# Incorporating Natural Language Processing into Virtual Assistants: An Intelligent Assessment Strategy for Enhancing Language Comprehension

Dr. Franciskus Antonius[1], Purnachandra Rao Alapati[2], Mahyudin Ritonga[3], Dr. Indrajit Patra[4], Yousef A.Baker El-Ebiary[5], Myagmarsuren Orosoo[6], Manikandan Rengarajan[7]
Lecturer at School of Business and Information Technology STMIK LIKMI, Bandung Indonesia[1]
Associate Professor of English, Prasad V Potluri Siddhartha Institute of Technology,
Kanuru, Vijayawada, Andhra Pradesh, India[2]
Universitas Muhammadiyah Sumatera Barat[3]
An Independent Researcher, PhD from NIT Durgapur, West Bengal, India[4]
Faculty of Informatics and Computing, UniSZA University, Malaysia[5]
School of Humanities and Social Sciences, Mongolian National University of Education, Mongolia[6]
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai[7]

***Abstract*—The study introduces a comprehensive technique for enhancing the Natural Language Processing (NLP) capabilities of virtual assistant systems. The method addresses the challenges of efficient information transfer and optimizing model size while ensuring improved performance, with a primary focus on model pertaining and distillation. To tackle the issue of vocabulary size affecting model performance, the study employs the SentencePiece tokenizer with unigram settings. This approach allows for the creation of a well-balanced vocabulary, which is essential for striking the right balance between task performance and resource efficiency. a novel pre-layernorm design is introduced, drawing inspiration from models like BERT and RoBERTa. This optimization optimizes the placement of layer normalization within transformer layers during the pretraining phase. Teacher models are effectively trained using masked language modeling objectives and the Deepspeed scaling framework. Modifications to model operations are made, and mixed precision training strategies are explored to ensure stability. The two-stage distillation method efficiently transfers knowledge from teacher models to student models. It begins with an intermediate model, and the data is distilled carefully using logit and hidden layer matching techniques. This information transfer significantly enhances the final student model while maintaining an ideal model size for low-latency applications. In this approach, innovative measurements, such as the precision of filling a mask, are employed to assess the effectiveness and quality of the methods. The findings demonstrate substantial improvements over publicly available models, showcasing the effectiveness of the strategy within complete virtual assistant systems. The proposed approach confirms the potential of the technique to enhance language comprehension and efficiency within virtual assistants, specifically addressing the challenges posed by real-world user inputs. Through extensive testing and rigorous analysis, the capability of the method to meet these objectives is validated.**

***Keywords*—Natural language processing; virtual assistants; smart evaluation approach; artificial intelligence; human-computer interactions***

## I. Introduction

Our everyday lives have become more reliant on virtual assistants, which provide efficiency and convenience for a variety of activities, from setting notifications and handling calendars to answering inquiries and managing smart home devices [1]. The capacity of such virtual assistants to understand and interpret user input in natural language is essential to their effectiveness. The core of this language comprehension process is Natural Language Processing (NLP), which enables virtual assistants to comprehend the purpose behind user inquiries and deliver pertinent and contextually suitable replies. Although NLP research has made great strides, conventional virtual assistants frequently have trouble understanding complicated and nuanced user inputs. When their inquiries are incorrectly translated, users may become frustrated, which can result in unsatisfying results and decreased user engagement [2]. These drawbacks highlight the demand for more intelligent and complex methods of language interpretation in virtual assistants. A software programme with artificial intelligence combined with NLP (natural language characteristics are known as a virtual assistant [3]. It acts as a virtual friend that can converse with users in a manner like that of a human and help them with a variety of jobs and enquiries. This technologically advanced system responds to speech or text-based instructions, deciphers user intentions, and offers pertinent data, services, or recommendations in order to expedite and simplify daily tasks [4]. Virtual assistants are becoming a common feature of contemporary digital experiences on a variety of platforms; including computers, smart speakers, wearable technology, and smartphones. These assistants change the way people engage with technology by utilizing NLP to manipulate smart devices, play musical instruments manage appointments, send reminders, obtain weather information, and more [5]. Virtual assistants are anticipated to develop further as artificial intelligence technology progresses, growing better at comprehending context, recognizing individual preferences, and completing difficult

tasks, ultimately changing how we traverse our linked and digital lives.

Virtual assistants' usability and efficacy are greatly influenced by natural language processing (NLP). The manner in which people engage with technology is being revolutionized by these AI-powered assistants' ability to understand, interpret, and reply to human language thanks to NLP [6]. Speech recognition is one of the fundamental elements underlying NLP in virtual assistants, where algorithms translate spoken language into text that can be understood by machines. This allows the assistant to interpret voice instructions. Another crucial element is intent identification, which enables virtual assistants to understand the rationale behind a user's question and tailor their responses. Entity extraction is made easier by NLP, which helps virtual assistants find crucial information in user inputs like places or names [7]. Additionally, NLP empowers virtual assistants to keep track of context during interactions, resulting in more suitable and natural replies. With the ability to generate language, virtual assistants may provide replies that seem human and are customized to the preferences and communication preferences of the user [8]. Sentiment analysis improves the experience by enabling assistants to recognize and understand user emotions. Additionally, NLP offers multilingual assistance, serving a variety of user bases globally. A few virtual assistants also use machine learning algorithms with NLP characteristics for continuous learning, improving their language comprehension and replies over time in response to user input. Virtual assistants' usefulness has been greatly enhanced by the addition of NLP, which makes interactions more natural, individualized, and conversational. Virtual assistants are anticipated to become increasingly smarter as NLP technology develops, interpreting complicated questions and providing contextually appropriate replies that meet the individual needs of users [9].

SEA for understanding language basics marks a significant advancement in the fields of artificial intelligence and natural language processing. The capacity of robots to understand the subtleties of human language is an essential hurdle in today's world when technology is ingrained more deeply into our everyday lives. To meet this problem, the SEA emerged as a revolutionary approach that makes use of current developments in machine learning to get beyond the drawbacks of traditional language understanding techniques. Modern technology is based on language understanding, which enables smooth interactions between humans and machines [10]. Traditional methods, however, frequently fail to adequately capture the subtleties of language, setting, and purpose. By fusing context analysis, entity extraction, and intelligent intent identification into one seamless framework, the SEA makes a brave step forward. By doing this, it not only aims to improve the sensitivity and accuracy of language processing devices but it additionally presents the possibility of reshaping user experiences in a variety of sectors [11]. The SEA has the ability to fundamentally alter the design of virtual assistant systems. These AI-powered friends have become an essential part of our lives, assisting us with everything from managing calendars to operating smart gadgets. Involving SEA, virtual assistants will be able to go beyond what they are now capable of, increasing the breadth of their understanding and raising the quality of their relationships with users. The main goal of SEA is to close the gap between computer interpretation and the subtleties of human language, therefore enabling more intuitive and natural communication. The SEA's importance goes beyond only technology. The SEA equips virtual assistants to act as knowledgeable guides, expertly leading users throughout a large sea of data in the age of overload of information and rapid technological advancement [12]. The SEA not only increases efficiency but also creates the foundation for establishing trust between people and computers by allowing virtual assistants to comprehend contextual and user intent more precisely. This study intends to give a thorough analysis of the Smart Evaluation Approach's methodology, implementation tactics, and the intriguing research directions it opens upon our delve into its complexities. By combining artificial intelligence and natural language processing, SEA aims to reinvent the fundamentals of language understanding, pushing the limits of what virtual assistants can accomplish and fundamentally altering how we engage with technology [13].

The study presents a Smart Evaluation Approach (SEA) that aims to improve virtual assistants' ability to understand language. The SEA uses new developments in artificial intelligence and machine learning to address problems with traditional virtual assistants. The SEA seeks to greatly improve virtual assistants' accuracy and response to user inputs using intelligent intent identification, entity extraction, and contextual analysis. The study's major goal is to expand the capabilities of virtual assistants' existing language understanding techniques in order to promote more intuitive and natural human-machine interactions. A thorough assessment of the literature on NLP, AI, virtual assistants, and comprehension of language approaches is part of the paper's framework. The conceptual framework for the virtual assistant network and its integration with SEA are then discussed, followed by the approach used for SEA installation and assessment. The merits and possible improvements of the suggested technique are highlighted by the detailed experimental results and comparison analysis with conventional approaches [14]. The conclusion summarizes the research findings and suggests new avenues of inquiry for improving language comprehension in virtual assistants. The goal of using NLP and the smart assessment method is to unlock the potential of virtual assistants, alter technology engagement, and enable smooth and efficient human-machine communication. The key contributions of the research models are as follows:

- This study focuses on enhancing Natural Language Understanding (NLU) capabilities in a large-scale virtual assistant through language model pretraining and distillation techniques, specifically targeting intent classification and slot filling, which are critical components of effective language comprehension.
- The research addresses the challenge of understanding user intentions and identifying relevant slots in user inputs. For instance, given a query like "can you call mom," our NLU model should discern the intention to

initiate a call and identify the corresponding slot, in this case, the contact's name, marked as "mom".

- Throughout the paper, the research consistently refers to our models and pipeline as "Virtual Assistant Teacher Model(s) (VATM)". This nomenclature reflects the unique aspects of our problem domain, which diverges from traditional research tasks in several ways.
- This approach leverage relatively extensive labeled datasets, which is noteworthy as it introduces challenges and opportunities distinct from typical pretraining approaches.
- The research emphasizes the importance of optimizing model efficiency, as our models must operate within stringent latency and memory limitations, ensuring their practical utility in real-world scenarios.
- Research tackle the unique challenge of processing primarily spoken language data, distinguishing our work from the more common "written form" text used in the pretraining of publicly available models.

This system's capability extends across multiple languages, adding an additional layer of complexity to the language understanding process. The study tackles the intricate challenge of improving language understanding within virtual assistants by leveraging language model pretraining and distillation techniques. We navigate through unique challenges, including a sizable labeled dataset, performance constraints, spoken language input, and multilingual support, to enhance the overall NLU capabilities of our Virtual Assistant Teacher Models (VATMs).

## II. Related Works

Ait-Mlouk and Jiang [15] introduces "KBot", a novel ChatBot designed to harness the power of knowledge graphs and linked data for enhancing natural language understanding. With the increasing availability of structured data in the form of knowledge bases on the semantic web, the objective of the ChatBot is to make this information accessible and beneficial for end-users. The authors address several challenges associated with building such a ChatBot, including user query comprehension, support for multiple knowledge bases, and multilingual capabilities. The authors present an architecture that facilitates an interactive user interface, enabling effective communication between users and the ChatBot. They propose a machine learning-based approach that employs intent classification and natural language understanding to interpret user intents and generate SPARQL queries for retrieving relevant information from knowledge bases. Notably, the authors extend their system by incorporating a new social network dataset, 'myPersonality,' into existing knowledge bases, enhancing the ChatBot's ability to handle analytical queries. The system allows for the incorporation of new domains, offers flexibility in supporting multiple knowledge bases, and is designed to handle multilingual interactions. The paper also emphasizes the user-friendly creation and execution of various tasks across a broad spectrum of topics. The paper supports its claims with evaluation and application cases that demonstrate KBot's practical utility. These examples underscore how the ChatBot effectively navigates semantic data to cater to diverse real-world scenarios. The approach taken by the authors is particularly notable for its data-driven nature, leveraging knowledge graphs to provide insightful responses. The paper makes a significant contribution to the field of natural language processing and knowledge graph utilization. KBot's architecture and machine learning-based approach, along with its demonstrated adaptability and practical application, position it as a valuable tool for interactive and data-rich interactions.

Jungbluth et al. [16] delves into the integration of popular virtual assistants like Alexa, Siri, Cortana, and Google Assistant with industrial robotics, focusing on their role in controlling components of an intelligent robot assistant system for disassembly tasks. The authors introduce the paper by highlighting the increasing presence of virtual assistants in daily life, particularly their use as intuitive human-machine interfaces for device control through natural language. The core contribution of the work lies in its exploration of using virtual assistants to manage individual elements within a sophisticated industrial robot assistance system. After a succinct introduction and a survey of available virtual assistants, the authors present their system architecture, which seamlessly incorporates Amazon's Alexa using an Echo Dot device. Leveraging the Alexa Skills Kit, they develop a voice user interface encompassing various device functionalities and assistive behaviors. The authors detail the technical setup, which involves linking Alexa Voice Service with Amazon's Lambda and IoT web services. This connectivity facilitates the customization of machine commands based on users' voice inputs. Through intermediary components like a Raspberry Pi, they establish communication between the internet and the robot's isolated network. A notable aspect is the bidirectional communication flow, enabling real-time updates of device statuses in Amazon Web Services IoT shadow. This status information is subsequently utilized in Lambda functions to generate speech output using Alexa Voice Services and relayed through the Echo Dot for user notifications. The conclusion offers a balanced perspective by highlighting both positive and negative experiences encountered during their endeavors. The paper provides a comprehensive case study of integrating virtual assistants, particularly Amazon's Alexa, into the realm of industrial robotics. The technical details, architecture overview, and demonstration of a use case collectively exemplify the potential benefits and challenges of combining virtual and robot assistants. The work underscores the practicality of leveraging virtual assistants in complex real-world scenarios and contributes to the growing understanding of their integration within industrial applications.

Alagha and Helbing [17] responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri" aims to assess and compare the accuracy and quality of responses provided by Amazon Alexa, Google Assistant, and Siri to consumer health queries regarding vaccine safety and usage. The study employs a rubric-based scoring system to evaluate the responses of each voice assistant across 54 questions related to vaccination. The

evaluation criteria include the accuracy of the answer given through audio output and the credibility of the source supporting the response. The findings of the study reveal significant differences in the performance of the three voice assistants. Siri obtains the highest average score of 5.16 points, followed closely by Google Assistant with an average score of 5.10 points. In contrast, Alexa lags behind with a notably lower average score of 0.98 points. The results indicate that Google Assistant and Siri excel in accurately interpreting voice queries and providing users with authoritative sources of information about vaccination. However, Alexa struggles to comprehend queries and relies on sources different from those used by the other two assistants. The authors conclude that those involved in patient education should be cognizant of the varying quality of responses provided by different voice assistants. They also suggest that developers and health technology experts should advocate for improved usability and transparency in terms of information partnerships as these devices continue to evolve in their capabilities to deliver health-related information. The article is available under the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, permitting others to share, adapt, and build upon the work non-commercially, provided appropriate credit is given, changes are indicated, and usage remains non-commercial. The study contributes valuable insights into the performance of voice assistants in delivering accurate and reliable health information to users. The findings underscore the importance of further refining these technologies to ensure consistent and high-quality responses, especially in critical domains such as health information dissemination.

Villegas-Ch et al. [18] explores the implementation of a virtual assistant for managing academic aspects within a university using artificial intelligence (AI). In light of the ongoing pandemic, private universities are encountering challenges across academic and financial domains. Learning difficulties have contributed to increased dropout rates, exacerbating financial strains. Additionally, economic impacts from the pandemic have led to a decline in students seeking private education. These circumstances necessitate support measures to enhance student enrollment, safeguard budgets, and optimize resources. The academic realm poses significant efforts to manage academic activities while prioritizing those interested in pursuing educational programs. To address these complex challenges, integrating technologies like Chatbots, powered by artificial intelligence, emerges as a solution. By leveraging AI-powered Chatbots, universities can delegate tasks such as providing information about academic courses. This offloads administrative burdens and simultaneously enhances the user experience, thereby encouraging greater participation in the university community. The integration of AI-powered Chatbots can offer multifaceted benefits. These tools can efficiently handle information dissemination about academic courses, freeing up human resources for more critical tasks. They contribute to a smoother and more engaging user experience, potentially attracting more prospective students. This technology aligns well with the broader trend of digital transformation in education, offering personalized and immediate support to users. However, it's important to consider the potential challenges and limitations of implementing such systems. Ensuring accurate and contextually relevant responses, maintaining data privacy, and addressing potential technical glitches are areas that require careful attention. The paper underscores the need for private universities to adapt to the current circumstances by embracing technological solutions like AI-driven Chatbots. These tools hold the promise of enhancing student enrollment, reducing administrative burdens, and ultimately improving the overall efficiency and effectiveness of academic management in the face of evolving challenges.

Dong et al. [19] provides a brand-new method called the Universal pre-trained linguistic model (UniLM), which is intended to handle both responsibilities of comprehending and producing natural language. Utilizing A Transformer network that is shared among multiple users goes through pre-training utilizing a variety of language modeling tasks, including unidirectional, bidirectional in nature, and sequence-to-sequence prediction, this is accomplished. Specific self-awareness masks are used to identify the pertinent context for predictions to achieve unified modeling. On the frequently utilized GLUE measure as well as on challenging tasks including SQuAD 2.0 and CoQA problem answering, UniLM performs favorably when compared to BERT. The study demonstrates how UniLM outperforms industry standards on five naturally language-generating datasets. The ROUGE-L scores for CNN/DailyMail abstract summarization and Gigaword abstractive summarizing both saw increases of 2.04 absolute points and 0.86 absolute points, respectively, reaching 40.51 and 35.75, respectively. UniLM makes major advancements in generative question-answering tasks in addition to summarization. It produces an astounding 82.5 per cent increase in the F1 score with CoQA generating question answering. Additionally, the article documents significant gains in the DSTC7 document-grounded dialogue answer generating NIST-4 rating (achieving 2.67, with individual performance at 2.65), as well as the SQuAD controversy generating BLEU-4 score (3.75 absolute enhancements, reaching 22.12). The article proposes UniLM, a brand-new unified already trained linguistic model that performs very well on challenges requiring both interpretation and creation of natural language. The algorithm's superiority over current state-of-the-art models and outstanding performance on numerous benchmarks and datasets emphasize its potential to enhance the study of the processing of natural language. The research approach and conclusions in this work provide a substantial contribution to the creation of more powerful and adaptable language representations for use in real-world situations.

An overview of the main points, benefits, and drawbacks of the relevant papers are provided in the Table I. While having trouble with language support and user query comprehension, Ait-Mlouk and Jiang [15] employ natural language processing and machine learning-based intent categorization, utilising knowledge graphs to deliver perceptive responses. In order to overcome the hurdles involved in this integration, Jungbluth et al. [16] incorporate Amazon's Alexa with industrial robot support systems for real-time communication. Voice assistants' answers to health-related questions are compared using a rubric-based rating

system by Alagha and Helbing [17], exposing differences in response quality. Chatbots driven by AI are used by Villegas-Ch et al. [18] to improve administrative effectiveness and information distribution in higher education. The Universal pre-trained linguistic model (UniLM), as introduced by Dong et al. [19], improves F1 scores and natural language generation performance on multiple benchmarks in natural language processing tasks.

TABLE I. OVERALL SUMMARY OF LITERATURE REVIEW

| Reference | Technique | Merits | Limitation |
|---|---|---|---|
| Ait-Mlouk and Jiang [15] | Natural language processing and intent classification powered by machine learning. | Applying knowledge graphs to provide perceptive answers and integrating other data areas | Difficulties with language support, multiple knowledge base support, and user query comprehension. |
| Jungbluth et al. [16] | Integration of Amazon's Alexa into industrial robot assistance systems, technical setup, and real-time communication. | Integration of popular virtual assistants like Alexa into industrial robotics, offering a use case in industrial applications. | Challenges associated with the integration of virtual assistants into industrial robotics and practical use cases in the real world. |
| Alagha and Helbing [17] | Rubric-based scoring system to evaluate the responses' accuracy and credibility. | Comparative assessment of Amazon Alexa, Google Assistant, and Siri in responding to health | Differences in the quality of responses provided by voice assistants to consumer health queries. |
| Villegas-Ch et al. [18] | Integration of AI-powered Chatbots to handle information dissemination and streamline administrative tasks. | Improved distribution of information regarding educational programs, increased efficiency, and an enhanced user experience within the context of higher education. | Leveraging AI-driven Chatbots for the administration of academic functions in privately-owned educational institutions. |
| Dong et al. [19] | Development of the Universal pre-trained linguistic model (UniLM) using shared Transformer networks, self-awareness masks, and pre-training using language modeling tasks. | Enhancements in performance across diverse natural language generation tasks, such as summarization and question-answering, resulting in impressive F1 scores and leading outcomes on various evaluation benchmarks. | Introducing the Universal pre-trained linguistic model (UniLM) for natural language processing (NLP) assignments. |

## III. PROBLEM STATEMENT

The domain of virtual assistant systems faces a substantial challenge in achieving both efficiency and accuracy in language understanding. Striking the right balance between model size and performance is paramount, particularly for applications that require low-latency responses. However, the current landscape is marked by models that tend to be overly large, leading to latency and resource constraints, or simplified versions that sacrifice language understanding capabilities [20]. The evaluation of these models poses significant challenges, particularly with regard to the limitations of conventional metrics like perplexity, which can be influenced by tokenization choices and may not accurately reflect real-world performance. Existing evaluation methods may not fully capture the intricacies of language comprehension necessary for virtual assistant tasks that involve precise understanding and response to user intents and slots. Therefore, the central challenge is to develop an approach that distills knowledge from larger models into smaller ones while either preserving or enhancing their language understanding capabilities. This involves addressing the delicate balance between reducing model size and maintaining performance standards. Innovative evaluation metrics are essential to align with the specific requirements of virtual assistants, offering a comprehensive gauge of language understanding quality. The ultimate goal is to establish a robust methodology that tackles the dual challenges of model efficiency and performance, while introducing novel evaluation techniques tailored to the demands of real-world virtual assistant applications. Successfully addressing this challenge promises the development of highly efficient yet accurate language understanding models, ultimately revolutionizing virtual assistant technology and elevating user experiences across a wide range of domains, including customer service and personal assistants.

## IV. PROPOSED METHODOLOGY FOR EVALUATION OF LANGUAGE UNDERSTANDING

The methodology employed in this study follows a comprehensive approach to enhance language understanding capabilities. The process starts with the selection of diverse pretraining datasets, including the multilingual Colossal Clean Common Crawl (mC4), CC100 dataset, and Wikipedia data. These datasets encompass various domains, languages, and tones. Incorporating twelve languages for pretraining, such as Arabic, English, French, and more, establishes a robust foundation for multilingual comprehension. The sampling process, guided by a multinomial distribution, ensures proportional representation of languages while up-sampling low-resource ones. Preprocessing involves organizing sentences into sequences and dynamic tokenization during training. Additionally, a Stage 2 pretraining dataset, comprising anonymized utterance text from the system, undergoes refinement through duplication reduction, length filtering, and integration with public datasets. Tokenization strategy employs a SentencePiece tokenizer with intrinsic metrics, optimizing vocabulary size for effective tokenization. During pretraining, a modified architecture, introducing pre-layernorm components, aids in capturing both intra- and inter-sentence structures. Stage 2 pretraining focuses on enhancing

the model's specialization for virtual assistant utterances. Distillation techniques are applied in two phases, with an intermediate-sized model distilled from the large teacher model, followed by the use of this distilled model as a teacher for the final, smaller student model. Validation leverages the "mask-filling accuracy" task to monitor progress and performance.

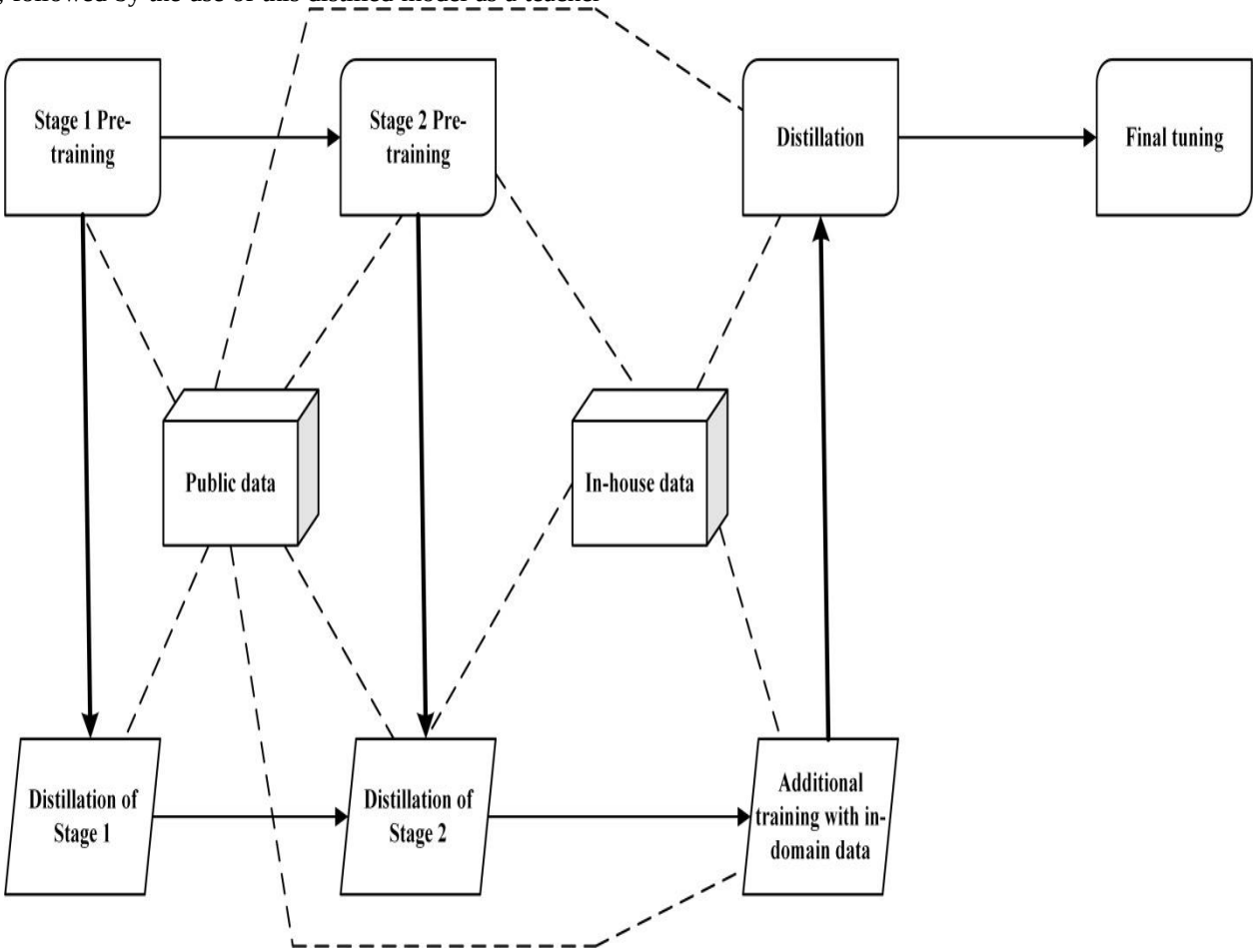


Fig. 1. Proposed model training and evaluation framework.

### A. Training Datasets

Pre-training datasets, which include a wide variety of data covering different areas, zones, languages, and more, are an essential basis for improving language processing abilities. This study takes into account a trio of primary sources of pre-training data: Wikipedia data, which helped train BERT, mBERT, and the BooksCorpus; the multiple-language Colossal Clean Common Crawl (mC4) a database, used for T5 and mT5 training; the CC100 a database, used to instruct XLM-R. Notably, Common Crawl data are used to create the mC4 and CC-100 datasets. A systematic strategy was used for selecting phrases from the training corpus, which included twelve languages for pre-training: the language of Arabic, English, French, German, Italian, Japanese, Marathi, Portuguese, Spanish, Tamil, and Telugu are all examples of supported languages. To ensure representation across languages, phrases were chosen for sampling using a multinomial distribution in accordance with predetermined rules. By up-sampling countries with fewer examples, the selection procedure, which was controlled by a multinomial distributing equation, aimed to achieve a balance. This technique successfully improved low-resource languages, resulting in a more varied dataset.

TABLE II. DATASET

| | Section 1 | Section 2 | Section 3 |
|---|---|---|---|
| data to Train | 95K | 84 K | 70 K |
| Validation data set | 15K | 12 K | 20 K |
| Size of test data | 25K | 24 K | 40 K |
| # of intents | 17 | 9 | 15 |
| # of slots | 97 | 28 | 58 |

Table II provides an overview of the dataset used for training, validation, and testing in the context of our language understanding model. The dataset is divided into three distinct domains, each representing a different aspect or category of language understanding.

The preprocessing method improved the dataset's quality. Sentences were organized into sequences of about 700 words,

and dynamic tokenization was carried out while training. This method made sure that sequences stayed inside the 1,024-token limit after tokenization while maintaining sequence integrity. In addition to publicly available datasets, a private Stage 2 prior instruction dataset made up of unlabeled and anonymous utterance text was included. There were several preparation stages for this dataset. Instances with lesser than five tokens were deleted, and duplicates have been minimized by keeping just a portion of their original total. In order to create an exhaustive collection for Stage 2 pretraining, this private data was pooled in a 1:2 ratio with the open Stage 1 pretraining database to reduce catastrophic forgetting. In the end, our efforts produced a Stage 2 pretraining database with over 50 million cases. These datasets were smoothly included in the training pipeline that was developed, as shown in Fig. 1. The need of creating a thorough and comprehensive pretraining database to improve language comprehension models is highlighted by this method of data collecting, curation, and preparation [21].

### *B. Text Pre-processing using Tokenization*

This approach revolves around the application of a SentencePiece tokenizer trained in the unigram setting, with the aim of enhancing Natural Language Processing within the realm of Virtual Assistants. The vocabulary size of the tokenizer emerged as a pivotal factor impacting the overall system performance. While larger vocabulary sizes often yield performance improvements in tasks like masked language modeling, they also introduce trade-offs, such as slower training convergence, heightened memory consumption during inference, and increased latency. Considering the resource-intensive nature of training an extensive teacher model with a wide-ranging tokenizer vocabulary, we introduced two intrinsic tokenizer metrics: the split-ratio and unk-token fraction. These metrics enabled us to meticulously balance performance and resource utilization without necessitating protracted teacher model training. The split ratio metric, grounded in the principle that a higher count of subword splits can compromise overall accuracy, guided our optimization efforts. Meanwhile, the unk-token fraction metric, which gauges the prevalence of unknown tokens in output, emerged as a pivotal performance determinant. A higher proportion of unk-token fractions negatively impacted the overall system performance. To align our tokenizer's split-ratio and unk-token fraction with baseline production models, we strategically incorporated an extensive set of 2,136 frequently used kanji characters in Japanese. This set was complemented by a comprehensive array of hiragana and katakana symbols, thus ensuring robust coverage of Japanese characters. Our strategy also involved partitioning data in a 70/30 ratio between spoken and written forms. This discerning distribution facilitated the attainment of a balanced vocabulary size, ultimately totaling 150,000 subword tokens. This vocabulary size was in harmony with the effective approach adopted in our pretraining corpus strategy. The foundation of our methodology for integrating Natural Language Processing capabilities into Virtual Assistants centers on a sophisticated SentencePiece tokenizer, meticulously trained in the unigram setting. The intrinsic tokenizer metrics, namely the split-ratio and unk-token fraction, were harnessed to strike a harmonious equilibrium between task performance and efficient resource utilization. Additionally, the thoughtful incorporation of diverse Japanese characters ensured comprehensive language coverage. This astute evaluation approach is poised to elevate language comprehension within the domain of Virtual Assistants.

### *C. Stage 1 and Stage 2 Pre-training Model*

In the initial pre-training phase, we drew inspiration from established models like BERT, RoBERTa, and XLM-R to shape our approach. While our Virtual Assistant models found their foundation in RoBERTa, a distinct innovation emerged through the implementation of pre-layernorm architecture. This architectural adjustment involved placing layer normalization immediately before the self-attention and feed-forward blocks within each transformer layer. Central to this training process was the masked language modeling objective, where 15% of tokens within the text were masked. Among these masked tokens, 10% were maintained unchanged, while an additional 10% were substituted with random tokens. Our teacher models underwent training with a focus on scalability, culminating in the management of up to 9.3 billion non-embedding parameters. To enhance training throughput, the Deepspeed framework came into play, capitalizing on its two-stage strategy. In the first stage, optimizer states were distributed across GPUs, followed by gradient partitioning in the second stage. Notably, this was executed without introducing network-based bottlenecks. Employing mixed precision training was pivotal in optimizing computational efficiency. This technique enabled us to achieve an impressive computational output of 107 TFLOP/sec per GPU for an encoder housing 9.3 billion parameters. Our infrastructure was grounded in AWS p4d.24xlarge instances, housing Nvidia a100 GPUs and leveraging Elastic Fabric Adapters to ensure steadfast network throughput. Throughout the pretraining journey, Deepspeed's mixed precision training mechanism remained our companion. Nonetheless, some model operations encountered challenges related to FP16 overflow. To address these concerns, two key modifications were introduced. Firstly, the baddbmm operation took the place of the matmul operation for query-key multiplication. Secondly, a conversion to FP32 was performed before variance computation during the layer normalization process. These changes, while slightly decreasing throughput by up to 20%, successfully mitigated instability issues within the model. It's worth noting that an alternative avenue to handle stability concerns entails the utilization of BFLOAT16. However, this path wasn't available within the Deepspeed framework during our experimentation phase. Transitioning to Stage 2 pretraining, our exploration delved into the Muppet system. Unlike the initial phase, Stage 2 pretraining employed a more direct approach. We extended the pretraining objective using our designated Stage 2 dataset. The primary objective here was to enhance the model's proficiency in handling virtual assistant-specific utterances, which are often brief and may deviate from strict grammatical norms. A careful balance was sought between enhancing specialized capabilities and retaining the broader language knowledge gained during Stage 1.

### *D. Distillation*

Given the imperative of modest model sizes for low-latency applications, a direct distillation from extensively

large teacher models to considerably smaller student models might hinder the effective transfer of the teacher's expertise. As a remedy, a two-stage teacher assistance setup was devised for the distillation process, as illustrated in Fig. 1. This strategy aimed to strike a balance between knowledge transfer and model size reduction. In the initial stage, the immense teacher model was compressed into an intermediate-sized model. Subsequently, the final student model was trained using this intermediate model as a guide. This approach ensured that the transfer of knowledge from the teacher to the student was well-optimized, despite the significant size reduction. Drawing inspiration from the teacher's pretraining methodology, a distillation process was initiated from a randomly initialized student model. Convergence in training signaled a seamless transition to the deployment of the Stage 2 teacher model, thus continuing the distillation process. Importantly, the data employed for distillation in both stages remained consistent with the data utilized for teacher pretraining in their respective stages. Within the intermediate student/teacher pairing, a balanced blend of categorical cross-entropy (MLM loss) and soft cross-entropy was applied, with equal weighting. Remarkably, experimentation indicated no substantial benefits from incorporating the attention and hidden layer outputs of the teacher model. Transitioning to the final student model, a dual-stage process was undertaken. First, the intermediate model underwent further pretraining, exclusively using Stage 2 data and without teacher involvement. Subsequently, a distillation procedure was executed to seamlessly transfer knowledge to the compact final student model. During this distillation phase, techniques mirrored those employed in the initial distillation, with the addition of hidden-layer output matching. In essence, the approach mirrors the core principles of the process outlined in the source paper. The process ensures effective knowledge transfer while mitigating the challenges arising from substantial model size reductions.

### *E. Validation of Model Performance without Fine-Tuning*

In order to effectively monitor the progress of our training efforts, a commonly employed technique is evaluating perplexity on a separate validation dataset. However, a notable drawback of perplexity measurements is their susceptibility to the tokenizer's specific characteristics. To overcome this limitation, this study introduced an innovative evaluation metric called "mask-filling accuracy", designed to enhance the comparability of different models. The formulation of these metric involved curating texts from diverse public tasks, encompassing resources like XNLI, PAWS-X, and Multi-lingual Amazon Reviews. It is worth noting that we deliberately excluded these specific examples from our training dataset. For each instance within this curated dataset, we leveraged the Stanza tagger to identify a noun word. Subsequently, all subword tokens corresponding to that noun were concealed. The model's task was to accurately predict all subword tokens associated with the hidden noun, with successful predictions deemed correct. A notable insight, highlighted in Fig. 2 and 3, reveals a robust correlation between perplexity measurements and mask-filling accuracy, as well as the model's performance on the XNLI benchmark. This correlation persists across various stages of model updates. This finding underscores the valuable potential of our mask-filling accuracy metric as a reliable indicator of model quality. Importantly, this metric transcends the challenges introduced by the nuances of different tokenization approaches. This novel approach to validating model performance without fine-tuning presents promising insights. It serves as an effective means of assessing the quality of models, bypassing the inherent limitations posed by tokenizer choices. This strategy, as outlined in the original paper, holds promise for similar applications in various contexts.

## V. Result and Analysis

Analysis of the performance and effectiveness of language understanding models within the realm of virtual assistant systems. Our exploration was underpinned by a meticulous evaluation process, which encompassed multiple stages and methodologies, ultimately yielding insightful results and findings. First, we investigated the relationship between XNLI accuracy and perplexity, as well as mask-filling accuracy, using the approach of a 2.3 billion parameter model. This analysis was crucial for understanding the interplay between these metrics and their correlation with model performance over various updates. Notably, we observed that the mask-filling accuracy exhibited a stronger correlation with XNLI accuracy during the no-fine-tune validation process, indicating its potential as a more informative gauge of model quality, less susceptible to the intricacies of tokenization choices. Moving forward, the approach focused on the evaluation of our distilled models in comparison to publicly available models, utilizing comprehensive training datasets specific to our system. Our assessment involved benchmarking against XLM-R Base with 85 million non-embedding parameters and the multilingual DistillBERT with 42 million non-embedding parameters. Notably, our distilled models consistently outperformed the public models in terms of exact match error rate. Most promisingly, our compact 17 million-parameter model demonstrated a remarkable 4.23% improvement over XLM-R, while retaining its performance margin over our larger 170 million-parameter model, showcasing an improvement of 4.82% over XLM-R. In the pursuit of holistic evaluation, we delved into the performance of our models within the full framework of a virtual assistant system. To accomplish this, an intermediate-sized model with 170 million non-embedding parameters acted as a teacher-assistant to distill the final student models. This compression journey involved multiple stages of distillation, leveraging diverse datasets and employing logit matching and hidden layer matching techniques. Our models underwent extensive testing via parallel A/B testing and sequential testing, simulating real-world scenarios. These evaluations encompassed automated measures of user dissatisfaction, particularly tail dissatisfaction, and the offline Semantic Error Rate (SemER) that evaluates intent and slot-filling performance.

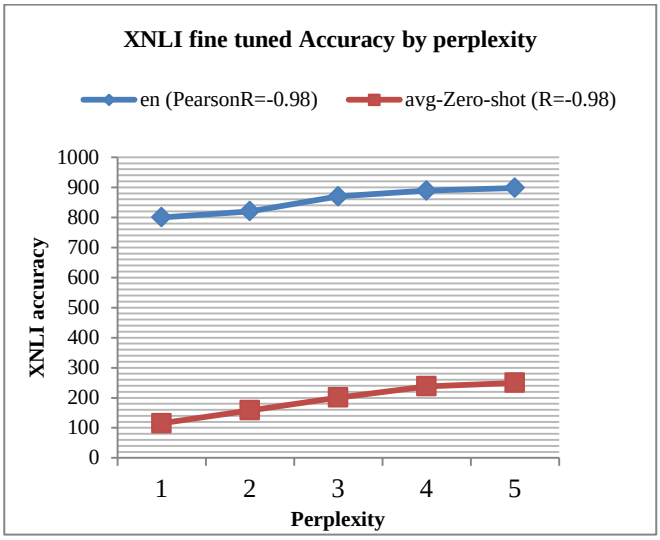


Fig. 2. XNLI accuracy from perplexity.

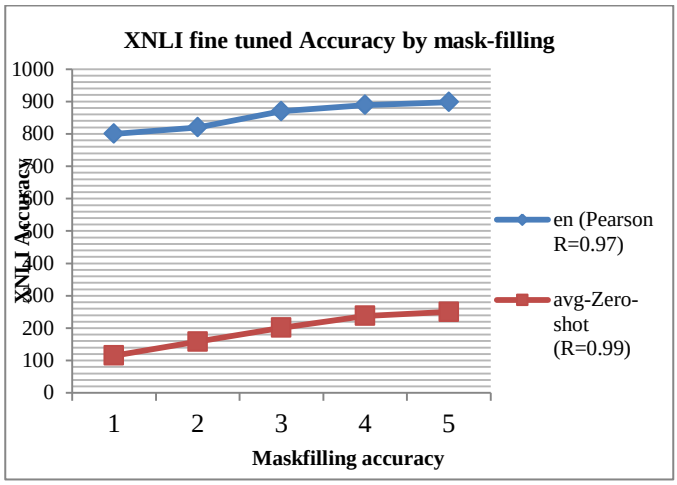


Fig. 3. XNLI accuracy as of mask-filling accuracy.

Fig. 2 and Fig. 3 utilizing the approach of 2.3Bparameter model, the relationship between XNLI accuracy and perplexity, as well as mask-filling accuracy, are being studied over model updates. The measure performs better for no-fine-tune validation the higher the correlation.

### A. NLU Results following Distillation

Our evaluation involved comparing the performance of our distilled models against publicly available models, using the complete training datasets specific to our system, as detailed in Section 3.4. The considered public models encompass XLM-R Base, featuring 85M non-embedding parameters, and the multilingual DistillBERT, comprising 42M non-embedding parameters. For an example to be counted as an exact match, the model must correctly predict both the intent and all associated slots. Encouragingly, both of our distilled models exhibit superior performance compared to the public models on average. Our 17M-parameter model, which shows an enhancement of 4.83% over XLM-R while showing only little decline when compared to our 170-M-parameter version (an increase of 4.63% over XLM-R), is particularly notable.

### B. Full System Results

This research follows the design described in Section 2.6 to provide a thorough evaluation of the model's effectiveness within the context of a whole virtual assistant system. The teacher assistant for the distillation of the final student models is an intermediate-sized model with 170M non-embedding parameters. This intermediate model underwent many steps of compression, including 160K entries of distillation from a 700M-parameter first-phase teacher, 105K updates from the Phase 1 2.3B-parameter teacher, and 300K updates from the second Stage 2 2.3B-parameter model. Further details on hyperparameters for the 700M-parameter model can be found in Appendix A. The 170M-parameter model, post fine-tuning with a task-specific dataset for 15,625 updates, acted as the teacher for distilling 17M-parameter student models. Distillation employed logit matching and hidden layer matching, following the student-teacher layer mapping (0, 1, 2, and 3) to (3, 7, 11, 15). Optimal performance emerged from utilizing two checkpoints within the same 17M-parameter model distillation process: one after 80M examples and the other after 200M examples. This process involved the inclusion of 9 languages: English, French, German, Hindi, Italian, Marathi, Spanish, Tamil, and Telugu. Two baseline models, each constituting a 5M-parameter monolingual encoder distilled from a BERT-Base architecture teacher, were considered. These baselines employed Wikipedia dumps in the relevant languages, following the text conversion to spoken form. Our comprehensive study occurred within a virtual assistant system experimentation platform. Our models were subjected to both parallel A/B testing, involving distinct user cohorts, and sequential testing with the same user cohort. The assessment encompassed automated measures of user dissatisfaction across the entire virtual assistant system, alongside considerations for tail dissatisfaction (related to less common utterances). The offline Semantic Error Rate (SemER) evaluations were conducted, jointly evaluating intent and slot-filling performance. SemER considers correct slots, deletion errors, insertion errors, and substitution errors, along with intent classification errors. The methodology employed aligns with the foundational principles outlined in the original paper. Our approach showcases promising potential for effectively assessing model performance within the intricate landscape of virtual assistant systems.

$$SemER = \frac{\#\ Deletion + \#\ Insertion + \#\ Substitution}{\#\ Correct + \#\ Deletion + \#\ Substitution} \quad (1)$$

The 2.3B-parameter Phase 2 model, the distilled 170-M-parameter Phase 2 model, and the 17-M-parameter Phase 2 model were assessed The results are shown in Tables III and IV, respectively. A negative number implies a lower error rate compared to the Stage 1 baseline model with 2.3B parameters.

TABLE III. FULL FINE-TUNING

| **Complete-fine-tuning** | | | | |
|---|---|---|---|---|
| Reduced Relative-Intent-Class-Error Versus 2.3-B Phase 1 | | | | |
| | *Section 1* | *Section 2* | *Section 3* | *Average* |
| 2.3-B Phase -2 | -4.72% | -1.78% | -6.79% | -2.76% |
| 170-M from 2.3-B | -3.28% | -3.98% | -2.36% | -1.97% |
| 17-M from 170-M | 10.56% | 9.78% | 9.75% | 9.98% |
| Reduced Relative-Slot-Filling-Error Versus 2.3-B Stage 1 | | | | |
| | *Section 1* | *Section 2* | *Section 3* | *Average* |
| 2.3-B Phase-2 | -6.02% | -10.05% | -6.68% | -8.01% |
| 170-M from 2.3-B | -1.62% | -11.03% | -9.53% | -9.68% |
| 17-M from 170-M | 28.07% | 3.11% | 4.36% | 11.51% |

TABLE IV. FROZEN-ENCODER RESULTS

| **Frozen Encoder** | | | | |
|---|---|---|---|---|
| Improvement of the Relative-Intent-Class-Error Versus 2.3B Stage 1 | | | | |
| | *Section* 1 | *Section* 2 | *Section* 3 | Average |
| 2.3-B Phase-2 | -11.61% | -4.61% | -2.78% | -6.98% |
| 170-M from 2.3-B | -16.09% | -17.23% | -12.09% | -16.70% |
| 17-M from 170-M | 12.99% | 7.49% | 11.89% | 11.78% |
| Improvement in Relative-Slot-Filling-Error Versus 2.3B Stage 1 | | | | |
| | *Section* 1 | *Section* 2 | *Section* 3 | Average |
| 2.3-B Phase-2 | -5.56% | -18.98% | -5.79% | -9.31% |
| 170-M from 2.3-B | -6.17% | -11.93% | -2.90% | -6.99% |
| 17-M from 170-M | 18.46% | -6.90% | 2.90% | 5.01% |

TABLE V. RESULTS FROM A PLATFORM FOR TESTING VIRTUAL ASSISTANTS

| | Experiment 1 | Experiment 2 |
|---|---|---|
| *Non-Embed Base Teacher Parameters* | 85-M | 85-M |
| *FF Size, Hidden Size, and Base Layers* | 4/312/1200 | 4/312/1200 |
| *Non-Embed Base Parameter Count* | 5-M | 5-M |
| *Support for Base Langs* | 1 | 1 |
| *Non-Embedded Parameters for Cand Teachers* | 2.3-B | 2.3-B |
| *Non-Embed Parameters for Cand Teachers* | 170-M | 170-M |
| *FF Size, Hidden Size, and Cand Layers* | 4/768/1200 | 4/768/1200 |
| *Non-Embed Cand Params* | 17-M | 17-M |
| *Cand Langs was backed* | 9 | 9 |
| *Cand Distil Illustrations* | 80-M | 200-M |
| *Testing Position* | 1 | 2 |
| *Entire Solution User Discontent A/B* | -3.74% | -4.91% |
| *Tail A/B for Entire Solution User discontent* | -10.3% | -7.50% |
| *Users' Overall discontent Score* | -14.9% | -7.2% |
| *Inactive SemER* | -15.6% | -2.98% |

In Table V, the findings presented are from two different experiments (Exp) carried out in distinct locations using a virtual assistant experimentation platform comparing our 17-M-parameter candidate model (Cand) is compared to the reference model (Base), which was developed by an 85-M-parameter teacher using Wikipedia data. Relative findings from an A/B test conducted concurrently with a distinct user cohort and an alternating test conducted using identical users are both shown for the computerized metric of whole-system user discontent. The tail A/B results from utterances beyond of the top 500 are also presented.

### C. Discussion

The presented findings and outcomes from the performance evaluation of language comprehension models in the context of virtual assistant systems represent a meticulous and thorough investigation of these models. In order to maximise the utility of these models for practical applications, this research emphasizes the complex nature of the fine-tuning and distillation processes [22]. The study started by exploring the connection between XNLI accuracy, perplexity, and mask-filling accuracy, offering insightful information about how these metrics relate to model performance across various updates [23]. Notably, the mask-filling accuracy showed strong model quality indicators, especially in the absence of fine-tuning, indicating its potential as a more accurate performance indicator that goes beyond tokenization intricacies. As the evaluation progressed, the emphasis shifted to evaluating the performance of the distilled models against publicly accessible models. The distilled models consistently outperformed public models in the assessment, especially in terms of exact match error rate, which took into account a variety of parameters, including non-embedding parameters. Particularly, the compact 17 million-parameter model showed striking improvements over reference models, demonstrating the possibility of developing more effective and efficient models in the context of virtual assistants. The research expanded its evaluation to take into account the overall performance of these models within the more general framework of a virtual assistant system, moving beyond model-centric assessments. This required using multiple datasets, logit matching, and hidden layer matching techniques in a multi-stage distillation process.

In order to simulate real-world situations and gauge user dissatisfaction, particularly for less frequent utterances, the evaluation included parallel A/B testing and sequential testing in addition to the offline Semantic Error Rate (SemER), which assesses intent and slot-filling performance [23]. The outcomes of these studies showed concrete advantages, with decreases in relative intent-class error and slot-filling error compared to baseline models, highlighting the efficiency of the distillation process in improving model performance within the complex environment of virtual assistant systems. User dissatisfaction scores significantly decreased for both the system as a whole and for less frequent utterances, highlighting the tangible enhancements in the user experience. This study offers a thorough and organized assessment of language comprehension models in the context of virtual assistant systems. The results point to a promising development in the development of virtual assistants and their function in enhancing human-machine interactions: the approach of fine-tuning and distillation can result in more effective, accurate, and user-friendly models.

The full virtual assistant system analyses that, our analysis showcased the comprehensive assessment we conducted. An intermediate-sized model with 170 million non-embedding parameters served as a teacher-assistant, distilling final student models. Our multi-stage approach to distillation, which involved leveraging diverse datasets and employing logit matching and hidden layer matching techniques, showcased the effectiveness of knowledge transfer from teacher to student models [24]. These distilled models were extensively tested using both parallel A/B testing and sequential testing, simulating real-world scenarios. The evaluations encompassed measures of user dissatisfaction across the entire virtual assistant system, particularly focusing on less common utterances (tail dissatisfaction), as well as the Semantic Error Rate (SemER) that evaluates intent and slot-filling performance. The study provides a comprehensive and nuanced understanding of how distilled models perform within the intricate landscape of virtual assistant systems. By addressing challenges related to model size reduction, evaluating performance across various metrics, and testing in real-world usage scenarios, our findings contribute to advancing language understanding technology and optimizing virtual assistant systems for enhanced user experiences.

### D. Challenges and Limitation

A prominent field of research and development is integrating Natural Language Processing (NLP) into virtual assistants since it has the potential to significantly improve these systems' functionality. The current approaches to NLP in virtual assistants, however, have drawbacks and limitations, just like any other technology. Here are some of these difficulties and restrictions:

*1) Challenges:* Natural language presents a significant problem due to its inherent ambiguity and reliance on context. Virtual assistants frequently have trouble understanding the complexities of context, which causes them to misread user requests. This restriction may make it more difficult to have natural discussions and give accurate responses. Another issue is support for several languages [25]. Virtual assistants powered by NLP must be proficient in a variety of languages, each with its own distinctive quirks. For users who speak uncommon languages or participate in multilingual conversations, some may fare very well in one language but fall short in others. For a flawless user experience, real-time processing is necessary. Nevertheless, NLP processing can be computationally demanding, making it difficult to provide immediate or close to real-time solutions [26]. This lag time may irritate users and lessen virtual assistants' general efficacy. Additionally, the difficulty of generalisation looms big. Many virtual assistants struggle to infer knowledge from certain user interactions. Instead, they might rely too heavily on pre-programmed reactions, which would make it harder for them to adjust to different user needs and would lessen their overall value. For virtual assistant technology to advance,

these problems must be solved. The improvement of NLP models, contextual understanding, multilingual support, real-time processing, implementation of strong privacy safeguards, and creation of more generalised and adaptable virtual assistants should be the main areas of research and development. These initiatives will assist NLP be more successfully and conveniently incorporated into virtual assistant technologies.

*2) Limitations:* Virtual assistants' dearth of common-sense reasoning is a key drawback. Their inability to participate in truly natural conversations is hampered by their frequent inability to comprehend fundamental, daily concepts and situations. When customers expect their virtual assistants to understand basic, contextual questions, this shortcoming might result in unsatisfactory and fragmented interactions. Another difficulty is managing lengthy talks. During lengthy conversations, context can be lost by existing NLP models, leading to responses that do not fit the general direction of the discourse. When consumers converse with virtual assistants in-depth or complex topics, this shortcoming may become especially apparent. Misinformation vulnerability is a serious issue, particularly in industries like news or healthcare. If virtual assistants don't have the tools to check the veracity of the information they provide, they might unintentionally spread harmful information or make false claims [27].

Virtual assistants might also have knowledge gaps in specific fields. When asked questions about specialized topics or industries, NLP models might not have current or in-depth knowledge, which results in answers that are incorrect or lacking. This restriction may limit the usefulness of virtual assistants in work-related or contexts requiring in-depth knowledge. Creating and maintaining NLP-driven virtual assistants can be prohibitively expensive and resource-intensive [28]. This restriction may make it difficult for smaller organizations and underserved communities to access this technology, potentially resulting in a digital divide. Another difficulty is the complexity and expense of integrating NLP into current systems. Because of this complexity, it may be challenging for businesses and organizations to adopt these technologies without major time and resource commitments. Continuous research and development efforts are crucial to overcoming these constraints and difficulties. This entails enhancing the capacity for common-sense reasoning, boosting the capacity for lengthy conversations, putting in place fact-checking procedures to thwart false information, honing domain-specific knowledge, and reducing response bias [29].

## VI. Conclusion and Future Prospect

This research offers a comprehensive approach to enhancing language understanding models within the realm of virtual assistant systems. Through a carefully structured series of experiments and in-depth analyses, we have substantiated the effectiveness of distilled models in achieving remarkable performance levels while upholding model efficiency. A noteworthy aspect of our findings lies in the recognition of the pivotal role played by meticulous evaluation metrics, particularly highlighting the superiority of metrics like mask-filling accuracy over conventional perplexity measurements. This innovation enhances the robustness and practical relevance of our evaluation methods in real-world applications. The results of our study also showcase the significant advantages of the distillation process. By carefully compressing large teacher models into intermediate-sized models, the successful transferred knowledge while overcoming challenges posed by model size reduction. Our distilled models consistently outperformed publicly available models, proving the efficacy of our approach in producing models that are not only more efficient but also more accurate in language understanding tasks. This has direct implications for the development of high-performing virtual assistant systems that can deliver prompt and accurate responses to user queries.

Looking forward to the research opens the door to exciting future prospects in the field of language understanding and virtual assistants. The success of our distillation process encourages further exploration into optimization techniques that can balance model size and performance. Additionally, the innovative evaluation metrics we introduced, such as mask-filling accuracy, offer new directions for evaluating model quality, which can be refined and extended in future studies. As virtual assistant systems continue to evolve, the insights from our research can guide the development of more efficient and effective models. The demonstrated techniques for distillation and evaluation provide a foundation for building even more advanced systems that can understand and respond to user input with increased accuracy and speed. As technology progresses, our work lays the groundwork for continuous improvements in virtual assistant capabilities, ultimately enhancing user experiences and interactions. The presented study contributes valuable knowledge to the field of language understanding in virtual assistant systems, offering practical solutions for optimizing model efficiency and performance. As the field continues to evolve, our research provides a strong stepping stone for future innovations that will shape the way virtual assistants understand and interact with users.

## References

[1] C. Lee, Y. Ko, and J. Seo, “A Simultaneous Recognition Framework for the Spoken Language Understanding Module of Intelligent Personal Assistant Software on Smart Phones,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China: Association for Computational Linguistics, 2015, pp. 818–822. doi: 10.3115/v1/P15-2134.

[2] A. P. Sam, B. Singh, and A. S. Das, “A Robust Methodology for Building an Artificial Intelligent (AI) Virtual Assistant for Payment Processing,” in 2019 IEEE Technology & Engineering Management Conference (TEMSCON), Atlanta, GA, USA: IEEE, Jun. 2019, pp. 1–6. doi: 10.1109/TEMSCON.2019.8813584.

[3] J. Luketina et al., “A Survey of Reinforcement Learning Informed by Natural Language,” 2019, doi: 10.48550/ARXIV.1906.03926.

[4] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, and G. Campbell, “BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics,” Int J Artif Intell Educ, vol. 24, no. 3, pp. 284–332, Sep. 2014, doi: 10.1007/s40593-014-0017-9.

[5] H. Mahmoudi, S. Camboim, and M. A. Brovelli, “Development of a Voice Virtual Assistant for the Geospatial Data Visualization

Application on the Web," Computer Science and Mathematics, preprint, Jul. 2023. doi: 10.20944/preprints202307.0413.v1.

[6] G. Dizon, "Evaluating intelligent personal assistants for L2 listening and speaking development," Feb. 2020, Accessed: Aug. 10, 2023. [Online]. Available: http://hdl.handle.net/10125/44705

[7] A. Mishakova, F. Portet, T. Desot, and M. Vacher, "Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes," in 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kyoto, Japan: IEEE, Mar. 2019, pp. 832–837. doi: 10.1109/PERCOMW.2019.8730721.

[8] E. Elshan, P. Ebel, M. Söllner, and J. M. Leimeister, "Leveraging Low Code Development of Smart Personal Assistants: An Integrated Design Approach with the SPADE Method," Journal of Management Information Systems, vol. 40, no. 1, pp. 96–129, Jan. 2023, doi: 10.1080/07421222.2023.2172776.

[9] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and Natural Language Processing of social media data for event detection in smart cities," Sustainable Cities and Society, vol. 85, p. 104026, Oct. 2022, doi: 10.1016/j.scs.2022.104026.

[10] Z. Yang, L. Shou, M. Gong, W. Lin, and D. Jiang, "Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System," in Proceedings of the 13th International Conference on Web Search and Data Mining, Houston TX USA: ACM, Jan. 2020, pp. 690–698. doi: 10.1145/3336191.3371792.

[11] J. Mariani, S. Rosset, M. Garnier-Rizet, and L. Devillers, Eds., Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice. New York, NY: Springer New York, 2014. doi: 10.1007/978-1-4614-8280-2.

[12] J. R. Bellegarda, "Spoken Language Understanding for Natural Interaction: The Siri Experience," in Natural Interaction with Robots, Knowbots and Smartphones, J. Mariani, S. Rosset, M. Garnier-Rizet, and L. Devillers, Eds., New York, NY: Springer New York, 2014, pp. 3–14. doi: 10.1007/978-1-4614-8280-2_1.

[13] C. Lee and Y. Ko, "Spoken Language Understanding with a Novel Simultaneous Recognition Technique for Intelligent Personal Assistant Software," Int. J. Artif. Intell. Tools, vol. 27, no. 03, p. 1850009, May 2018, doi: 10.1142/S0218213018500094.

[14] W. Hariri, "Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing," 2023, doi: 10.48550/ARXIV.2304.02017.

[15] A. Ait-Mlouk and L. Jiang, "KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data," IEEE Access, vol. 8, pp. 149220–149230, 2020, doi: 10.1109/ACCESS.2020.3016142.

[16] J. Jungbluth, K. Siedentopp, R. Krieger, W. Gerke, and P. Plapper, "Combining Virtual and Robot Assistants—A Case Study about Integrating Amazon's Alexa as a Voice Interface in Robotics," 2018.

[17] E. C. Alagha and R. R. Helbing, "Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri," BMJ Health Care Inform, vol. 26, no. 1, p. e100075, Nov. 2019, doi: 10.1136/bmjhci-2019-100075.

[18] W. Villegas-Ch, J. García-Ortiz, K. Mullo-Ca, S. Sánchez-Viteri, and M. Roman-Cañizares, "Implementation of a Virtual Assistant for the Academic Management of a University with the Use of Artificial Intelligence," Future Internet, vol. 13, no. 4, p. 97, Apr. 2021, doi: 10.3390/fi13040097.

[19] L. Dong et al., "Unified Language Model Pre-training for Natural Language Understanding and Generation," 2019.

[20] M. Syromiatnikov and V. Ruvinskaya, "Natural Language Processing for Intelligent Virtual Assistant System," 2020.

[21] J. FitzGerald et al., "Alexa Teacher Model: Pretraining and Distilling Multi-Billion-Parameter Encoders for Natural Language Understanding Systems," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC USA: ACM, Aug. 2022, pp. 2893–2902. doi: 10.1145/3534678.3539173.

[22] M. U. Hadi et al., "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects." TechRxiv, Sep. 21, 2023. doi: 10.36227/techrxiv.23589741.v3.

[23] J. FitzGerald et al., "Alexa Teacher Model: Pretraining and Distilling Multi-Billion-Parameter Encoders for Natural Language Understanding Systems," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, in KDD '22. New York, NY, USA: Association for Computing Machinery, Aug. 2022, pp. 2893–2902. doi: 10.1145/3534678.3539173.

[24] H. B. Essel, D. Vlachopoulos, A. Tachie-Menson, E. E. Johnson, and P. K. Baah, "The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education," International Journal of Educational Technology in Higher Education, vol. 19, no. 1, p. 57, Nov. 2022, doi: 10.1186/s41239-022-00362-6.

[25] K. Affolter, K. Stockinger, and A. Bernstein, "A comparative survey of recent natural language interfaces for databases," The VLDB Journal, vol. 28, no. 5, pp. 793–819, Oct. 2019, doi: 10.1007/s00778-019-00567-8.

[26] M. Hagiwara, Real-World Natural Language Processing: Practical applications with deep learning. Simon and Schuster, 2021.

[27] J. Anderson and L. Rainie, "The Future of Truth and Misinformation Online," Pew Research Center: Internet, Science & Tech. Accessed: Oct. 25, 2023. [Online]. Available: https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

[28] A. Piñeiro-Martín, C. García-Mateo, L. Docío-Fernández, and M. del C. López-Pérez, "Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models," Electronics, vol. 12, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/electronics12143170.

[29] Read "How People Learn II: Learners, Contexts, and Cultures" at NAP.edu. doi: 10.17226/24783.