DOI: https://doi.org/10.34123/jurnalasks.v17i1.792

POLITEKNIK STATISTIKA STIS

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika &
Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Early Study of LLM Implementation in Survey Interviews

Lailatul Hasanah [1*], Budi Yuniarto[2]

[1,2]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address:* 222011364@stis.ac.id

ARTICLE INFO

Article history:
Received 30 August, 2024
Revised 12 December, 2024
Accepted 1 February, 2025
Published 24 February, 2025

Keywords:
ChatGPT; Interview; LLM; Official Statistics; Survey

Abstract

Introduction/Main Objectives: This research aims to conduct a preliminary study into the use of LLMs for extracting information to fill out questionnaires in survey interviews. **Background Problems:** BPS-Statistics Indonesia used paper-based questionnaires for interviews and is recently utilizing the Computer Assisted Personal Interviewing (CAPI) method. However, the CAPI method has some drawbacks. Enumerators must input data into the device, which can be burdensome and prone to errors. **Novelty:** This study uses a large language model (LLM) to extract information from survey interviews. **Research Methods:** This study utilizes a text-to-speech application to translate interview results into text. Translation accuracy is measured by the Word Error Rate (WER). Then the text was extracted using the ChatGPT 3.5 Turbo model. GPT-3.5 Turbo is part of the GPT family of algorithms developed by OpenAI. **Finding/Results:** The extraction results are formatted into a JSON file, which is intended to be used for automatic filling into the database and then evaluated using precision, recall, and F1-score. Based on research conducted by utilizing the Speech Recognition API by Google and the ChatGPT 3.5 Turbo model, an average WER of 10% was obtained in speech recognition and an average accuracy of 76.16% in automatic data extraction.

1. Introduction

Large Language Model (LLM) is a relatively new technology, which utilizes artificial intelligence. LLMs are a class of Artificial Intelligence (AI) that can understand, interpret, and generate texts. LLMs are capable of understanding and generating texts at a level indistinguishable from humans [1]. Large Language Models (LLMs) are deep learning models trained on vast amounts of data, enabling them to understand and generate natural language [2]. Recent studies have demonstrated that LLMs have achieved significant success in various natural language tasks, including automatic summarization (creating a condensed version of a text), machine translation (automatically translating text from one language to another), and question answering (developing automated systems that respond to questions based on a given text) [3].

Statistical organizations can use the advantages of an LLM. They could be helpful in automating several duties within a statistical organization because of their exceptional comprehension of textual data, ability to summarize vast amounts of information, and ability to provide responses that like those of a human [1]. Badan Pusat Statistik (BPS-Statistics Indonesia) is the national statistical office of Indonesia. BPS plays a crucial role in providing accurate and reliable statistical information for both the public and the government [4]. To fulfill this role, BPS collects data through censuses and surveys, employing interviewers to gather information from respondents. Initially, BPS used paper-based questionnaires for interviews. However, with technological advancements, BPS began utilizing the

Computer Assisted Personal Interviewing (CAPI) method [5]. Interviews are now conducted using Android-based devices, allowing respondents' answers to be directly stored in a database and automatically uploaded to a central server.

The implementation of CAPI has proven to have numerous advantages. Using Android devices for CAPI minimizes the costs associated with printing paper questionnaires. Additionally, this system enables the direct transmission of respondents' answers to the database, expediting the data collection process. The validation features in the CAPI application also help prevent entry errors by enumerators during interviews [6]. However, the CAPI method has some drawbacks. Enumerators must input data into the device, which can be burdensome and prone to errors [7]. Concurrent interviewing, where the same questions are asked to different household members, can increase the workload on enumerators, leading to physical fatigue [6]. Furthermore, enumerators must multitask—conducting interviews, listening to respondents, and entering data—which can result in typing errors and incomplete documentation of respondents' information, as enumerators often only type key points [8].

With the advancement of technology, human-computer interaction has evolved from keyboard input to voice input [9]. Collecting data through voice input for survey interviews is a promising method compared to using paper questionnaires for interviews. For responses, the enumerator only needs to press the record button to capture the answer [10]. Collecting voice data also facilitates in-depth interviews, resulting in richer and more comprehensive information [7].

The recorded results must be converted into text. Therefore, using an automatic speech recognition application, the recordings are transcribed into text. However, the output of the automatic speech recognition process is still in the form of unstructured text. To fill out the interview questionnaire, an information extraction process is required. This process involves several natural language processing methods, such as categorizing information by identifying parts of the text that contain words matching the fields in the questionnaire.

Based on the research conducted by [12], speech recognition, commonly referred to as Speech-to-Text, can be achieved through various processing techniques. Speech-to-Text application by dividing the voice processing process into two stages: feature extraction and feature matching. She employed the Mel Frequency Cepstral Coefficients (MFCC) method for feature extraction, while the Dynamic Time Warping (DTW) method was used for feature matching. After conducting experiments on 217 data samples, she achieved an accuracy rate of 95.85%. Although Dinata et al.'s research demonstrated high accuracy, their experiments were limited to words and sentences containing only five words. This limitation makes their approach unsuitable for voice conversion in interview processes.

Indonesian Speech-to-Text on self-recorded voices and voices from YouTube, with durations ranging from a minimum of 17 seconds to a maximum of 2 minutes [13]. The Speech-to-Text process in this study involved several stages, including feature extraction, voice detection, and language detection. At the feature extraction stage, noise correction is performed to isolate the desired sound from background noise using the FastICA algorithm. For the voice and language detection stages, this study employs the speech recognition library module in Python, utilizing the Google Speech Recognition API. Experiments combining the FastICA and Google Speech Recognition algorithms achieved an accuracy of 94.75%. Buana's research demonstrates that the use of the speech recognition library module for extensive word sets yields high accuracy.

The effectiveness of the speech recognition library module is further supported by research conducted by Adnan et al. [14]. This study compares the accuracy of the Speech-to-Text process for audio recorded in real time versus audio files. The experiments revealed that the accuracy of the Speech-to-Text process for real-time audio recordings was slightly lower than for audio files, with respective accuracies of 93% and 97%. To obtain category information, the LLM can be utilized. Recent studies have demonstrated that LLMs have achieved significant success in various natural language tasks, including automatic summarization (producing a condensed version of a text), machine translation (automatically translating text from one language to another), and question answering (developing automated systems that respond to questions based on a given text) [11].

Gartlehner et al. [15] conducted a study on data extraction, experimenting with extracting data elements from published research. The experiment involved 10 English-language controlled trial publications, each containing 16 data elements. Utilizing the Large Language Model Claude 2, the study achieved an overall accuracy of 96.3%, with high test-retest reliability (replication 1: 96.9%; replication 2: 95.0%). In contrast to the controlled trial sample used by Gartlehner et al., Zou et al. [16] conducted research on data extraction from ESG reports published by companies in 12 industries listed on the Hong Kong Stock Exchange in 2022. This study utilized Large Language Models (LLM), specifically ChatGPT-4, combined with the Retrieval Augmented Generation (RAG) technique. The trial, which included a sample of 166 companies, achieved an accuracy rate of 76.9% in extracting structured data.

The currently popular Large Language Models include: BERT developed by Google, T5 developed by Google, and GPT developed by OpenAI. Although they all implement Transformer, these three models have different architectures. Therefore, this study aims to conduct a preliminary study into the use of LLMs for extracting information to fill out questionnaires in survey interviews.

2. Material and Methods

2.1. Collection and Preparation

As an early study on the application of LLM in survey interviews, this research focuses on the voice conversion system from interview recordings, using sample questions from the 2020 Population Census. The interviews conducted in this study used the Indonesian language. All respondents are students of Politeknik Statistika STIS. Here is the profile of the respondents of this study.

1. Respondent 1: female, a student from Kudus, Central Java which has 4 family members.
2. Respondent 2: female, a student from Palembang, South Sumatra which has 4 family members.
3. Respondent 3: female, a student from Banyuwangi, East Java which has 3 family members.
4. Respondent 4: female, a student from Makassar, South Sulawesi which has 2 family members.
5. Respondent 5: male, a student from Bantul, DI Yogyakarta which has 4 family members.
6. Respondent 6: male, a student from Agam, West Sumatra which has 3 family members.

The questions used in this interview can be seen in Table 1. In one-way interviews, respondents directly speak to answer written questions, while in two-way interviews, the enumerator asks each question, and the respondent provides the answers. The questions asked relate to 11 fields whose information will be extracted as shown in table 1.

Table 1. List of questions

Fields **(English / *Indonesian*)**	**Type**
Province (*Provinsi*)	Text
Regency/City (*Kabupaten/kota*)	Text
Religion (*Agama*)	Text
Head of household name (*Nama kepala rumah tangga*)	Text
Email (*Email*)	Text
Name of respondent (*Nama responden*)	Text
Cellphone number (*Nomor HP*)	Number
Number of household member (*Jumlah anggota rumah tangga*)	Number
ID (*NIK*)	Number
Address (*Alamat*)	Text
Business classification (*Lapangan usaha*)	Text

Six respondents were interviewed across several different scenarios. In addition, the results of interviews conducted in one direction during the day will be given additional noise using Gaussian noise from the Python library. This repetition with different scenarios on the same respondents aims to see the consistency of speech recognition.

The scenarios are as follows:

1. One-way interviews in the morning using external recording.
2. One-way interviews in the afternoon using external recording.
3. One-way interviews in the evening using external recording.
4. One-way interviews with respondents adding regional accents using external recording.
5. One-way interviews with randomized questions using external recording.
6. One-way interviews using a recorder from web-based application.
7. Two-way interviews using external recording.
8. Two-way interviews with randomized questions using external recording.
9. One-way interviews during the day with the addition of Gaussian noise from the python library.

For the purposes of case number 6, a simple web-based application will also be developed using python, javascript, and HTML programming.

2.2. Audio Translation with Speech Recognition

The collected recordings from scenarios 1-5 and 7-9 above are saved in MP3 format. Then the recorded audio format is converted from MP3 to WAV. This is necessary because the Automatic Speech Recognition process requires the library to accept voice input only in WAV and FLAC formats. The conversion of audio format from MP3 to WAV is performed using the AudioFileClip module in the *moviepy.editor* library.

The AudioFileClip module can convert all types of audio supported by ffmpeg. The advantage of using the moviepy library for converting sound formats is that it does not require a separate *ffmpeg* installation. Things to consider when converting sound with *moviepy* are the metadata of the audio. Audio metadata contains, among others: duration, format, bitrate, audio channel, and audio frequency. The recorder provided by the cellphone usually already has complete metadata. However, recordings in web-based applications still use the default recorder from the browser application, such as Mozilla Firefox, Google Chrome, and so on. Recording with this browser application produces audio with incomplete metadata. Therefore, metadata is refined using the *ffmpeg* library.

The translation process using the SpeechRecognition library begins with inputting audio by entering the file directory. The audio file will be recorded or read using the recognizer available in the SpeechRecognition library function. The recorded file is detected using Google Speech Recognition with the Indonesian language filter. After the detection process is complete, the output will be in the form of a text string which is stored in the input_text variable. The audio that is already in WAV format is then used as input in the speech recognition process. Several popular libraries can be used for Automatic Speech Recognition, including SpeechRecognition, PyAudio, and Librosa [17]. This study will use the SpeechRecognition library which utilizes the Google Speech Recognition API and translates into text.

2.3. Evaluation of Automatic Speech Recognition

In the next step, translation accuracy will be measured by the *Word Error Rate* (WER). WER is the percentage of words, which are to be inserted, deleted or replaced in the translation in order to obtain the sentence of reference [18]. WER is the most popular metric for Automatic Speech Recognition evaluation, which measures the percentage of word errors (Substitution (S), Insertion (I), Deletion (D)) against the number of words processed (N). WER can be written as the following equation 1.

$$WER = \frac{(S + D + I)}{N} \times 100\% \tag{1}$$

2.4. Model Fine Tuning

In the next process, the text is used as input into the fine-tuned Chat-GPT3.5 Turbo model. This study chooses the OpenAI and ChatGPT-3.5 Turbo tools as models in this study. This is because this model does not require a computer with high computer capabilities and easy free trial access. GPT3.5 Turbo is part of the GPT family of algorithms developed by OpenAI [19]. GPT is an autoregressive model that uses an attention mechanism to predict the next token in a sequence based on previous tokens. This process is conducted to obtain data extraction results from the entered text input.

The process of fine-tuning involves several stages, including:

1. Preparing and Uploading Training Data:
   Training data must be stored in a JSONL file with the following format.

   {"messages": [{"role": "system", "content": " *The contents of the task/prompt/command you want to run* "}, {"role": "user", "content": " Contents of Speech Recognition text results "}, {"role": "assistant", "content": " Desired output result "}]}

   It is important to note that a minimum of 10 training data samples is required for fine-tuning.
2. Uploading Training Data:

Once the training data has been created, upload it to the OpenAI API web by selecting the model used, in this case GPT-3.5 Turbo-0125.

3. Confirmation:

   If the upload is successful, the name of the fine-tuned model used in this study will be displayed.

2.5. Data Extraction Using Fine-Tuned Model

The next step is to extract information from translated text using the fine-tuned model above. To use the Chat-GPT3.5 Turbo Fine Tuned model in Python, the openai library is required. OpenAI provides an API that gives users worldwide access to the LLM model so that developers can build interactions with OpenAI applications. In performing information extraction using ChatGPT, this study uses Indonesian language prompt tuning with several commands as follows.

- *Tugas dari model adalah untuk melakukan pengekstrak data dari variabel input_text dan data yang diperlukan, antara lain: Provinsi, Kabupaten/Kota, Nama Kepala Keluarga, Email, Jumlah Anggota Keluarga, Nomor Handphone, dan Alamat Rumah*
- *Untuk data Nama, NIK, Agama, dan Deskripsi Pekerjaan harus diekstrak secara berulang sesuai dengan jumlah anggota keluarga.*
- *Pengekstrakan Provinsi harus sesuai dengan penulisan Provinsi, Kabupaten/Kota, dan Agama yang ada di Badan Pusat Statistik dan apabila provinsi tidak disebutkan dalam audio rekaman, maka bisa didekati dengan melihat Kabupaten/Kota yang disebutkan dalam audio rekaman.*
- *Pengekstrakan jumlah Anggota Keluarga dan NIK memiliki data berupa numerik.*
- *Pengekstrakan Alamat Rumah harus serinci mungkin.*
- *Pengekstrakan NIK harus memiliki 16 digit angka.*
- *Pengekstrakan Deskripsi Pekerjaan harus sedetail mungkin dan berisi kegiatan yang dilakukan, bahan yang digunakan, output yang dihasilkan, dan tempat bekerja.*

2.6. Formatting into JSON Form

The extraction results are formatted into a JSON file. Formatting into JSON format is intended so that it can be used for automatic filling into the database. Subsequently, the precision, recall, and accuracy of the model in recognizing answer entities according to the questions are calculated. Additionally, a matching process is performed to assess the alignment of the extracted text with real answer.

2.7. Evaluation of Extraction Results

Evaluation of information extraction results using fine-tuned models for each question in questionnaire is done by calculating precision, recall and F1-score of each respondent's answer. Precision, recall and F1-score are calculated using the following formula [20].

$$Precision: P = \frac{TP}{TP + FP} \quad (2)$$

$$Recall: R = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2\,\frac{P * R}{P + R} \quad (4)$$

Meanwhile, to evaluate the accuracy of the answers recognized by ChatGPT compared to the actual answers, the accuracy of the answers will be evaluated.

$$Accuracy \ = \frac{correct\ answer\ recognized}{number\ of\ questions} \times 100\% \quad (5)$$

3. Result and Discussion

Based on the results of Automatic Speech Recognition (ASR), it is evident that less common words often result in translation errors, such as “NIK” being interpreted as “nikah.” Additionally, most symbols are translated phonetically, for example, “@” being rendered as “ath.” Translation errors are also prevalent in the pronunciation of names and email addresses, which often differ from the intended spelling.

This issue arises because names and emails are created by individuals, allowing for variations in spelling despite identical pronunciation. Consequently, there is no standardized spelling for names and email addresses in ASR. Examples of incorrect spellings from the Automatic Speech Recognition results include Indriani instead of Indriyani, Sumiati instead of Sumiyati, and Zainab instead of Zaenab, among others. Repeated mention of digits also often experiences translation errors. This is because the same number but mentioned repeatedly will be read as a single digit number. For example; 0001 becomes 001. Two examples of the speech recognition results for the fourth respondent during the afternoon interview can be seen in Table 2.

Table 2. Two examples of speech recognition results

Respondent, case	**Result of *Speech Recognition***
Respondent 1, One-way interviews in the afternoon using external recording	**Saya tinggal di kabupaten Kudus provinsi Jawa Tengah nama kepala keluarga Ibu Jariyah alamat rumah di desa damaran nomor 26 RT 3 RW 1 Kecamatan Kota kabupaten Kudus nomor handphone 0815 7561 1774 untuk jumlah anggota rumah tangga ada 4 email 222011275 ~~etis~~ @stis.ac.id Kemudian untuk kepala anggota rumah tangga untuk anggota Rumah Tangga pertama itu Edwin Brian Rahmanto Nik 3319020 905080003 Kemudian untuk agama Islam dan dia masih bersekolah di SMK Muhammadiyah 1 Kudus jurusan teknik sepeda motor Kemudian untuk nama ibu ~~Sumiati~~ Sumiyati ~~nikah~~ nik 331 9025 ~~10445001~~ 104450001 kemudian agama Islam dan merupakan pensiunan di Rumah Sakit Islam Kudus Rumah Sakit Islam Kudus Kemudian untuk nama Ibu Jariyah agama Islam merupakan guru di Sekolah Dasar Muhammadiyah 1 Kudus Nik 3319024 60672 ~~001~~ 0001 Kemudian untuk nama untuk agama Islam kemudian untuk nama Karina Cindy Rahmanto merupakan mahasiswa di ~~Polsek~~ Polstat Stis beragama Islam nikah 331 902540302 0002**
Respondent 2, One-way interviews in the afternoon using external recording	**Provinsi Sumatera Selatan Kota Palembang nama kepala keluarga Bapak Amrin Joni email Niken Yuliana 853 @ gmail.com nomor HP 0895 0310 3779 jumlah anggota rumah tangga 4 alamat rumah jalan bungaran Nomor 61 RT 09 kota Palembang nama anggota rumah tangga pertama Bapak Amrin Joni agama Islam Nik 1671 09151062 ~~005~~ 0005 pekerjaan membuka usaha di depan rumah anggota rumah tangga kedua nama ibu ~~Zainab~~ Zaenab agama Islam Nik 1671 0941 0165 0008 lapang Pekerjaan ibu rumah tangga anggota rumah tangga ketiga nama Elisa Anggraini agama Islam ~~nikah~~ nik 1671 0943 0299 0007 rumah tangga keempat Niken Yuliana nama nama Niken Yuliana agama Islam ~~nikah~~ nik 1671 ~~09.58~~ 0958 0701 ~~005~~ 0005 pekerjaan mahasiswa di politeknik statistika Stis pekerjaan anggota rumah tangga ketiga sebagai mahasiswa di Universitas Negeri Malang**

Notes: Words in green, red, and blue indicate correct, incorrect, and corrective words, respectively.

Based on the Word Error Rate (WER) values provided, the one-way recording condition yields a smaller WER value than the two-way recording. When considering the recording method, audio files from external recorders exhibit a smaller WER value compared to recordings obtained directly from the web-based application. This difference can be attributed to the superior sound quality of external recorder files. Additionally, the presence of noise significantly impacts the results of Automatic Speech Recognition (ASR) translation. Specifically, audio files with noise tend to have a higher WER value than those without noise.

Table 3. Word error rate of speech reconition result

Case	Respondent					
	1	2	3	4	5	6
1	6%	5%	9%	6%	-	7%
2	7%	5%	7%	5%	11%	11%
3	11%	7%	11%	6%	-	5%
4	5%	7%	26%	6%	11%	8%
5	7%	9%	7%	6%	8%	11%
6	14%	13%	9%	8%	7%	7%
7	12%	12%	18%	11%	13%	11%
8	10%	7%	15%	7%	-	15%
9	10%	7%	11%	6%	18%	15%
Average : 10%						

Overall, the results of Speech Recognition in translating voice into text produce good values as can be seen in Table 3. This can be seen from the average Word Error Rate of 10%. This means that Speech Recognition has succeeded in correctly translating 90% of words. The next step is to extract information from the text that has been translated with ASR. The extracted information corresponds to the questions in the questionnaire. An example of the information extraction results on one of the respondents can be seen in Table 4.

Table 4. Example of information extraction result

Respondent, cases	**Results of information extraction using ChatGPT3.5 Turbo**
Respondent 1, One-way interviews in the afternoon using external recording	"provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "anggota_keluarga": "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK"

	"nama": "Karina Cindy Rahmanto", "nik": 3319025403020002, "agama": "Islam", "deskripsi_pekerjaan": "Sedang berkuliah di Politeknik Statistika (STIS) jurusan komputasi statistik" "nama": "Sumiati", "nik": 3319025104450001, "agama": "Islam", "deskripsi_pekerjaan": "Pensiunan perawat di RS Rumah Sakit Islam Kudus"

Then the results of this information extraction stored in JSON format as can be seen in Table 5. Formatting into JSON is done in this study by adding the symbols { and , to the output results of data extraction. An example is as follows.

Table 5. Extracted information from survey interviews in JSON format

"provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "anggota_keluarga": "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK"	{ "provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "Anggota_keluarga": [{ "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" }, { "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK" }]}

The next step involves classifying the respondents' answers (information extraction results) according to their questions using a fine-tuned ChatGPT model. As shown in Figure 1, the fine-tuned ChatGPT-3.5 model successfully classified the answers from the Speech Recognition text according to the appropriate question entity. This is evidenced by the precision, recall, and F-1 Score, all of which reached 100%.

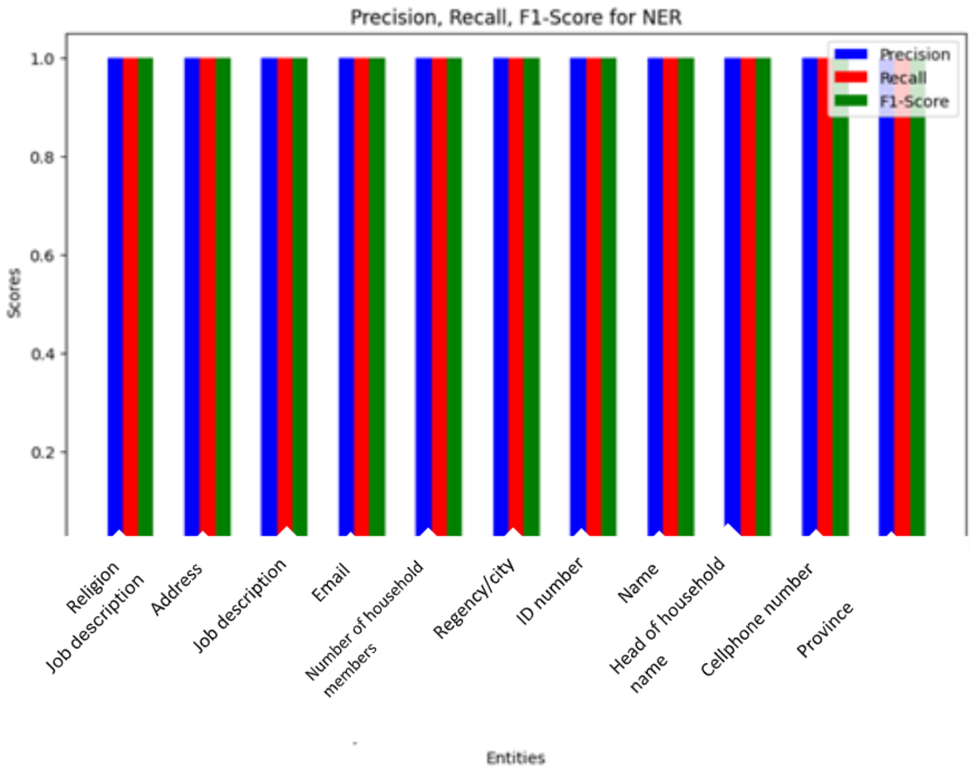

Figure 1. Precision, recall, dan f1-score

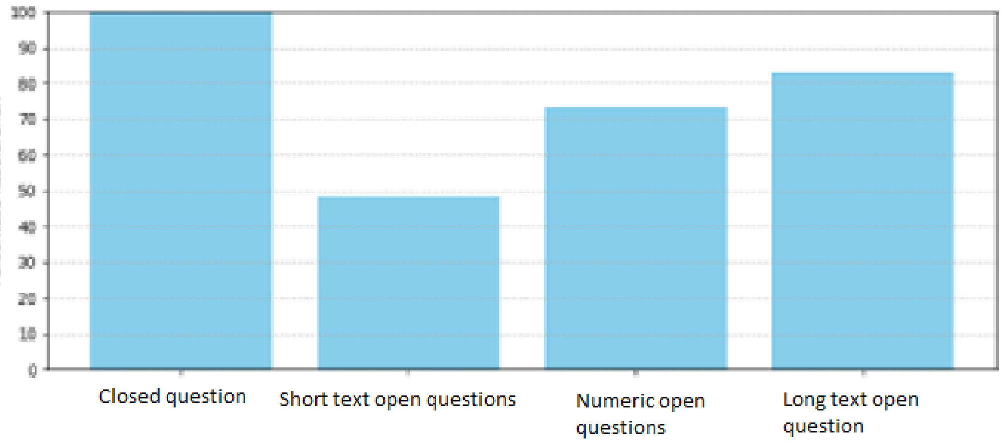

Figure 2. Percentage of correct answer accuracy by question type

The next step is to evaluate the accuracy of filling in the questions on the questionnaire based on the extraction results compared to the expected answers. To evaluate the results of information extraction, the questions in the questionnaire are grouped into several categories. The first category includes closed questions, covering province (provinsi), regency/city (kabupaten/kota), and religion (agama). The second category consists of short-text open questions, including head of household name (nama kepala rumah tangga), name (nama), and email. The third category encompasses numerical open questions, such as cellphone numbers (nomor HP), number of household members (jumlah anggota rumah tangga), and ID number (NIK). Finally, the fourth category includes long-text open questions, covering home addresses (alamat) and job descriptions (deskripsi pekerjaan). Figure 2 illustrates that the fine-tuned ChatGPT-3.5 model can perfectly answer closed-ended questions with clear standards, achieving a 100% success rate. However, the model is less effective at providing correct answers for open-ended questions without clear standards, especially in short-text open questions. This discrepancy is due to errors in the Speech Recognition translation, which result in incorrect extracted answers. Additionally, the most frequent errors occur with NIK digits that contain repeated numbers. But nevertheless, for all open-ended questions groups, the fine-tuned ChatGPT-3.5 model has produced satisfactory answers, achieving an accuracy rate of 83.01%.

4. Conclusion

Based on the results and discussion in the previous section, here are some conclusions that can be drawn: (i) ASR Performance: The Google Speech Recognition API effectively translates voice into text with an average Word Error Rate (WER) of 10%. Errors are more common with names, email addresses, and NIK (National Identification Numbers) due to their non-standardized spellings and the way numbers are read. (ii) Recording Conditions: One-way recordings and recordings with unscrambled questions yield lower WER values compared to direct web recordings and recordings with added noise. (iii) Data Extraction with ChatGPT-3.5 Turbo: The fine-tuned model successfully classifies answers based on question entities. However, the accuracy is affected by errors in the ASR output, particularly with names, emails, and NIKs. These findings highlight the importance of optimizing recording conditions and addressing specific challenges in ASR to improve overall performance. Additionally, refining the input data for models like ChatGPT can enhance the accuracy of automatic data extraction. This study is preliminary, so these things will be improved in subsequent research.

Then suggestions based on the research results are as follows: (i) Automatic Speech Recognition (ASR): Google Speech Recognition is not sufficiently effective for translating voices containing people's names and long digits. It is better suited for translating voices containing common words. It is recommended to use ASR for one-way recordings to achieve better accuracy. (ii) Large Language Model (LLM) API: Most LLM APIs are paid services. For data with large tokens, it is advisable to use open-source applications that provide LLMs, such as lmstudio. Ensure that the computer used has high capabilities to handle the processing requirements.

Ethics approval

This study was conducted in accordance with the ethical standards. Informed consent was obtained from all individual participants included in the study.

Acknowledgments

In this study, we would like to thank all respondents who were willing to participate in this study.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Lailatul Hasanah: Methodology, Software, Data Curation, Writing- Original Draft Preparation, Visualization. **Budi Yuniarto:** Conceptualization, Methodology, Supervision, Reviewing and Editing.

References

[1] UNECE HLG-MOS, *Large Language Models for Official Statistics*, United Nations Economic Commission for Europe – High Level Group on Modernisation of Official Statistics, Dec. 2023. https://unece.org/.

[2] R. Peng, K. Liu, P. Yang, Z. Yuan, and S. Li, “Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data,” *arXiv preprint*, arXiv:2308.03107 [cs.AI], 2023.

[3] J. Gao, H. Zhao, C. Yu, and R. Xu, “Exploring the feasibility of ChatGPT for event extraction,” *arXiv preprint*, arXiv:2303.03836, 2023.

[4] Pemerintah Republik Indonesia, *Undang-Undang Republik Indonesia Nomor 16 Tahun 1997 tentang Statistik [Law of the Republic of Indonesia Number 16 of 1997 concerning Statistics]*, 1997

[5] BPS, *Pemanfaatan Aplikasi CAPI (Berbasis Android) dalam Pendataan Updating PODES 2020 [Utilization of CAPI (Android-based) Application in Data Collection for Updating PODES 2020]*, 2020. https://bukittinggikota.bps.go.id/news/2020/07/15/41/pemanfaatan-aplikasi-capi--berbasis-android--dalam-pendataan-updating-podes-2020.html

[6] T. Takdir, “Analisis Kinerja, Kualitas Data, dan Usability pada Penggunaan CAPI untuk Kegiatan Sensus/Survey,” *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 10, no. 1, pp. 9–26, 2018.

[7] J. K. Höhne, K. Gavras, and J. Claassen, “Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys,” *Social Science Computer Review*, vol. 08944393231160961, 2022.

[8] W. Wicara, *Meningkatkan Produktivitas dengan Menggunakan STT [Enhancing Productivity by Using STT]*, 2023. https://widyawicara.com/meningkatkan-produktivitas-dengan-menggunakan-speech-to-text/

[9] Y. Feng, “Intelligent speech recognition algorithm in multimedia visual interaction via BiLSTM and attention mechanism,” *Neural Computing and Applications*, vol. 36, no. 5, pp. 2371–2383, 2024.

[10] T. Lenzner and J. K. Höhne, “Who is willing to use audio and voice inputs in smartphone surveys, and why?,” *International Journal of Market Research*, vol. 64, no. 5, pp. 594–610, 2022.

[11] J. Gao, H. Zhao, C. Yu, and R. Xu, “Exploring the feasibility of chatgpt for event extraction,” *arXiv preprint*, arXiv:2303.03836, 2023.

[12] C. Dinata, D. Puspitaningrum, and E. Erna, “Implementasi Teknik Dynamic Time Warping (DTW) pada Aplikasi Speech To Text [Implementation of Dynamic Time Warping (DTW) Technique in Speech-to-Text Applications],” *Jurnal Teknik Informatika*, vol. 10, no. 1, pp. 49–58, 2017, doi: 10.15408/jti.v10i1.6816.

[13] I. K. S. Buana, “Implementasi Aplikasi Speech to Text untuk Memudahkan Wartawan Mencatat Wawancara dengan Python [Implementation of Speech-to-Text Application to Facilitate Journalists in Recording Interviews using Python],” *Jurnal Sistem Dan Informatika (JSI)*, vol. 14, no. 2, pp. 135–142, 2020, doi: 10.30864/jsi.v14i2.293.

[14] F. Adnan and I. Amelia, “Implementasi Voice Recognition Berbasis Machine Learning [Implementation of Machine Learning-Based Voice Recognition],” *Edu Elektrika Journal*, vol. 11, no. 1, pp. 24–29, 2022.

[15] G. Gartlehner et al., “Data Extraction for Evidence Synthesis Using a Large Language Model: A Proof-of-Concept Study,” medRxiv, 2023-10, 2023

[16] Y. Zou et al., “ESGReveal: An LLM-based approach for extracting structured data from ESG reports,”arXiv *preprint*, arXiv:2312.17264, 2023

[17] T. Ricketts, *Speech Recognition Application With Tone Analyzer (Doctoral dissertation).* Alabama Agricultural and Mechanical University, 2023

[18] E. Vidal, “Finite-State Speech-to-Speech Translation,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

[19] S. Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley Professional, October, 2023.

[20] H. Dalianis, “Evaluation Metrics and Evaluation,” in *Clinical Text Mining*, Springer, Cham, 2018, doi: 10.1007/978-3-319-78503-5_6.