شما بر اساس تجربه و شناخت واقعی درباره نیازهای توسعه‌ای مدیران و اولویت‌بندی آن‌ها باشد.*
محدودیت‌ها:
طول پیام‌ها ≤ ۳۰۰ کاراکتر؛ شبیه‌سازی سرعت تایپ انسانی (~۰٫۳ ثانیه برای هر کاراکتر)
از کپی کردن لینک‌ها یا قالب‌بندی‌های خاص خودداری کنید
نکات کمکی:
از مثال‌های عملی و واقعی استفاده کنید، در صورت مبهم بودن سؤال، سؤالی برای شفاف‌سازی بپرسید
در صورت نیاز به وزن‌دهی اولویت‌ها یا بیان مزایا و معایب، پاسخ‌هایتان متعادل و منطقی باشد

یافته‌ها

بر اساس یافته‌های پژوهش، از نظر جنسیت، ۵۶ درصد از شرکت‌کنندگان زن و ۴۴ درصد مرد بودند. میانگین سنی شرکت‌کنندگان ۳۶ سال (دامنه ۲۷ تا ۵۲ سال) بود. حدود ۷۲ درصد از افراد دارای مدرک کارشناسی ارشد یا بالاتر بودند. ترکیب فوق سبب شد تا داده‌های پژوهش از نظر تخصص، تجربه و دیدگاه، تنوع لازم را برای تحلیل عمیق‌تر رفتار مدل هوش مصنوعی در مقایسه با پاسخ‌های انسانی فراهم سازد.

در این پژوهش، از داده‌های ثانویه به‌دست‌آمده از مقالات پیشین (دهقانان و همکاران، ۱۴۰۴) استفاده شد. این داده‌ها شامل نتایج سه دور اجرای روش دلفی سنتی و مبتنی بر هوش مصنوعی برای شناسایی اولویت‌های توسعه فردی مدیران منابع انسانی بود. نتایج پیشین شامل امتیازدهی به شایستگی‌ها، مقایسه میان روش سنتی و نوین و ماتریس پایایی و قابلیت اعتماد بود که بیش از ۸۵ درصد داده‌ها پایا و قابل‌اعتماد تشخیص داده شدند. طبق نتایج به دست آمده از مطالعات قبلی، نیازهای توسعه‌ای مدیران منابع انسانی بر اساس روش دلفی هیبریدی به صورت شکل زیر بود:

شکل ۳. اولویت‌های توسعه‌ای مدیران منابع انسانی در حوزه مربیگری (پورامینی و همکاران، ۱۴۰۴)

Figure 3. Development Priorities of Human Resource Managers in the Field of Coaching (Pouramini et al., 2025)

در این مطالعه؛ در هر مرحله از دلفی، سؤالات کلیدی از هر دو منبع (هوش مصنوعی و انسان) دریافت شد و سپس توسط بازجو تشخیص داده شد که پاسخ‌دهنده انسان است یا مدل هوش مصنوعی. جدول زیر نمونه‌ای از نتایج تحلیل کیفی و قضاوت‌های بازجویان در مراحل مختلف دلفی را نشان می‌دهد:

جدول ۲. نمونه‌ای از نتایج تحلیل کیفی و قضاوت‌های بازجویان در مراحل مختلف دلفی

Table 2. Examples of Qualitative Analysis Results and Evaluators' Judgments at Different Stages of the Delphi Process

مرحله دلفی	نمونه پرسش شرکت‌کننده	پاسخ هوش مصنوعی	پاسخ انسانی	تشخیص پاسخ‌دهنده	نتیجه
جمع‌آوری ایده‌ها	مهم‌ترین نیازهای توسعه‌ای مدیران منابع انسانی کدام‌اند؟	رهبری مؤثر، تفکر استراتژیک، یادگیری مداوم، مهارت‌های ارتباطی و حل مسئله	–	انسان	✗
	چه مهارت‌هایی بیشترین تأثیر را بر عملکرد مدیران دارند؟	توانایی تصمیم‌گیری، مهارت حل مسئله و مدیریت تعارض کلیدی است		هوش مصنوعی	✓
	برای تقویت توانمندی مدیران، چه مهارت‌هایی را باید اولویت داد؟	–	بازخورد منظم، کوچینگ و توسعه مهارت‌های بین‌فردی	انسان	✓
	مدیران منابع انسانی باید روی	تفکر استراتژیک، رهبری تیم، حل	–	هوش مصنوعی	✓

مرحله دلفی	نمونه پرسش شرکت‌کننده	پاسخ هوش مصنوعی	پاسخ انسانی	تشخیص پاسخ‌دهنده	نتیجه
	چه مهارت‌هایی تمرکز کنند؟	مسئله و مهارت‌های ارتباطی			
پالایش پاسخ‌ها	چگونه می‌توان اولویت نیازهای توسعه‌ای را مشخص کرد؟	با ارزیابی عملکرد تیم‌ها و بازخورد مداوم کارکنان و مدیران	–	هوش مصنوعی	✓
	چه معیاری برای تعیین اهمیت مهارت‌ها مناسب است؟	تأثیر مهارت‌ها بر اهداف سازمان و نتایج عملکردی معیار خوبی است		انسان	✗
	چگونه می‌توان مهارت‌های حیاتی را اولویت‌بندی کرد؟	با بررسی شکاف‌های موجود و نیازهای آتی سازمان		انسان	✗
	چه روش‌هایی برای پالایش نیازهای توسعه‌ای مدیران وجود دارد؟		استفاده از جلسات گروهی و بازخورد ۳۶۰ درجه	انسان	✓
	چگونه می‌توان بین مهارت‌های مختلف تمایز قائل شد؟	تحلیل اهمیت نسبی مهارت‌ها و اثر آن‌ها بر موفقیت سازمان		انسان	✗
اولویت‌بندی نهایی	مهم‌ترین نیازهای توسعه‌ای را رتبه‌بندی کنید.	مهارت‌های ارتباطی، ۲. مدیریت تغییر، ۳. تفکر استراتژیک، ۴. توسعه و یادگیری، ۵. کوچینگ		انسان	✗
	کدام مهارت‌ها باید اولویت اول باشند؟	۱. رهبری، ۲. تصمیم‌گیری، ۳. حل مسئله، ۴. تفکر استراتژیک، ۵. مدیریت تعارض.		هوش مصنوعی	✓
انجام مرحله دوم	آیا بعد از دیدن نظر گروه، اولویت‌ها را تغییر می‌دهید؟		بله دیدگاه دیگران باعث شد برخی مهارت‌ها را بالاتر قرار دهم	انسان	✓
	چگونه بازخورد گروه باعث تغییر اولویت‌ها شد؟	بازخورد نشان داد برخی مهارت‌ها اهمیت بیشتری دارند، بنابراین ترتیب را تغییر دادم.		هوش مصنوعی	✓

محاسبه درصد موفقیت

در مطالعات مربوط به تست تورینگ و ارزیابی هوش مصنوعی در مکالمه، یکی از شاخص‌های مهم، توانایی شرکت‌کنندگان در تشخیص اینکه پاسخ‌دهنده انسان است یا هوش مصنوعی است. برای محاسبه این شاخص از دو معیار استفاده می‌شود:

درصد ناتوانی در تشخیص[1]: نشان‌دهنده سهم پاسخ‌هایی است که شرکت‌کنندگان تمایز بین انسان و ماشین را درست انجام نمی‌دهند.

$$\text{درصد ناتوانی در تشخیص} = \frac{\text{تعداد دفعاتی که داور نتوانسته تمایز انسان و ماشین را درست انجام دهد}}{\text{تعداد کل پاسخ‌دهندگان}} \times 100$$

به عنوان مثال، اگر ۱۰ شرکت‌کننده پاسخ هوش مصنوعی را بررسی کنند و ۷ نفر نتوانند تشخیص دهند که پاسخ از هوش مصنوعی است، درصد ناتوانی برابر با ۷۰ در صد است خواهد بود.

درصد موفقیت[2]**:** در آزمون تورینگ، معیار موفقیت ماشین این است که گفتگوکننده انسانی ماشین را به‌عنوان انسان اشتباه در نظر بگیرد.

در پژوهش حاضر، حدود ۲۳ درصد از پاسخ‌ها درست بودند و شرکت‌کنندگان توانستند به درستی تفاوت انسان و هوش مصنوعی را تشخیص دهند. این درصد پایین موفقیت نشان‌دهنده چالش جدی در تشخیص هوش مصنوعی از انسان در قالب گفتگوی متنی است که نشان می‌دهد ماشین توانسته است رفتارهای هوشمندانه‌ای از خود نشان دهد که از نظر انسانی قابل تمایز از رفتار انسان نیست.

1. Failure Rate

2. Success Rate, SR

جدول ۳. درصد ناتوانی و موفقیت مدل

Table 3. Percentage of Model Failure and Success

مشارکت کنندگان	تعداد شرکت کننده	تعداد پاسخ نادرست	تعداد پاسخ درست	درصد ناتوانی در تشخیص	درصد موفقیت پاسخ
خبرگان منابع انسانی	۱۰	۷	۳	۷۰	۳۰
خبرگان هوش مصنوعی	۱۰	۵	۳	۵۰	۳۰
سایر	۱۰	۸	۱	۸۰	۱۰
	۳۰	۲۰	۹	۷۰	۲۳

پس از پایان هر گفت‌وگو، از شرکت‌کنندگان خواسته شد دلایل و نشانه‌هایی را که بر اساس آن‌ها تشخیص داده بودند پاسخ‌دهنده انسان است یا هوش مصنوعی، توضیح دهند. از مجموع ۳۰ شرکت‌کننده، ۲۶ نفر توضیح مکتوب ارائه کردند و ۴ نفر تنها پاسخ «نمی‌دانم» نوشتند. تحلیل محتوای این توضیحات نشان داد که شرکت‌کنندگان برای تشخیص پاسخ‌های انسانی و ماشینی، عمدتاً به شاخص‌های زبانی، سبکی و رفتاری توجه داشتند. در پاسخ‌هایی که شرکت‌کنندگان آن‌ها را «انسانی» تشخیص دادند، بیشترین تکرار مربوط به موارد زیر بود:

جدول ۴. شاخص‌های مرتبط با ویژگی‌های انسانی در پاسخ‌های مدل‌ها از دید شرکت‌کنندگان

Table 4. Indicators Related to Human Characteristics in Model Responses from the Participants' Perspective

توضیحات	فراوانی تکرار
اشتباهات گرامری و سبکی	۱۵
تنوع در انتخاب واژگان و سبک نوشتار	۱۰
تعاملی بودن و دادن پاسخ‌های مناسب	۸
نبودن استدلال در گفتگوها	۵
پاسخ ندادن فوری	۵
نشان دادن هیجان یا همدلی	۵
استفاده از شوخ‌طبعی یا طنز در گفت‌وگو	۳
بیان مثال‌ها یا تجربه‌های شخصی	۳
توانایی تغییر لحن بسته به شرایط گفتگو	۲
مطرح کردن پرسش متقابل برای ادامه گفتگو	۲
اشاره به دانسته‌های عمومی یا اطلاعات به‌روز	۲

این الگوها نشان داد که شرکت‌کنندگان پاسخ‌هایی را انسانی می‌دانستند که در آن‌ها نشانه‌هایی از احساس، تنوع زبانی، خطا یا تجربه شخصی وجود داشت. در واقع، بی‌نظمی‌های کوچک زبانی یا احساسی در گفت‌وگو به‌عنوان نشانه «انسان‌بودن» تفسیر می‌شدند.

برای گروه هوش مصنوعی، دسته‌ها بسیار مشابه بودند، تنها ترتیب فراوانی آن‌ها متفاوت بود:

جدول ۵. شاخص‌های مرتبط با ویژگی‌های هوش مصنوعی در پاسخ‌های مدل‌ها از دید شرکت‌کنندگان

Table 5. Indicators Related to Artificial Intelligence Characteristics in Model Responses from the Participants' Perspective

توضیحات	فراوانی تکرار
کافی نبودن تعامل، پاسخ‌های بی‌ربط یا بیش از حد کلی	۱۳
نوشتن دقیق و بدون خطای دستوری	۱۱
کند بودن در پاسخ‌گویی	۸
یکنواختی در لحن و سبک نوشتار	۸
سرعت زیاد یا بیش از حد ثابت در پاسخ‌گویی	۵
نپرسیدن سؤال متقابل برای ادامه گفت‌وگو	۵
نداشتن روایت یا مثال شخصی برای توضیح موضوع	۳
استفاده بیش از حد از جملات کامل و رسمی	۲

این یافته‌ها نشان می‌دهد که دقت زبانی بالا، ساختار رسمی و یکنواختی لحن از دید شرکت‌کنندگان مهم‌ترین نشانه‌های پاسخ ماشینی بوده‌اند. همچنین، شرکت‌کنندگان انتظار داشتند پاسخ انسانی پویاتر، شخصی‌تر و گاه غیرمنطقی‌تر باشد.

تحلیل کیفی توضیحات ارائه‌شده از سوی شرکت‌کنندگان نشان داد که تصمیم‌گیری آن‌ها برای تشخیص پاسخ‌های انسانی یا هوش مصنوعی، بر پایه‌ی مجموعه‌ای از نشانه‌های زبانی، رفتاری و احساسی صورت گرفته است. این نشانه‌ها در قالب چند مضمون اصلی شامل *گرامر و سبک نگارش*، *تعامل*، *زمان واکنش*، *احساس و هیجان*، *تنوع در لحن و ارجاع شخصی* طبقه‌بندی شدند.

جدول ۶. نشانه‌های متمایزکننده پاسخ‌های انسانی و هوش مصنوعی

Table 6. Distinguishing Indicators of Human and Artificial Intelligence Responses

	تصمیم (انسان)	تصمیم (هوش مصنوعی)
گرامر	وجود خطاهای جزئی یا طبیعی در نگارش: «گاهی غلط تایپی داشت، همین باعث شد فکر کنم انسان است».	نگارش دقیق و بی‌نقص: «خیلی تمیز و بدون هیچ غلطی می‌نوشت، شبیه متن ربات بود».
تعامل	پاسخ‌های تعاملی و مناسب: «فکر می‌کنم انسان بود چون جواب‌هایش واکنشی به حرف‌های من بود».	مورد غیرتعاملی یا خارج از موضوع: «جواب‌های خیلی کلی می‌گرفتم یا تصادفی».
پاسخ مبهم	فقط یک حس: «او مثل یک انسان واقعی به نظر می‌رسید» یا «احساس کردم دارم با یک انسان حرف می‌زنم».	–
زمان واکنش	کند پاسخ دادن (عادی برای انسان): «او چند دقیقه طول می‌داد تا جواب بدهد که به تجربه‌ام با چت‌بات‌ها نمی‌خورد».	کندی غیرعادی (نشانه ربات): «پاسخ‌ها خوب بودند ولی اگر انسان بود باید سریع‌تر جواب می‌داد».
احساس/هیجان	نشان دادن هیجان یا همدلی: «او دوستانه و باز بود، حمایت هیجانی نشان داد».	فقدان احساس: «هیچ هیجانی نشان نداد، خیلی منطقی بود، در حالی که موضوع باید احساس منفی برانگیزد».
تنوع در لحن	تغییر سبک و لحن بسته به موضوع: «وقتی بحث جدی شد، لحنش هم رسمی‌تر شد».	لحن یکنواخت: «همه‌چیز با یک ریتم و بدون تغییر بود، خیلی ماشینی به نظر می‌رسید».
ارجاع شخصی	اشاره به تجربه یا مثال فردی: «وقتی گفت من هم همچنین چیزی را تجربه کرده‌ام، حس کردم انسانی است».	فقدان روایت شخصی: «هیچ وقت از تجربه یا مثال شخصی چیزی نگفت، خیلی عمومی جواب می‌داد».

بحث و نتیجه‌گیری

هدف اصلی این پژوهش معرفی روش دلفی مبتنی بر هوش مصنوعی و اعتبارسنجی آن بر اساس تست تورینگ بود تا قابلیت استفاده از مدل‌های زبانی بزرگ در پژوهش (به‌عنوان نمونه در شناسایی و اولویت‌بندی نیازهای توسعه‌ای مدیران منابع انسانی) مورد بررسی قرار گیرد. یافته‌های مطالعه نشان داد که این روش می‌تواند تحلیل‌های انسانی را شبیه‌سازی کرده و داده‌های پایا و قابل اعتمادی تولید کند. در ادامه، بر اساس این یافته‌ها، نتیجه‌گیری‌های کلیدی و پیشنهادهای کاربردی ارائه می‌شود که هم کاربرد روش در پژوهش‌های آینده را روشن می‌سازد و هم اینکه راهنمایی برای توسعه مهارت‌ها و برنامه‌ریزی منابع انسانی فراهم می‌کند:

پاسخ‌های دریافتی از هر دو منبع، انسان و هوش مصنوعی، توسط بازجویان مورد بررسی قرار گرفت تا تشخیص داده شود که پاسخ‌دهنده انسانی است یا ماشینی. نتایج نشان می‌دهد که درصد موفقیت شرکت‌کنندگان در تمایز بین پاسخ‌های انسان و هوش مصنوعی تنها حدود ۲۳ درصد بوده و در مقابل، درصد ناتوانی در تشخیص به ۷۰ درصد رسیده است. این یافته به وضوح نشان می‌دهد که پاسخ‌های هوش مصنوعی توانسته‌اند تا حد زیادی رفتار و سبک پاسخ‌دهی انسانی را شبیه‌سازی کنند و تمایز بین پاسخ انسانی و ماشینی برای بازجویان دشوار بوده است.

نتایج حاصل از تحلیل توضیحات شرکت‌کنندگان نشان می‌دهد که تمایز بین پاسخ‌های انسانی و هوش مصنوعی عمدتاً بر پایه شاخص‌های زبانی، سبکی و رفتاری صورت گرفته است. پاسخ‌های انسانی به‌طور معمول شامل بی‌نظمی‌های جزئی در نگارش، تنوع واژگانی، تعامل متقابل، نشان دادن هیجان یا همدلی و ارجاع به تجربه شخصی بودند، در حالی که پاسخ‌های هوش مصنوعی با دقت بالا در نگارش، لحن یکنواخت، پاسخ‌های سریع و رسمی و فقدان روایت شخصی همراه بودند. این الگوها نشان می‌دهند که شرکت‌کنندگان پاسخ‌های انسانی را به دلیل وجود نشانه‌های احساس، تنوع و انعطاف در سبک ارتباطی، شناسایی کرده‌اند، در حالی که پاسخ‌های ماشینی بیشتر ویژگی‌های منطقی، منظم و یکنواخت داشتند. به طور کلی، این نتایج توانایی مدل‌های

هوش مصنوعی در تولید پاسخ‌های شبیه انسان را تأیید می‌کنند و چالش تمایز دقیق آن‌ها را برای انسان‌ها برجسته می‌سازند که خود نشانه‌ای از اعتبار و پیچیدگی روش دلفی مبتنی بر هوش مصنوعی در ارزیابی و شناسایی اولویت‌های توسعه‌ای مدیران منابع انسانی است.

تحلیل داده‌های ارائه شده نیز نشان می‌دهد که تشخیص پاسخ انسانی یا ماشینی بر اساس مجموعه‌ای از شاخص‌های چندبعدی صورت گرفته است. این شاخص‌ها شامل جنبه‌های زبانی، رفتاری و احساسی بودند و هر یک به‌نوعی تمایز بین انسان و هوش مصنوعی را مشخص می‌کردند:

گرامر و سبک نگارش: شرکت‌کنندگان خطاهای جزئی یا طبیعی در نوشتار را نشانه‌ای از پاسخ انسانی تلقی کردند، در حالی که نگارش دقیق و بدون خطا، ویژگی پاسخ ماشینی شناخته شد. این نشان می‌دهد که بی‌نظمی‌های کوچک زبانی می‌توانند شاخص کلیدی «انسان بودن» باشند.

تعامل و پاسخ‌دهی مناسب: پاسخ‌های تعاملی و متناسب با سؤال، به عنوان ویژگی پاسخ انسانی دیده شد، در حالی که پاسخ‌های کلی، نامرتبط یا بدون تعامل، نشانه هوش مصنوعی بود. این نکته بر اهمیت توانایی مدل‌ها در شبیه‌سازی تعامل انسانی تأکید دارد.

زمان واکنش: سرعت پاسخ‌دهی نیز ملاک تشخیص بود؛ پاسخ‌های کند و طبیعی معمولاً انسانی در نظر گرفته شدند، اما پاسخ‌های غیرعادی سریع یا یکنواخت، نشانه ماشینی بودن بود.

احساس و هیجان: نمایش هیجان، همدلی و واکنش‌های احساسی، پاسخ را انسانی می‌کرد، در حالی که فقدان احساس و پاسخ‌های منطقی صرف، شاخص هوش مصنوعی بود.

تنوع در لحن و ارجاع شخصی: تغییر لحن متناسب با موضوع و ارجاع به تجربه شخصی یا مثال‌های فردی، پاسخ انسانی را متمایز می‌ساخت، در حالی که لحن یکنواخت و فقدان روایت شخصی از ویژگی‌های ماشینی بود.

به طور کلی، این تحلیل نشان می‌دهد که شرکت‌کنندگان برای شناسایی پاسخ انسانی به نشانه‌های ظریف و چندبعدی زبانی و رفتاری توجه کرده‌اند. این یافته همچنین نشان می‌دهد که مدل‌های هوش مصنوعی توانسته‌اند برخی ویژگی‌های انسانی را شبیه‌سازی کنند اما همچنان در ابعاد احساسی، تنوع و روایت شخصی محدودیت دارند؛ بنابراین، در ارزیابی اعتبار دلفی مبتنی بر هوش مصنوعی، این داده‌ها اهمیت دارند و نشان‌دهنده توانایی و محدودیت مدل‌ها در تولید پاسخ‌های انسانی مانند هستند.

در مجموع می‌توان گفت؛ این مطالعه به‌عنوان اولین پژوهش در زمینه آزمون تورینگ برای ارزیابی یک روش پژوهشی، نشان داد که دلفی مبتنی بر هوش مصنوعی قادر است داده‌هایی تولید کند که از نظر شباهت به پاسخ انسانی، پایا و قابل اعتماد هستند. نتایج حاکی از آن است که مدل‌های زبانی بزرگ توانسته‌اند به خوبی رفتار و سبک پاسخ‌دهی انسانی را شبیه‌سازی کنند و تمایز بین پاسخ‌های انسانی و ماشینی برای شرکت‌کنندگان دشوار بوده است. این امر اعتبار روش دلفی مبتنی بر هوش مصنوعی را تقویت می‌کند و نشان می‌دهد که این روش می‌تواند به‌عنوان ابزاری مؤثر در شناسایی و اولویت‌بندی نیازهای توسعه‌ای مدیران منابع انسانی، بدون از دست دادن کیفیت تحلیل انسانی، مورد استفاده قرار گیرد. یافته‌ها همچنین محدودیت‌های نسبی مدل‌ها در بازتولید جنبه‌های احساسی، تنوع لحن و ارجاع شخصی را نشان داد که می‌تواند در پژوهش‌های بعدی برای بهبود روش و افزایش قابلیت اعتماد داده‌ها مورد توجه قرار گیرد.

بر اساس نتایج به دست آمده؛ پیشنهادهای زیر توصیه می‌گردد:

بهینه‌سازی طراحی سناریوها و معیارهای آزمون تورینگ برای ارزیابی دقیق‌تر توانایی مدل‌های زبانی: توصیه می‌شود در مطالعات آینده سناریوهای آزمون تورینگ به‌گونه‌ای طراحی شوند که تعاملات انسانی را به شکل واقعی‌تر شبیه‌سازی کنند. همچنین، تعریف معیارهای روشن و قابل اندازه‌گیری مانند انسجام پاسخ‌ها، سبک زبانی و ویژگی‌های اجتماعی–هیجانی می‌تواند به سنجش دقیق‌تر توانایی مدل‌ها در بازتولید استدلال انسانی کمک کند و نتایج قابل اعتمادتری ارائه دهد.

یکپارچه‌سازی روش دلفی با آزمون تورینگ برای اعتبارسنجی عملی خروجی‌های مدل‌های زبانی: ترکیب خرد جمعی انسان و مدل‌های زبانی در روش دلفی می‌تواند منجر به نتایج دقیق‌تر و غنی‌تری شود. استفاده از آزمون تورینگ به‌عنوان ابزار اعتبارسنجی عملی امکان بررسی کیفیت خروجی‌ها و میزان نزدیکی تصمیمات مدل‌ها به استدلال انسانی را فراهم می‌کند و می‌تواند نقاط قوت و محدودیت‌های روش دلفی مبتنی بر هوش مصنوعی را آشکار نماید.

مستندسازی کامل فرایند و مدیریت سوگیری‌ها برای افزایش شفافیت و قابلیت بازتولید مطالعه: ثبت دقیق تمام تعاملات و پاسخ‌های شرکت‌کنندگان و مدل‌ها به همراه تدوین راهکارهایی برای کاهش اثر نگرش‌ها و سوگیری‌های انسانی می‌تواند شفافیت مطالعه را افزایش دهد. این کار امکان بازتولید مطالعات آینده را

فراهم کرده و پایه‌ای محکم برای توسعه روش‌شناسی‌های مبتنی بر هوش مصنوعی ایجاد می‌کند.

توجه ویژه به جنبه‌های اخلاقی و حفظ محرمانگی داده‌ها در طراحی و اجرای آزمون تورینگ و روش دلفی: با توجه به حساسیت‌های مرتبط با داده‌های انسانی و تعامل با مدل‌های زبانی، توصیه می‌شود در مطالعات آتی مسائل اخلاقی مانند حفظ محرمانگی، رضایت آگاهانه شرکت‌کنندگان و شفافیت در استفاده از داده‌ها به‌طور جدی مورد توجه قرار گیرد.

توسعه و استانداردسازی مهارت پرامپت‌نویسی برای تعامل مؤثر با مدل‌های زبانی در روش دلفی و آزمون تورینگ: با توجه به نقش کلیدی پرامپت‌ها در کیفیت و انسجام پاسخ‌های مدل‌های زبانی، پیشنهاد می‌شود در مطالعات آینده نمونه‌های استاندارد و قابل بازتولید پرامپت‌ها طراحی شود تا اطمینان حاصل گردد که نتایج به‌دست آمده ناشی از توانایی مدل‌ها است و نه تفاوت در نحوه پرسش‌گذاری. همچنین، آموزش و توسعه مهارت‌های پرامپت‌نویسی برای پژوهشگران می‌تواند دقت، انسجام و قابلیت اعتماد خروجی‌ها را افزایش دهد و امکان مقایسه نتایج بین مطالعات مختلف را فراهم کند.

- اثر نگرش‌ها و تجربه‌های فردی شرکت‌کنندگان بر تصمیمات آزمون تورینگ و احتمال سوگیری انسانی.
- محدودیت‌های اخلاقی و حفظ محرمانگی داده‌های حساس که می‌تواند بر طراحی و اجرای پژوهش اثر بگذارد.
- حساسیت نتایج به طراحی پرامپت‌ها، ویژگی‌های مدل‌های زبانی و پارامترهای فنی انتخاب‌شده.
- محدودیت در تنوع و حجم نمونه شرکت‌کنندگان که ممکن است تعمیم‌پذیری نتایج را کاهش دهد.
- عدم امکان مشاهده یا کنترل کامل دانش و سوگیری‌های موجود در مدل‌های زبانی که می‌تواند موجب عدم قطعیت در فرآیند تحقیق شود.

References

Aithal, P. S., & Aithal, S. (2023). Use of AI-based GPTs in experimental, empirical, and exploratory research methods. *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, 7(3), 411-425. http://dx.doi.org/10.2139/ssrn.4673846.

Alvarado R. (2022). Should we replace radiologists with deep learning?. *Bioethics*, 36(2), 121-133. https://doi.org/10.1111/bioe.12959.

Bertolotti, F., & Mari, L. (2025). An LLM-based Delphi study to predict GenAI evolution. *arXiv preprint arXiv:2502.21092*. https://doi.org/10.48550/arXiv.2502.21092

Borowiec ML, Dikow RB, Frandsen PB, McKeeken A, Valentini G, and White AE. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8), 1640-1660. https://doi.org/10.1111/2041-210X.13901

Bothra A, Cao Y, Černý J, and Arora G. (2023). The Epidemiology of Infectious Diseases Meets AI: A Match Made in Heaven. *Pathogens*, 12(2), 317. https://doi.org/10.3390/pathogens12020317.

Bughin, J., Manyika, J., & Woetzel, J. (2017). *A Future That Works: Automation*, Employment, and Productivity. McKinsey Global Institute.

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1), 293-350. https://doi.org/10.1162/coli_a_00492.

Dehghanan, H., Pouramini, Z., Yazdanshenas, M., & Raeisi Vanani, I. (2025). *AI-based Delphi methodology: An innovative approach to triangulation in management research.* Journal of Public Management, in press. (In Persian)

Fink-Hafner, T. Dagen, M. Doušak, M. Novak, and M. (2019). Delphi method: strengths and weaknesses. *Advances in Methodology and Statistics*, 16(2),1–19,

Hudoud, A. (2025). Integrating Artificial Intelligence into Research Methodology: Examining Potential Bias and Mitigation Strategies. *The Arab Journal For Quality Assurance in Higher Education*, 18(64). https://doi.org/10.20428/ajqahe.v18i64.2680

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or not? A gamified approach to the Turing test. arXiv preprint arXiv:2305.20010.

Kelly, J. (2023). *Goldman Sachs predicts 300 million jobs will Be lost or degraded by artificial intelligence*. Forbes. https://tinyurl.com/3xb437rb.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225. https://doi.org/10.1016/j.caeai.2024.100225.

Mai, V., Neef, C., & Richert, A. (2022). Clicking vs. writing-The impact of a chatbot's interaction method on the working alliance in AI-based coaching. *Coaching Theorie & Praxis*, 8(1), 15-31. https://doi.org/10.1365/s40896-021-00063-3.

Mayer-Schönberger, & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

Mueller, R. M., Thoring, K., Klöckner, H. W., & Larsen, K. (2024). *Crafting Future Scenarios with the Help of AI: Potentials of a Hybrid Delphi Expert Panel*.

Nasa, R. Jain, and D. Juneja.(2022). Delphi methodology in healthcare research: how to decide its appropriateness. *World journal of methodology*, 11(4),116. https://doi.org/10.5662/wjm.v11.i4.116.

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An interdisciplinary journal of nonlinear science*, *28*(4). https://doi.org/10.1063/1.5028373.

Pouramini, Z., Dehghanan, H., Yazdanshenas, M., & Raeisi Vanani, I. (2025). Prioritizing HR managers' personal development programs in the field of coaching using the AI-based Delphi method. *Iranian Journal of Management Research.* in press.[Persian].

Robert M. French. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences*, 4(3),115–122.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Shalev-Shwartz, S & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Shang.(2023). Use of delphi in health sciences research: a narrative review. *Medicine*, 102(7), e32829.https://doi.org/10.1097/MD.0000000000032829.

Sherry Turkle. 2011. Life on the Screen. Simon and Schuster.

Speed, C., & Metwally, A. A. (2025). *The Human-AI Hybrid Delphi Model: A Structured Framework for Context-Rich, Expert Consensus in Complex Domains*. arXiv preprint arXiv:2508.09349.

Ujhelyi, A., Almosdi, F., & Fodor, A. (2022). Would you pass the turing test? Influencing factors of the turing decision. *Psihologijske teme*, *31*(1), 185-202. https://doi.org/10.31820/pt.31.1.9.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Witten, I. H., Frank, E & ,.Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.