

Overview of the progression of state-of-the-art language models

Asmae Briouya, Hasnae Briouya, Ali Choukri

Laboratory of Computer Sciences, Faculty of Sciences Kenitra, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Dec 22, 2023

Revised Jan 7, 2024

Accepted Jan 15, 2024

Keywords:

Artificial intelligence

BERT model

Generative pre-trained transformer

Machine learning

Question-answering

ABSTRACT

This review provides a concise overview of key transformer-based language models, including bidirectional encoder representations from transformers (BERT), generative pre-trained transformer 3 (GPT-3), robustly optimized BERT pretraining approach (RoBERTa), a lite BERT (ALBERT), text-to-text transfer transformer (T5), generative pre-trained transformer 4 (GPT-4), and extra large neural network (XLNet). These models have significantly advanced natural language processing (NLP) capabilities, each bringing unique contributions to the field. We delve into BERT's bidirectional context understanding, GPT-3's versatility with 175 billion parameters, and RoBERTa's optimization of BERT. ALBERT emphasizes model efficiency, T5 introduces a text-to-text framework, and GPT-4, with 170 trillion parameters, excels in multimodal tasks. Safety considerations are highlighted, especially in GPT-4. Additionally, XLNet's permutation-based training achieves bidirectional context understanding. The motivations, advancements, and challenges of these models are explored, offering insights into the evolving landscape of large-scale language models.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Asmae Briouya

Laboratory of Computer Sciences, Faculty of Sciences Kenitra, Ibn Tofail University

Kenitra, Morocco

Email: asmae.briouya@uit.ac.ma

1. INTRODUCTION

Recent years have seen remarkable advancements in natural language processing (NLP) with models like bidirectional encoder representations from transformers (BERT) and the generative pre-trained transformer (GPT) series. BERT, introduced by Google in 2018, revolutionized NLP with its bidirectional approach, while the GPT series, evolving from GPT-1 to refined models like InstructGPT and the latest GPT-4, demonstrated significant scale and capabilities. BERT excels in tasks like text classification and question answering [1] influencing subsequent models. The GPT series, known for its generative capabilities, includes InstructGPT, fine-tuned for user alignment. Both BERT and GPT models face challenges in computational demands. Newer models like XLNet, RoBERTa, ALBERT, and T5 push boundaries with innovative approaches. GPT-4, the latest in the series, boasts multimodal capabilities and fine-tuning with reinforcement learning from human feedback (RLHF), achieving human-level proficiency. In conclusion, these models collectively shape the dynamic landscape of NLP, influencing how computers comprehend and generate human language. The exploration of diverse transformer-based language models, including BERT [2], GPT-3 [3], XLNet [4], RoBERTa [5], ALBERT [6], T5 [7], and the groundbreaking GPT-4 [8], is motivated by a multifaceted drive encompassing advancements in NLP, model efficiency, task-specific optimization, and the pursuit of safer and more

responsible AI applications. Models like BERT and RoBERTa have revolutionized language understanding by leveraging bidirectional context, contrasting with unidirectional models and advancing the field significantly. ALBERT stands out for its focus on reducing the number of parameters while maintaining high efficiency and performance, a crucial consideration in resource-limited scenarios. T5 introduces a unique text-to-text framework, treating all NLP problems as variations of a single text conversion task, simplifying the processing pipeline. GPT-4 marks a notable development with its expanded scale and multimodal capabilities, handling both text and image inputs. Its training phase incorporates rule-based reward models, exploring the integration of vision and language. Understanding GPT-4's advancements sheds light on the complexities and potential of multimodal AI systems, including safety features and ethical implications. XLNet's introduction of a permutation-based training approach combines bidirectional context understanding with autoregressive models effectively. InstructGPT, developed through the fine-tuning of GPT-3, focuses on aligning language models more closely with user intent, emphasizing user preference and training efficiency. These models represent various advancements in language understanding and model development [9].

2. METHOD

Several state-of-the-art language models have significantly influenced the landscape of NLP. BERT introduced bidirectional context understanding by predicting masked words in a sentence, revolutionizing contextual language modeling. GPT-3 stands out with its massive 175 billion parameters, showcasing unparalleled versatility across various language tasks. RoBERTa, an optimized version of BERT, improves performance through larger batch sizes and extended training times. ALBERT, emphasizing efficiency, achieves competitive results with fewer parameters. T5, adopting a text-to-text framework, demonstrates robust performance and efficient multi-task learning. XLNet innovatively combines bidirectional context understanding with an autoregressive model, achieving state-of-the-art results on the GLUE benchmark. InstructGPT, achieved through fine-tuning, introduces motivation for aligning language models more closely with user intent [9]. GPT-4, the latest iteration, introduces multimodal capabilities, processing both text and image inputs, with a colossal 170 trillion parameters. These models collectively represent the forefront of language understanding, each contributing unique advancements and addressing specific challenges in the NLP domain.

2.1. Bidirectional encoder representations from transformers

BERT, which stands for bidirectional encoder representations from transformers [10], is a groundbreaking NLP technique introduced by researchers at Google, including Jacob Devlin, in 2018. BERT is built upon the transformer architecture, a neural network architecture designed for processing sequential data. Unlike earlier NLP models that processed text input in a unidirectional manner, BERT is distinguished by its bidirectional approach to understanding word context. The key innovation lies in its use of a "masked language model" (MLM) pre-training objective. In the pre-training phase, BERT processes input text by randomly masking out some of the words. The model is then trained to predict the masked words based on the context provided by the surrounding words in both directions, i.e., to the left and right of the masked word. This bidirectional context understanding allows BERT to capture the nuances and meanings of words in a more comprehensive way. BERT's bidirectionality is achieved through two main pre-training tasks: masked language modeling and next sentence prediction. In the masked language modeling task, BERT predicts the masked words in a sentence, while in the next sentence prediction task, the model learns to predict whether a given pair of sentences are consecutive or not. This bidirectional contextual understanding enables BERT to excel in various NLP tasks, including text classification, named entity recognition, question answering, and more. BERT has significantly advanced the state of the art in NLP, and its impact extends to inspiring subsequent models and improvements in the broader field of language representation models. In summary, BERT is a transformer-based NLP model that, through bidirectional context understanding via masked language modeling, has revolutionized how computers comprehend and process human language.

2.2. GPT series

The GPT series, developed by OpenAI, represents a series of revolutionary steps in the field of NLP. These models, based on the innovative transformer architecture, have not only grown in size and complexity with each iteration but have also significantly expanded the scope and capabilities of NLP applications, they have become pivotal tools in various domains such as text generation, translation, and sentiment analysis, revolutionizing how we interact with language in the digital age.

2.2.1. GPT-1 (2018)

Debuting in 2018, GPT-1 emerged as a pioneering force in the realm of pre-trained language models, marking the inception of a new era in NLP. Its innovative approach and architecture represented a significant departure from previous models, establishing a foundational shift in how language models were developed and utilized [11]. Equipped with 117 million parameters, GPT-1 stood as a groundbreaking development in the field. It demonstrated remarkable proficiency in tasks such as text completion and machine translation, showcasing the vast potential of large-scale language models in understanding and generating text that closely resembles human language [12]. The success of GPT-1 was instrumental in paving the way for subsequent, more advanced models in the GPT series. It set a new benchmark in the field, both in terms of technical capability and the potential applications of such models. The strides made by GPT-1 laid the groundwork for ongoing advancements in NLP, driving forward the development of more sophisticated and powerful language models.

2.2.2. GPT-2 (2019)

Introduced in 2019, GPT-2 marked a significant step forward in the GPT series, boasting an impressive model size of 1.5 billion parameters, a substantial increase from its predecessor [12]. This expansion in scale enabled GPT-2 to generate text that was not only coherent but also contextually relevant, pushing the envelope of what was achievable in automated text generation. However, the advanced capabilities of GPT-2 led to a cautious approach by OpenAI. Concerned about the potential misuse of the technology, especially in the generation of misleading or fake content, access to the full model was initially restricted [13]. This decision underscored the importance of ethical considerations in the development and deployment of powerful AI tools.

The release of GPT-2 and the discussions surrounding it brought to the forefront the ethical implications of such advanced NLP tools. It highlighted the need for responsible development and deployment of AI, ensuring that these technologies are used in ways that are beneficial and do not pose risks to society. This conversation was crucial in shaping the future directions of AI research and development, emphasizing the balance between innovation and responsible usage.

2.2.3. GPT-3 (2020)

The release of GPT-3 in 2020 represented a landmark achievement in the GPT series. This model, equipped with an astonishing 175 billion parameters, significantly outperformed its predecessors in both scale and capabilities [14]. The sheer size and sophistication of GPT-3 set a new benchmark in the field of NLP.

GPT-3's prowess extended across a wide range of language tasks. It demonstrated exceptional skills not only in translation and question answering but also in creative writing, among other tasks. This level of versatility was unprecedented, establishing GPT-3 as a highly adaptable and multifaceted tool in the NLP.

A notable advancement with GPT-3 was the introduction of few-shot and zero-shot learning capabilities. This feature enabled the model to effectively handle various tasks with minimal or even no specific training, showcasing its remarkable ability to adapt and generalize across different language applications. The launch of GPT-3 also reignited discussions about the ethical implications of advanced AI technologies. It highlighted the growing need to focus on the safe and responsible development of AI systems. These conversations were crucial in understanding the broader impacts of such powerful technologies and ensuring their beneficial use in society.

2.2.4. GPT-Neo

GPT-Neo emerged as a community-driven initiative, responding to the increasing demand for open-source and more accessible versions of large-scale language models akin to the GPT series [15]. This development was particularly significant as it opened up the advanced capabilities of such models to a broader audience. Although these models are scaled down in comparison to the official GPT versions, they are designed to deliver comparable performance. This aspect of GPT-Neo is especially beneficial for researchers and developers who seek a balance between computational efficiency and the effectiveness of the model [16].

By offering a more accessible version of cutting-edge NLP technology, GPT-Neo represents a major stride towards democratizing the use of advanced language processing tools. It enables a wider spectrum of users, from academic researchers to independent developers, to experiment and innovate with these potent AI tools. This initiative plays a crucial role in fostering a more inclusive environment in the field of AI, allowing for greater participation and collaboration in the development and application of these technologies.

2.2.5. GPT-4 (2023)

Unveiled in 2023, GPT-4 represents a monumental advancement in the GPT series, showcasing OpenAI's unwavering commitment to advancing the frontiers of AI language models. This latest iteration marks a significant leap forward from its predecessors, particularly in its ability to handle a broader range of inputs. GPT-4 introduces multimodal capabilities, a notable evolution from the earlier models that were solely text-based. This advancement allows it to process and understand both text and image inputs, greatly expanding the range of applications and functionalities of the model.

The architecture and training methodologies of GPT-4 have undergone significant refinement, building on the established strengths of the Transformer model. This model has been subjected to rigorous evaluation across a diverse array of academic and professional benchmarks, where it has demonstrated its versatility and robustness in handling complex language tasks. While the exact size of GPT-4 has not been publicly disclosed, it is understood to be substantially larger than GPT-3. This increase in scale is believed to enhance its capacity for processing and generating language, enabling it to tackle more complex and nuanced tasks.

In addition to its enhanced capabilities, GPT-4 has incorporated significant improvements in the realm of AI safety. These include advancements in steerability, which allow for more precise control over the model's outputs, and a refined refusal behavior designed to mitigate risks associated with the generation of inappropriate or harmful content. Despite these significant advancements, GPT-4 is not without its limitations. It continues to undergo refinement, as it occasionally exhibits inaccuracies and a tendency to provide cautious or hedged responses, particularly in complex or ambiguous scenarios. This ongoing development reflects the continuous effort to improve the model, ensuring it remains a reliable and effective tool in the ever-evolving landscape of artificial intelligence.

3. RESULT AND DISCUSSION

3.1. GPT vs BERT: bidirectional vs generative approaches

BERT and GPT [17] two influential language models, differ fundamentally in their approaches to language modeling. BERT is a bidirectional model trained on a large corpus to understand context from both left and right sides of a word. It excels in fine-tuning on tasks like question answering and sentiment analysis [18]. In contrast, GPT, or generative pre-trained transformer, is an autoregressive language generation model that generates text based on a given prompt without task-specific fine-tuning. GPT's strength lies in its ability to produce coherent and contextually relevant text for diverse applications, such as summarization [19] and dialogue generation. Both models have achieved state-of-the-art results in their respective domains, showcasing the impact of their unique approaches on NLP tasks as shown in Table 1.

3.2. RoBERTa vs ALBERT: optimized versions of BERT with efficiency emphasis

Before delving into the detailed comparison table, let's summarize the key differences between RoBERTa and ALBERT. RoBERTa, which stands for robustly optimized BERT pretraining approach, is an enhanced version of BERT, focusing on optimizing the training process. It achieves this by eliminating the next sentence prediction task and introducing new masking techniques, while also utilizing a wider range of training data and a larger vocabulary size. ALBERT, on the other hand, is a more efficient variant of BERT, characterized by its use of cross-layer parameter sharing and a reduced model size. It also introduces sentence-order prediction to better handle multi-sentence encoding tasks. These modifications result in ALBERT having far fewer parameters compared to BERT-large, making it more efficient yet still powerful. Overall, while both models build upon the foundations of BERT, they each introduce distinct improvements and optimizations, catering to different aspects of NLP tasks [20].

The Table 2 provides a comprehensive comparison between the RoBERTa and ALBERT models, detailing various aspects such as their names, approaches, training procedures, data sources, and unique features. It contrasts RoBERTa's emphasis on optimized BERT pretraining and use of large datasets with ALBERT's focus on efficiency and parameter reduction. The table also highlights differences in their tokenization methods, parameter sharing strategies, model sizes, and their respective objectives and applications in NLP tasks.

Table 1. This table summarizes the differences in approach, advantages, and applications between BERT and GPT in the context of language modeling

Aspect	BERT	GPT
Model type	BERT	GPT
Training approach	Pre-trained in an unsupervised manner and fine-tuned on specific tasks	Pre-trained in an unsupervised manner for language generation without fine-tuning
Architecture	Transformer with bidirectional context	Transformer with a decoder-only setup for autoregressive language generation
Pre-training objective	MLM for bidirectional context understanding	Autoregressive language modeling to predict the next token in a sequence
Context consideration	Captures context from both left and right sides of a word	Considers the entire context, including preceding and following words
Application focus	Fine-tuning on specific tasks such as question answering, sentiment analysis	Generating coherent and contextually relevant text for various applications
Task examples	Question answering, sentiment analysis, named entity recognition	Text completion, summarization, dialogue generation, story generation
Advantages	Bidirectional context, fine-tuning capability, versatility	Language generation, flexibility, state-of-the-art performance
Performance achievements	State-of-the-art results in NLP tasks like question answering and sentiment analysis	State-of-the-art results in language generation tasks like story generation and translation
Versatility	Suitable for various NLP tasks	Versatile for a wide range of language generation applications
Contextual understanding	Precise understanding of context due to bidirectional approach	Captures context with a focus on autoregressive language generation

Table 2. Comparison of RoBERTa and ALBERT

Aspect	RoBERTa	ALBERT
Model name	Robustly optimized BERT pretraining approach	A lite BERT
Approach and optimization	Optimizes BERT with adjustments in fine-tuning, data, and input handling	A "lite" version of BERT, focusing on efficient parameter
Training procedure	Removes the next sentence prediction task; Introduces static and dynamic masking	Employs cross-layer parameter sharing; uses sentence-order prediction (SOP) loss
Data sources	Large datasets: 16 GB of books corpus and English Wikipedia, common crawl news, web text corpus, stories from common crawl	Similar to BERT but with a focus on efficiency and fewer parameters
Tokenization	Byte-level byte-pair encoding (BPE) with 50,000 subword units	Utilizes BERT's tokenization method with optimizations for parameter efficiency
Unique features	Larger batch training without NSP objective; enhanced pre-training with diverse datasets	Significantly fewer parameters than BERT-large; factorized embedding parameterization
Parameter sharing strategy	Standard parameter usage as in original BERT	Extensive use of cross-layer parameter sharing to reduce model size
Model size and efficiency	Larger model size compared to BERT, due to extended training data and procedure	Much smaller in size due to parameter sharing, about 18 times fewer parameters than BERT-large
Objective and application	Focused on improving training efficiency and accuracy	Designed for tasks requiring efficient multi-sentence encoding and smaller model sizes

3.3. GPT-2 vs GPT-3: Evolution within the GPT series

The primary distinctions between GPT-2 and GPT-3 are centered on their scale, model size, the breadth and diversity of their training data, and the overall scope of their capabilities. GPT-3 stands as a significant progression from GPT-2, highlighting advancements in multiple critical areas, which are outlined as:

- Model size and parameters: GPT-2, released with 1.5 billion parameters, was a considerable advancement in the field of language models at its time of introduction [21]. In stark contrast, GPT-3 represents a monumental leap forward with its massive 175 billion parameters, solidifying its position as one of the most advanced and sophisticated language models ever developed. This dramatic increase in parameters underpins GPT-3's enhanced processing power and its ability to handle more complex language patterns.
- Training data and scale: the training data used for both models encompass a wide range of internet text, but GPT-3's dataset is far more extensive and diverse. This larger and more varied training corpus significantly

bolsters GPT-3's ability to understand and respond to a vast spectrum of topics and linguistic styles, making it more versatile and effective in different contexts.

- Capabilities and applications: while GPT-2 was proficient in various NLP tasks like text generation and summarization, GPT-3 takes these capabilities to new heights. It excels in a broader array of tasks, including advanced language understanding, context-based reasoning, and sophisticated problem-solving. This expansion of capabilities demonstrates GPT-3's adaptability and effectiveness in a wide range of applications, from simple text generation to complex decision-making scenarios.
- Fine-tuning and adaptability: GPT-2 provided the flexibility to be fine-tuned for specific tasks, offering customization based on the task at hand. However, GPT-3, with its immense scale and comprehensive training, showcases remarkable adaptability, often requiring little to no fine-tuning to excel in a wide variety of tasks. This feature of GPT-3 makes it an incredibly versatile tool for developers and researchers, as it can be applied to numerous tasks without the need for extensive retraining or adjustment.
- Few-shot and zero-shot learning: among the most impressive features of GPT-3 is its few-shot learning ability. This capability allows GPT-3 to perform effectively with minimal training examples, often surpassing models specifically fine-tuned for those tasks. Moreover, its zero-shot learning abilities enable GPT-3 to tackle new tasks it wasn't explicitly trained for, using just natural language prompts. This aspect of GPT-3 underscores its potential to revolutionize how AI models are trained and deployed, offering efficiency and flexibility in learning new tasks.
- Language generation: in terms of language generation, the output of GPT-3 is markedly more coherent and contextually relevant compared to GPT-2. The improvement can be attributed to GPT-3's larger model size and the extensive and diverse training it has undergone. This allows GPT-3 to better understand and process the nuances of language, resulting in outputs that are more refined, accurate, and context-aware. This enhancement in language generation makes GPT-3 an invaluable tool for a wide range of applications, from automated content creation to interactive conversational agents.

To provide a clearer and more comprehensive understanding of the differences and advancements from GPT-2 to GPT-3, I will include both a table and a figure in the following sections. The Table 3 is designed to offer a detailed, side-by-side comparison of these two models. It will cover various critical aspects such as the model size, the extent and nature of the training data, the specific capabilities each model possesses, and the learning techniques employed.

Additionally, a Figure 1 will accompany the table to provide a graphical representation of these differences and improvements. This visual aid aims to simplify the complex information, making it more accessible and easier to comprehend at a glance. It will serve as an effective tool for visually capturing the scale of advancement from GPT-2 to GPT-3, highlighting the key areas where GPT-3 has pushed the boundaries of what's possible in language modeling.

Table 3. Comparison of GPT-2 and GPT-3

Aspect	GPT-2	GPT-3
Model size and parameters	Large language model with 1.5 billion parameters.	Significantly larger with 175 billion parameters.
Training data and scale	Trained on diverse internet text data, including news articles, websites, and other sources.	Trained on an even larger and more diverse dataset, covering a wide array of subjects.
Capabilities and applications	Strong performance in text generation, translation, summarization, and more.	Advanced capabilities in language understanding and generation, context-based reasoning, and problem-solving.
Fine-tuning and adaptability	Effective fine-tuning on specific tasks, adaptable to various applications.	Generalizes across a wide range of tasks and domains without extensive fine-tuning.
Few-shot learning	–	Demonstrates ability to perform well on tasks with few examples.
Zero-shot learning	–	Capable of performing on tasks without explicit training, using natural language prompts.
Language generation	Generates coherent text.	Generates more coherent and contextually relevant text compared to GPT-2.

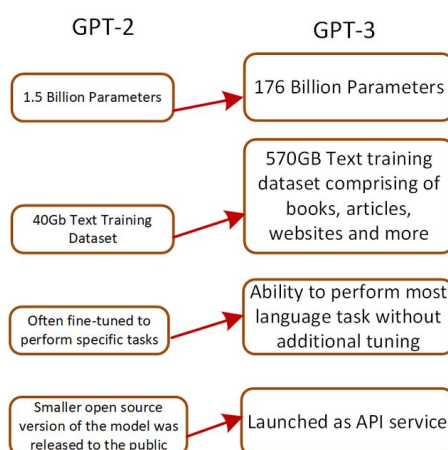


Figure 1. Comparison of GPT-2 and GPT-3 [22]

3.4. GPT-3 vs InstructGPT: generative vs fine-tuned approach

This Table 4 provides a detailed comparison between GPT-3 and InstructGPT, focusing on several critical aspects such as their training methodologies, alignment with user preferences, adaptability, resource efficiency, and the overall quality of outputs, particularly in terms of truthfulness, reliability, and control. The table clearly outlines the distinctions between the two models, emphasizing InstructGPT's advancements, especially in aligning more closely with user intent and operating with greater efficiency. Overall, this table not only contrasts the technical specifications of GPT-3 and InstructGPT but also provides insights into the practical implications of these differences, offering a comprehensive view of their strengths and limitations in various applications.

Table 4. Comparison of GPT-3 and InstructGPT

Aspect	GPT-3	InstructGPT
Model parameters	175 billion parameters	1.3 billion parameters (100x fewer than GPT-3)
Training Method	Trained on diverse internet data	Fine-tuned from GPT-3 with human feedback as illustrated in (Figure 2)
User preference	Less preferred in evaluations	Preferred over GPT-3 in $85 \pm 3\%$ of cases
Truthfulness	Lower performance on TruthfulQA benchmark	Generates truthful answers about twice as often as GPT-3
Hallucination rate in closed-domain tasks	41% hallucination rate	21% hallucination rate (about half of GPT-3's rate)
Toxicity in outputs	Higher rate of toxic outputs	Generates 25% fewer toxic outputs when prompted to be respectful
Generalization capabilities	Generalizes well but may require specific prompts	Shows promising generalization to non-English language tasks and code-related tasks
Performance in customer assistant context	–	More appropriate and reliable in customer assistant roles, better adherence to instructions
Performance regressions (Alignment tax)	Consistent performance on public NLP datasets	Performance regressions on datasets like SQuAD, DROP, HellaSwag, WMT 2015 (French to English translation)

The development of InstructGPT, an AI model with 1.3 billion parameters, presents a notable advancement in the field of artificial intelligence, especially when compared to its predecessor, GPT-3, which boasts 175 billion parameters. Despite its significantly smaller size, InstructGPT demonstrates several key enhancements, making it a remarkable achievement in AI technology. The improvements are manifold:

- User preference: InstructGPT has received notably more favorable feedback from users compared to GPT-3. This increased preference is attributed to its enhanced capacity to comprehend and respond to user instructions more accurately, resulting in a more user-friendly experience. The model's improved interaction dynamics make it particularly suitable for applications requiring high levels of user engagement and satisfaction.

- Truthfulness: a key advancement of InstructGPT is its ability to generate responses that are not only more truthful but also more informative. This improvement is especially significant in benchmark tasks where the accuracy and reliability of information are paramount. The model's enhanced truthfulness is vital in applications where decision-making depends on the integrity and quality of information provided.
- Reduced hallucination: InstructGPT exhibits a substantial reduction in generating incorrect or misleading information, particularly in closed-domain tasks. This decrease in erroneous outputs, often referred to as 'hallucinations', enhances the model's reliability in providing accurate and trustworthy information. This attribute is essential in scenarios where factual accuracy is crucial, such as in educational or informational contexts.
- Lower toxicity: InstructGPT is designed to produce fewer toxic outputs, an important feature for maintaining a respectful and positive interaction tone. This characteristic is crucial when the AI is tasked with ensuring safe and constructive user interactions, particularly in public-facing or sensitive communication scenarios. The focus on reducing toxicity is a step towards more ethical and user-centric AI development.
- Better generalization: InstructGPT shows exceptional adaptability across a diverse range of tasks, including challenges in non-English languages and coding-related tasks. Its ability to perform effectively across various domains highlights its versatility and broad applicability. This generalization makes InstructGPT a valuable tool in multilingual and interdisciplinary applications.
- Customer assistant performance: in customer service scenarios, InstructGPT surpasses its predecessor in performance. Its ability to adhere closely to instructions enhances its effectiveness and reliability in customer support roles. This improvement is particularly beneficial in industries where accurate and responsive customer service is crucial for client satisfaction.
- Performance trade-offs: while InstructGPT boasts numerous advantages, it's important to recognize some performance compromises. The model sometimes shows reduced performance in certain public NLP datasets, a trade-off resulting from its alignment with human feedback. This balance between general performance and specialized, human-centric improvements is key to the model's design and application strategy.

In conclusion, InstructGPT stands out not only for its alignment with user preferences but also for its improved performance in truthfulness, reliability, and a range of other areas. This achievement is largely attributed to its fine-tuning process, which incorporates human feedback, ensuring that the model better meets the needs and expectations of its users. The development of InstructGPT marks a significant step forward in the realm of AI, showcasing the potential for smaller, more focused models to achieve high levels of performance and user satisfaction [23].

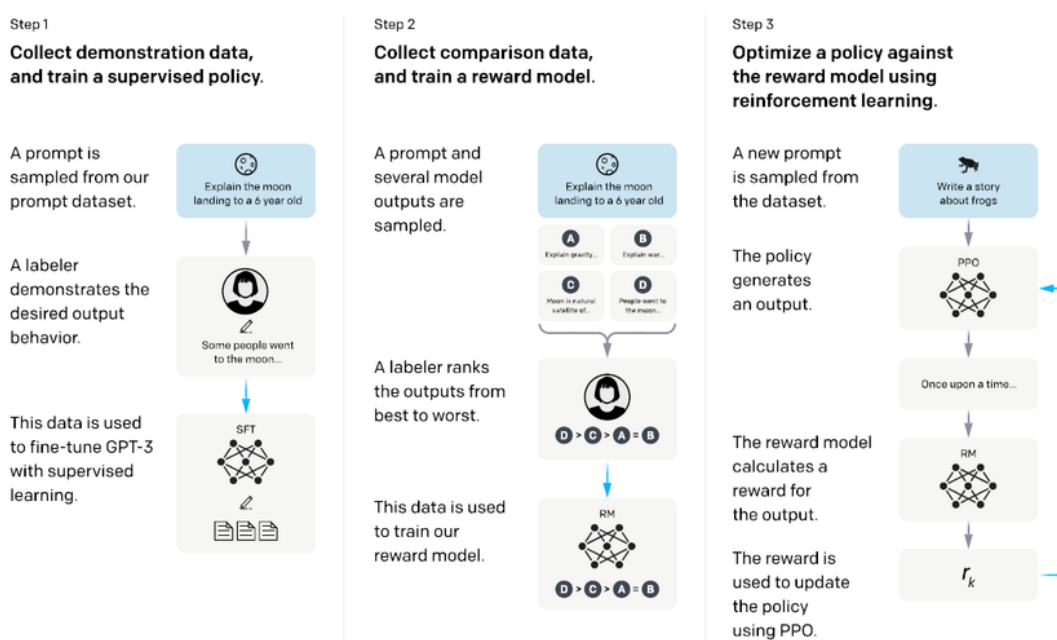


Figure 2. Three-step training process of InstructGPT: Demonstration, comparison, and optimization [9]

3.5. XLNet vs T5: permutation-based vs text-to-text framework

In the realm of NLP, two advanced transformer-based models, XLNet and T5, stand out for their unique approaches. XLNet utilizes a permutation-based training strategy, which allows it to predict the ordering of words within a sequence, thus enabling a deeper understanding of context and improving performance on tasks such as text completion and sentiment analysis. This model, referenced in (Figure 3), benefits from a diverse training background including Wikipedia and BooksCorpus. In contrast, T5, cited in (Figure 4), operates on a text-to-text [24] basis, treating every NLP task as a text generation problem, where inputs are transformed into outputs, excelling in translation and summarization due to its training on the C4 dataset.

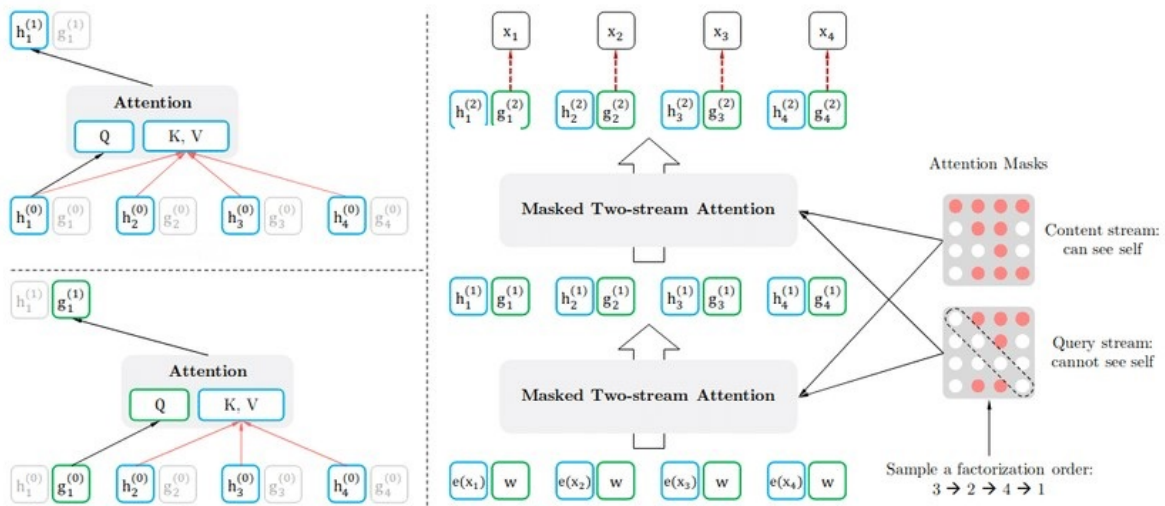


Figure 3. Xlnet architecture [4]

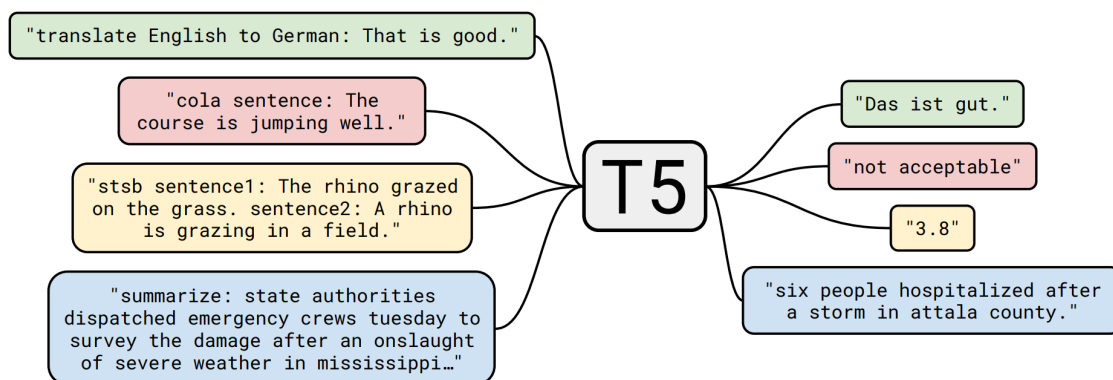


Figure 4. The text-to-text framework used by T5 [7]

This Table 5 offers a cursory look into the functionalities and capabilities of T5 and XLNet. It presents a side-by-side comparison that highlights their methodological differences and how these variations impact their performance in various NLP tasks. The table aims to provide a straightforward overview rather than an in-depth analysis, making it a useful reference for quickly grasping the fundamental distinctions between these two models. Such a comparative view is instrumental in understanding the basic operational frameworks of T5 and XLNet, offering insights into their respective strengths and limitations in handling NLP challenges.

Table 5. Simplified comparison of T5 and XLNet

Feature	T5	XLNet
Training objective	Trained on tasks like language modeling, translation, summarization	Maximizes likelihood over all permutations of the input sequence [25]
Training data	Colossal cleaned crawled corpus (C4)	Combines sources like Wikipedia, BooksCorpus, Giga5
Architecture	Encoder-decoder transformer with separate components for input and output	Uses segment-level recurrence to capture bidirectional context
Performance	Excels in diverse NLP tasks including question answering and summarization	Strong in understanding context, useful in sentiment analysis, and natural language inference

3.6. GPT-3 vs GPT-4: comparison within the GPT series

The evolution from GPT-3 to GPT-4 marks significant advancements in NLP models. While GPT-3 made waves with 175 billion parameters, GPT-4 has taken a colossal leap to 170 trillion parameters. This increase facilitates a greater understanding of context, allows for more accurate and relevant responses, and expands the model's capabilities. One of the notable upgrades in GPT-4 is the expanded context window length, increasing from 2048 tokens in GPT-3.5 to up to 32768 tokens [26], depending on the version. This enhancement greatly improves the model's ability to process and generate longer passages of text. Furthermore, GPT-4 introduces multimodality, supporting inputs that include both text and images, unlike its predecessors, which processed text only. The training process also sees refinement with the incorporation of a rule-based reward model (RBRM) alongside the reinforcement learning with human feedback (RLHF) used in GPT-3.5. In terms of output, GPT-4 can generate up to 24000 words—equivalent to 48 pages—which is a substantial increase from the 3000-word limit of GPT-3.5. In benchmarking performance, GPT-4 showcases impressive results in professional and academic assessments, reflecting its robust capabilities in specialized domains such as the legal field and across various languages. For a more detailed comparison of these models, please refer to the enhanced Table 6 provided which encapsulates the key features and evolutions of GPT-3, GPT-3.5, and GPT-4. This table presents an organized view of the progression in model parameters, training sophistication, and the broadening scope of applications and technological integrations.

Table 6. Enhanced comparative analysis of GPT-3, GPT-3.5, and GPT-4 language models

Feature	GPT-3	GPT-3.5	GPT-4
Model size (parameters)	175 billion	175 billion	170 trillion
Context window length	2048 tokens	2048 tokens	Up to 32768 tokens, depending on the version
Modality	Text-only inputs	Text-only inputs	Multimodal inputs (text and images)
Training data	Diverse internet text up to 2021	More recent internet text than GPT-3	Most extensive, up-to-date internet text up to early 2023
Training process	RLHF	RLHF with improvements	RLHF and RBRM approach
Output length	Up to a few thousand tokens	Constrained by 3000 words (approx. 6 pages)	Can generate up to 24000 words (approx. 48 pages)
Performance	Advanced for its time	Incremental improvements	Notably higher in accuracy, relevance, and extended context
Benchmarking performance	–	–	Significantly improved performance in professional, academic, and multilingual evaluations
Multimodality	–	–	Supports text and image inputs, broadening application scope
Performance in professional domains	–	–	Exhibits impressive capabilities in legal and other professional domains
Multilingual capabilities	Supports multiple languages	Improved multilingual performance	Advanced proficiency in a wide range of languages
Technology integration	Wide integration into apps and services	Further integration improvements	Comprehensive integration capabilities for complex systems

3.7. Panoramic model comparison

This panoramic comparison serves as an overarching summary of the key distinctions and evolutionary trajectories discussed earlier across various groundbreaking language models. It encapsulates the contrasts and

developments from bidirectional versus generative approaches in GPT versus BERT [27], to the efficiency-focused enhancements in RoBERTa and ALBERT as optimized versions of BERT [28]. The table also traces the progression within the GPT series, highlighting the notable advancements from GPT-2 to GPT-3, and further to the significant leaps made with GPT-4. Additionally, it contrasts the generative prowess of GPT-3 with the fine-tuned, user-aligned capabilities of InstructGPT. Furthermore, the comparison includes an exploration of XLNet’s permutation-based methodology against T5’s text-to-text framework, showcasing diverse approaches in handling complex language tasks. Overall, the comprehensive Tables 7 and 8 offer a consolidated view by amalgamating the individual discussions of each model’s unique characteristics, advancements, and their respective roles in advancing the field of NLP.

Table 7. Comparison of language models - group 1

Feature	BERT	GPT-1	GPT-2	GPT-3
Released	2018	2018	2019	2020
Model size	Base: 110 M, Large: 340 M	117 M	1.5 B	175 B
Architecture	Transformer, bidirectional	Transformer, autoregressive	Transformer, autoregressive	Transformer, autoregressive
Training data	3.3 B words	40 GB text	40 GB text	570 GB text
Capabilities	Text classification, NER, QA [29]	Language generation	Advanced language generation	Advanced NLP tasks, translation, coding
Pre-training tasks	MLM, NSP	Unsupervised learning	Unsupervised learning	Unsupervised learning and permutation-based LM
Special features	Bidirectional context understanding	Generative text	Style mimicry, generative text	Few-shot learning, style adaptation
Multilingual capabilities	Yes	Limited	Limited	Yes
Safety and bias	Context understanding	–	–	Prone to biases
Context window	–	–	–	2048 tokens
Modality	Text	Text	Text	Text

Table 8. Comparison of language models - group 2

Feature	XLNet	RoBERTa	ALBERT	InstructGPT/GPT-4
Released	2019	2019	2019	InstructGPT: 2021, GPT-4: 2023
Model size	–	Optimized BERT	Smaller than BERT-Large	InstructGPT: Fewer than GPT-3, GPT-4: 170T
Architecture	Transformer, permutation-based	Optimized BERT	Efficient BERT variant	Transformer, multimodal for GPT-4
Training data	Large corpora, diverse sources	Larger datasets than BERT	Similar to BERT, with efficiency	Extensive, includes text and images for GPT-4
Capabilities	NLU benchmarks like GLUE	Improved over BERT	Efficient, competitive performance	Advanced NLP, multimodal for GPT-4
Pre-training tasks	Permutation-based LM	MLM like BERT [30]	MLM like BERT	RLHF, RBRM for GPT-4
Special features	Captures bidirectional context	Larger batches, more data	Parameter reduction techniques	Multimodal, human feedback for InstructGPT, longer outputs for GPT-4
Multilingual capabilities	Yes	Yes	Yes	Yes, advanced for GPT-4
Safety and bias	–	Reduced biases	Reduced biases	Improved safety in InstructGPT, bias concerns for GPT-4 [31]
Context window	–	–	–	Up to 32768 tokens for GPT-4
Modality	Text	Text	Text	Text and images for GPT-4

4. CONCLUSION

In conclusion, the exploration of various language models, from BERT’s bidirectional approach to the latest advancements in GPT-4, underscores a significant trajectory in the evolution of natural language processing. Each model, be it RoBERTa, ALBERT, the GPT series, or XLNet, has contributed distinctively to the

field, pushing the boundaries of how machines understand and generate human language. GPT-4, in particular, marks a monumental leap forward. With 170 trillion parameters and the ability to process both text and image inputs, it stands as a titan in the realm of language models. Its capabilities extend impressively into the medical field, suggesting transformative potential in medical education, clinical reasoning, and research assistance. However, its prowess in passing medical competency examinations and supporting interactive learning scenarios comes with a caveat. The need for expert oversight and cautious application in high-stakes domains like healthcare is paramount, given the model's limitations, and potential risks.

Despite GPT-4's groundbreaking achievements, it inherits limitations from its predecessors, such as the propensity to "hallucinate" facts, generate biased or harmful content, and make reasoning errors. Its tendency to hedge responses and potential misuse in disinformation campaigns, privacy breaches, and cybersecurity threats highlight the dual-edged nature of this technological marvel. As the model finds applications in various domains, ranging from education to potential assistance in clinical settings, the emphasis on safety mitigations and ethical considerations becomes increasingly crucial.

The journey from BERT's foundational bidirectional processing to GPT-4's multimodal capabilities encapsulates a remarkable progression in AI's language capabilities. However, this journey is also a reminder of the responsibility that comes with such powerful technology. As we venture into an era where AI models like GPT-4 can simulate human-like text and reasoning, the importance of balancing innovation with safety, ethical considerations, and responsible usage cannot be overstated.

Ultimately, while models like GPT-4 represent significant strides in AI, they are not infallible. Their outputs must be critically evaluated, especially in contexts where accuracy and reliability are non-negotiable. The future of AI in language processing is not just about creating more advanced models but also about developing robust frameworks for their safe and ethical application.




REFERENCES

- [1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2383–2392, Jun. 2016, doi: 10.18653/v1/d16-1264.
- [2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2019.
- [3] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 2020, no. 1, pp. 9–15, May 2020.
- [4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, Jun. 2019.
- [5] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>.
- [6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: a Lite Bert for Self-Supervised Learning of Language Representations," *8th International Conference on Learning Representations, ICLR 2020*, Sep. 2020.
- [7] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, Oct. 2020.
- [8] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.08774>.
- [9] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.
- [10] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5999–6009, Jun. 2017.
- [11] A. Radford, K. Narasimhan, I. Sutskever, and T. Salimans, "Improving Language Understanding by Generative Pre-Training," *Homology, Homotopy and Applications*, vol. 9, no. 1, pp. 399–438, 2007.
- [12] M. Zhang and J. Li, "A commentary of GPT-3 in MIT Technology Review 2021," *Fundamental Research*, vol. 1, no. 6, pp. 831–833, Nov. 2021, doi: 10.1016/j.fmre.2021.11.011.
- [13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2018.
- [14] R. Dale, "GPT-3: What's it good for?," *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, Jan. 2021, doi: 10.1017/S1351324920000601.
- [15] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," *Zenodo*, 2021.
- [16] X. Yang, E. Peynetti, V. Meerman, and C. Tanner, "What GPT Knows About Who is Who," *Insights 2022 - 3rd Workshop on Insights from Negative Results in NLP, Proceedings of the Workshop*, pp. 75–81, May 2022, doi: 10.18653/v1/2022.insights-1.10.
- [17] B. Ghojogh, A. L. I. Ghodsi, and U. Ca, "Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey," *ArXiv*, vol. 1, no. 1, pp. 1–14, 2017, doi: 10.31219/osf.io/m6gen.
- [18] A. M. Dai and Q. V. Le, "Semi-supervised Sequence Learning," *Advances in neural information processing systems*, vol. 34, no. 4, pp. 227–230, 2015.
- [19] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *ACL 2018 - 56th Annual Meeting of the*




- Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 328–339, Jan. 2018, doi: 10.18653/v1/p18-1031.
- [20] P. Basu, T. S. Roy, R. Naidu, Z. Muftuoglu, S. Singh, and F. Mireshghallah, “Benchmarking Differential Privacy and Federated Learning for BERT Models,” Jun. 2021, [Online]. Available: <http://arxiv.org/abs/2106.13973>.
- [21] X. Zheng, C. Zhang, and P. C. Woodland, “Adapting GPT, GPT-2 and BERT Language Models for Speech Recognition,” Jul. 2021, [Online]. Available: <http://arxiv.org/abs/2108.07789>.
- [22] P. Banerjee, Anurag K. Srivastava, D. Adjeroh, R. Reddy, and N. Karimian, “Understanding ChatGPT: Impact Analysis and Path Forward for Teaching Computer Science and Engineering,” *TechRxiv*, pp. 0–20, 2023.
- [23] C. Zhou *et al.*, “A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT,” Feb. 2023, [Online]. Available: <http://arxiv.org/abs/2302.09419>.
- [24] M.-H. Hwang, J. Shin, H. Seo, J.-S. Im, H. Cho, and C.-K. Lee, “Ensemble-NQG-T5: Ensemble Neural Question Generation Model Based on Text-to-Text Transfer Transformer,” *Applied Sciences*, vol. 13, no. 2, p. 903, Jan. 2023, doi: 10.3390/app13020903.
- [25] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016, [Online]. Available: <http://arxiv.org/abs/1609.08144>.
- [26] A. Koubaa, “GPT-4 vs. GPT-3.5: A Concise Showdown,” Preprints, no. March, pp. 1–5, 2023, doi: 10.36227/techrxiv.22312330.v2.
- [27] R. Wolfe and A. Caliskan, “Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models,” *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 518–532, Oct. 2021, doi: 10.18653/v1/2021.emnlp-main.41.
- [28] T. Schomacker and M. Tropsman-Frick, “Language Representation Models: An Overview,” *Entropy*, vol. 23, no. 11, p. 1422, Oct. 2021, doi: 10.3390/e23111422.
- [29] M. V. Koroteev, “BERT: A Review of Applications in Natural Language Processing and Understanding,” Mar. 2021, [Online]. Available: <http://arxiv.org/abs/2103.11943>.
- [30] M. M. Abdul Qadar and V. Mago, “A Survey on Language Models,” *Association for Computing Machinery*, vol. 1, no. September, pp. 1–31, 2020.
- [31] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, “Capabilities of GPT-4 on Medical Challenge Problems,” Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.13375>.

BIOGRAPHIES OF AUTHORS






Asmae Briouya    is a dedicated software engineer who specializes in web and mobile application development. She successfully completed her Bachelor’s degree in Computer Science in 2021, marking the beginning of her academic and professional journey. Alongside her bachelor’s degree, she pursued a Ph.D. program at Ibn Tofail University. In her Ph.D. research, she focuses on two exciting areas: natural language processing (NLP) and 3D data segmentation. NLP involves the development of algorithms and models that enable computers to understand and process human language. Simultaneously, she explores the intricate domain of 3D data segmentation, aiming to divide three-dimensional data into meaningful and distinct parts. She can be contacted by email at: asmae.briouya@uit.ac.ma.



Hasnae Briouya    is a committed software engineer specializing in the development of web and mobile applications. In 2021, she successfully obtained her Bachelor’s degree in Computer Science, which marked the beginning of her academic and professional journey. Concurrently, she pursued a Ph.D. program at Ibn Tofail University. Her Ph.D. research focuses on the advanced exploration of computer science, with specific emphasis on semantic segmentation of 3D data and natural language processing (NLP). This area involves the analysis and comprehension of the structure and meaning of 3D data to improve the process of segmentation. She can be contacted by email at: hasnae.briouya@uit.ac.ma.



Ali Choukri    is an assistant professor at the National Academy of Applied Sciences, having completed his Master’s in Computer Science and Telecommunications at the University of Ibn Tofail, Kenitra, Morocco, in 2008. He holds a Ph.D. from the School of Computer Science and Systems Analysis (ENSIAS) and a degree from ENSET (Higher Normal School of Technical Teaching). Currently contributing to the MIS team in the SIME laboratory, his research focuses on mobile intelligent ad hoc communication systems and wireless sensor networks. His diverse research interests include ubiquitous computing, the internet of things (IoT), QoS routing, mathematical modeling, game theory, and optimization, among others. He can be contacted by email at: ali.choukri@uit.ac.ma.