https://ijrcsit.org

International Journal for Research in Computer Science and Information Technology (IJRCSIT)

Original Article

Natural Language Processing in the Age of Foundation Models: From BERT to GPT-5

Shenroy Varatharaj[1], Kamal Viswanathan[2], Arunai Ganesan[3], Dr. Siva Sankari[4]

[1,2,3]UG Scholars, [4]Associate Professor.,
Don Bosco college of arts and science keela eral, Ettiayapuram, Tamilnadu, India.

Received Date: 03 October 2025 Revised Date: 17 October 2025 Accepted Date: 31 October 2025

Abstract

Natural Language Processing (NLP) has transformed from systems relying on handcrafted linguistic knowledge and co-occurrences, to statistical approaches with feature-based pattern matching, and recently into a paradigm shift based on foundations models—large scale pre-trained architectures learning representations over large amounts of data in an unsupervised way expected to capture general properties of the language across task and domain. Those transformation has revolutionalized the nature of machines understanding, representing and generation human language which was firstly infiringed by Transformer architecture. The pre-trained models: BERT, GPT-3, T5, PaLM and GPT-5 are the giant leap in computational linguistics. They replace narrow-eyed learning with one-shot universal linguistic understanding as are reasoning machines!

This paper conducts a thorough investigation on the evolution of NLP in the era of foundation models, highlighting both advancement in technology and philosophical reflection behind it. It studies how the bidirectional pretraining of BERT brought deep contextual understanding and how later autoregressive models such as GPT-3 and GPT-4 evolved it into creative and reasoning abilities. architectures Agentic AI9 The shift toward GPT-5 suggests a new mental model for agentic AI systems, which might be defined more directly in terms of the ability to perform high-level reasoning and metalearning, using internal goal decompositionalgorithms, planning algorithms,andhownovel agents learn about their own reasoning processes through meta-learning or feedback loops based on its own actions.

Moreover, we try to provide some understanding the core architecture of foundation models (e.g., self-attention), along with scaling laws by exploring pretraining objectives and fine-tuning methods including RLHF and Constitutional AI. The work compares the most relevant models in terms of number of parameters, diversity of data, modality and general downstream. It demonstrates that NLP and multimodal intelligence are increasingly joining forces, as modern models unify text, vision and audio processing under one roof to approximate general cognitive reasoning.

Much of this work involves applying foundation models in practice — including machine translation, summarization, sentiment analysis, code generation, education tutoring and healthcare communication. These features further reveal how NLP models are emerging as cognitive artifacts that augment human abilities to discover and communicate knowledge.

At the same time, the paper addresses important ethical, interpretability and environmental issues related to training and using such large models. Problems such as biased data, hallucination, invasion of privacy and high power consumption suggest that responsible regulation and sustainable strategies for AI are needed. We focus on the recent trends such as accurate/efficient fine-tuning, model compression and XAI (Explainable AI) techniques to make models more transparency/fairness.

In the end, I believe that it reminds us that even when expanded to BERT-based and further GPT-5-like systems, the history of NLP is one leading towards foundation-driven AI: language models don't just help people analyze and generate language but may become fully autonomous agents capable of reasoning or acting in a world. The paper claims that in the next phase of NLP research, we must pay special attention to the generalization and interpretability of these models when evolving foundation models—balancing model size with transparency and control—and advocate for a more rigorous *treatment of actual as well as potential risks: ethical, systemic or emergent which pertain not only to biases but also to ecological impacts such models may have in their environments.*

Keywords

NLP, BERT, GPT-5, Transformer, Foundation Models, Large Language Models, Deep Learning, Self-supervised Learning.

Introduction

In the past decade or so, NLP has rapidly evolved to redefine an era in AI. From previously rule-based systems and statistical models, NLP has been developing from data-starving models for specific tasks to data-saturated large-scale foundation models that are able to capture rich linguistic and semantic information as well as context from vast amounts of data. This revolution—from explicit programming to representation learning and self-supervised pretraining—has dramatically leveled up what can be done not just in the technical stories of how machines understand language, but also in the real-world tasks we can create across industries.

Until now, most of the NLP research was focused on handcrafted features and surface models, including Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) as well as bag-of-words representations. Though these approaches perform fairly well at particular tasks such as part-of-speech tagging and named entity recognition, their performance significantly deteriorates when generalized to various contexts and domains. The development of word embeddings like Word2Vec, GloVe and more recently FastText was the first step toward distributed meaning representation: words can be embedded into a dense vector space that captures semantic relations. But these word embeddings were context-insensitive: a word had the same fixed representation, (formally, its embedding), no matter where it appeared in a sentence.

Inspired by the Transformer model of Vaswani et al. In 2017 is when the real NLP revolution began. The transformer achieved the above task by removing recurrence and introduced a self-attention mechanism so that models could process input sequences in parallel and effectively capture long-term dependencies. This breakthrough eventually led pretrained language models (PLMs) like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer), which acquired language representations by training on huge corpora without any supervision and then could be fine-tuned on downstream tasks even with scarce labelled data.

BERT proposed by Devlin et al. first to apply bidirectional context modeling, which allowed finding deeper relationships between words in text. The proposed M-BERT pretraining used masking language model (MLM) and next sentence prediction (NSP) in BERT as the pretrain tasks which achieved new state-of-the-art results on the downstream language understanding tasks. GPT's auto-regressive architecture meant text generation was modeled uni-directionally – great for coherent and contextually relevant completion of text. These two complementary paradigms, with BERT for understanding and GPT for generation, laid out the foundation of today's NLP field.

As model size, data diversity, and compute power has increased, subsequent variants—GPT-2, GPT-3, GPT-4 and now GPT-5—have seen quantum leaps in generative fluency, reasoning capacity and domain adaptation. HAI's definition of foundation models characterizes these systems as general-purpose, flexible models trained on large available data covering a wide range of use cases. In contrast to task-specific models, foundation models have emergent properties including few-shot learning, zero-shot generalization and cross-domain portability that allow them to perform tasks for which they were never explicitly trained.

In the age of GPT-5, NLP has just exited yet another paradigm in which AI models are not only generating/analyzing text, but they also plan, reason and act autonomously. Such models utilize RLHF, memory-based retrieval systems, and multi-modal alignment to allow the modelling of text, images, audio and structured data in a coherent representation. Systems that result go beyond traditional NLP boundaries and into the domains of reasoning, simulation, and decision-making with applications to education, healthcare, legal analysis as well as robotics.

These developments have a number of effects. Technologically, the base model paradigm opens up a possibility of extremely large-scale transfer learning that requires less data and computation for domain adaptation. Societally, it levels the playing field for who can have access to advanced AI capabilities, enabling conversational systems and translation engines available now to millions. Simultaneously, such exponential growth raises key questions about bias, interpretability, data privacy and sustainability. Large-scale models are expensive to train in terms of computational resources and high carbon emissions PMIDAIN2021TransformersWB. Further, the black-box nature of such systems hiding their decision-making can make it difficult to hold them accountable, and use them ethically.

Another important trend is the rise of multimodal NLP where models are not only exposed to text streams but also visual and auditory ones. The integration of language and perception is what allows models such as GPT-4 and Gemini to reason about the world with shared linguistic and visual understanding. This multimodality expands the landscape of human–AI communication and offers opportunities for intelligent tutoring systems, visual question-answering, creative design, or robotics.

Aside from the remarkable demonstration of their abilities, a new era of linguistic theory and AI cognition is now incited by foundation models. They also call into question traditional concepts of meaning and syntax, proposing that intelligence can emerge from massive pattern matching and not pure symbolic manipulation. So, the study of NLP in the time of foundation models is not merely an issue of engineering performance targets but something that reaches out across linguistics, cognitive science, computer science and ethics.

In this work, we hope to offer a detailed study on the paradigm shifting era of NLP; the advent of foundation models: from BERT to GPT-5. This study aims to address three specific goals:

- To study design and algorithmic advances that have shaped the development of base models;
- To investigate strengths, weaknesses, and ethics of these models on a wide range of NLP tasks; and
- To talk about trends we observe in the field and directions to make NLP systems interpretable, efficient and sustainable.

The structure of the rest of this paper is as follows: In Section II, we review historical progression and several landmark achievement for NLP model. :- Section III describes the building of architecture from BERT to GPT-5 that underlines progress in self-supervised learning, scaling laws and reinforcement learning. In Section IV, we discuss key applications of foundation models across different fields. In Section V, we discuss the issues of bias, energy and explainability. The final Section VI concludes the paper providing perspectives on the next stage of laugauge models and their implication for AGI.

Evolution of NLP Models

The advancement of Natural Language Processing (NLP) has been a story along the trend from bottom-up to top-down, or abstraction/representation level by representation level, or state-of-the-art in terms of means and purposes, starting from symbolic system and reaching neural network that can do reasoning on its own. For the last 30 years, the field evolved through different technological stages, each with its distinct modeling philosophy, computational power and data)omosphere to nowcast51-52 but it is a challenge for models because of two-facto availability. In the following, we first provide a brief historical overview of NLP models along a timeline, highlighting landmark events that culminated in the era of foundation models and GPT-5.

A. Early RB and Statistical Work

To its first generations, from the 1950s through the 1990s, NLP was other words for handcrafted rules (formal grammars plus symbolic logic). These systems sought to capture the structure of human language by specifying explicit syntax trees and pattern matching algorithms. Examples include ELIZA (1966) which was the first example of a conversation program and SHRDLU (1970)--a robot who could understand simple natural language input in English regarding graphs.

But these techniques were fragile and narrow, unable to handle linguistic uncertainty or contextual diversity. The advent of statistical NLP in the 1990s also changed the direction of research. As text corpora grew and computational power increased, language itself began to be modelled as a probabilistic process. N-gram language models, Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs) came to dominate a range of tasks like part-of-speech tagging, speech recognition and machine translation.

Despite their effectiveness, data-driven models were limited by two factors.

- Data sparsity — they performed poorly on unseen combinations of words; and
- Feature engineering overhead — systems had many handcrafted design of linguistic features.

These challenges led to the shift towards distributed representations and neural embeddings that are capable of capturing richer semantic relationships between words.

B. The Era of Word Embeddings

The next milestone was the discovery of neural word embeddings representing words as dense, continuous vectors in high-dimensional space. The groundbreaking work of Word2Vec (Mikolov et al., 2013) showed that simple neural networks with little to no supervision were able to learn meaningful relationships between words directly from raw text data, such as analogies like "king – man + woman ≈ queen". Other methods such as GloVe (Global Vectors) and FastText further enhanced the semantic quality and robustness for morphologically rich languages.

Word embeddings transformed the NLP field, where semantic similarity, clustering, and compositional reasoning can be conducted between different tasks. However, these were context-independent models—a single vector representation was learned for each word type irrespective of the particular contextual role it played. For instance, "bank" was represented by the same word in "river bank", and "financial bank". This limitation motivated the trend to develop a more contextualized embeddings by using deeper neural architectures.

C. The Emergence of Deep Contextual Models

The proliferation of Recurrent Neural Network (RNNs) families like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) allowed models to work with sequential data maintaining context across many time steps. LSTM-based models led early breakthroughs in the fields of neural machine translation (NMT), speech recognition, and text classification.

RNNs, however, suffered from long-range dependencies because of the vanishing gradient and inefficient sequential processing. In the face of these challenges, Vaswani et al. introduced the Transformer architecture. in 2017, in the milestone paper "Attention Is All You Need." The Transformer replaced the recursion with self-attention mechanisms that enable models to selectively focus on all words in a sentence together. This invention naturally endowed the model with both parallelization capacity and long range attention, changing the landscape of NLP architecture design.

D. The Emergence of Pretrained Language Models (BERT and GPT Family)

Thanks to the Transformer architecture, the Pretrained Language Models (PLMs) are created which generalize language representations through self-supervised learning over large scale corpora. These models are then fine-tuned for specific downstream tasks using relatively small amounts of labeled data.

a) BERT: Bidirectional Understanding

BERT (Bidirectional Encoder Representations from Transformers) introduced by Google AI in 2018, brought great advancements in the field of contextual language processing. Its bidirectional attention mechanism allowed the model to learn representations that depended on both left and right context, which made it applicable for tasks such as question answering and sentiment analysis while achieving performance close to that of NER. The pretraining tasks of BERT, including Masked Language Modeling (MLM)and Next Sentence Prediction (NSP), enabled the model to capture deep bidirectional context. Models such as RoBERTa, ALBERT and DistilBERT enhanced BERT's performance and efficiency.

b) GPT: Autoregressive Generation

Meanwhile, OpenAI's GPT (Generative Pretrained Transformer) series further developed the autoregressive approach and output models that predict the next token in a sequence given previous context. At understanding, BERT models have done well, but at generation GPT models are superior for language completion and dialogue. GPT-2 (2019) presented multi-paragraph text generation, whereas GPT-3 (2020) with 175 billion outcomes illustrated few-shot learning capability, where a model may perform certain tasks given only a small number of examples without explicit fine-tuning.

E. Uptake of Foundation Models (GPT-4 and GPT-5)

Formally defined by Stanford HAI (2021), foundation models is a term that describes large pretrained models that can be fine-tuned for a tremendous variety of downstream tasks. GPT-4 and GPT-5 are exemplary of this paradigm, capable as they are of thinking, planning and adapting over linguistic and multimodal domains.

The GPT-4 (2023) responded to this by integrating knowledge representation and reasoning capabilities directly into its architecture of combined text and image comprehension. It has been a good demonstration of emergent reasoning abilities superseding human-level performance in many standardized tests.

GPT-5 (2025) is the next leap, which introduces agentic intelligence; here, goal-setting is emergent (\emph{i.e.}, complex objectives get broken down on its own), self-planning eschews in a natural manner and output interacts with tools and APIs to achieve real-world tasks. It also combines reinforcement learning, retrieval-augmented generation (RAG), persistent memory – closing the gap between static text understanding and dynamic decision making.

F. Comparison Among the Key Models

The comparison of leading language models over different periods is listed in Table.

Table 1. Evolution of NLP Models — Key Characteristics and Innovations

Model	Year	Architecture	Parameters	Core Objective	Key Innovation
Word2Vec	2013	Neural Embedding	100M	Word Co-occurrence	Distributed word vectors
ELMo	2018	Bi-LSTM	94M	Contextual Language Modeling	Dynamic word representations
BERT	2018	Transformer Encoder	340M	MLM, NSP	Bidirectional pretraining
GPT-2	2019	Transformer Decoder	1.5B	Autoregressive	Generative fluency
T5	2020	Encoder–Decoder	11B	Text-to-Text	Unified text generation tasks
GPT-3	2020	Transformer Decoder	175B	Autoregressive	Few-shot learning
PaLM	2022	Transformer	540B	Self-supervised	Scaling and reasoning
GPT-4	2023	Multimodal Transformer	1T (approx.)	Multimodal understanding	Vision-language fusion
GPT-5	2025	Agentic Multimodal Transformer	>1.5T	Reinforced self-learning	Goal decomposition & planning

G. The Trend to a Unified Generalized Language Intelligence

The leap from BERT to GPT-5 is a move not only in size but in a whole paradigm—from language model to general intelligence. By adding memory, reasoning and multi-agent co-ordination, such systems cease to be passive processors of text and become agents with cognitive capabilities. Furthermore, recent work on retrieval-augmented architectures, parameter-efficient fine-tuning (PEFT), and low-rank adaptation (LoRA) has shown that intelligence can be scaled without increasing resources.

The history of NLP therefore reflects the larger evolution of AI: from pattern recognition to cognitive gesture; from task-specific modelling to domain general learning. The foundation, up to GPT-5, is a model can reason and infer on its own as it processes language, thus taking us closer to what I in this series have defined linguistic AGI.

H. Image Placeholder

[Fig. 1:The Evolutionary Trajectory of NLP Models from Statistical to GPT-5 (Illustrative Placeholder)

Figure caption: A flow chart summarizing the chronological progression of NLP models, from statistical methods through embedding-based techniques and Transformer based architectures to contemporary parent models.

Foundation Models: Concept and Properties

A. Defining Foundation Models

The concept of foundation model originated by Stanford's Institute for Human-Centered Artificial Intelligence (HAI) in 2021, refers to the latest generation of large-scale AI systems that are trained on massive, diverse self-supervised data. E.g., in Figure 1, EA can be a solid and robust model that is effectively used as a base which can then be fine-tuned (or adapted) for many downstream tasks without doing back the task-specific retraining.

Compared to conventional NLP models which are designed for narrow and predefined tasks, foundation models show stronger performance in generalizing and transferring. They serve to be the "building blocks" which can be used as components to structure various specialized systems (such as text generation, question answering, dialog or translation) including code synthesis. Their relative flexibility is owed to their capacity to learn universal language representations that encode not only lexical semantics but also more complex patterns of reasoning, discourse structure, and pragmatic inferences.

More formally, we denote a base model as Fθ which is trained with parameters θ on a large corpus D under some self-supervised or unsupervised objective Lself (x;θ), so that the learned represenation h(x) can be easily adapted to any downstream task T with little fine-tuning. It is this transferability property that sets foundation models apart from traditional pretrained models.

B. Characteristics of Foundation Models

Property of foundation models Foundation model exhibits multiple characteristics which makes them significantly different from previous NLP systems. These aspects are detailed in the next subsections.

a) Scale and Generalization

Foundation models are powerful because of their scale — data and parameters. The scaling laws of Kaplan et al. (2020) show that model quality scales monotonically with increased number of parameters, dataset size, and compute. Models like PaLM with 540B parameters, as well as GPT-4 (1T parameters) and GPT-5 (>1.5T parameters), draw on trillions of tokens in multilingual, multimodal and code-based data to learn high-level linguistic and world knowledge.

This scaling results in emergent generalization – models can perform natural tasks at test time even if they were not trained on them, e.g., analogical reasoning (Mikolov et al., 2013a; Padi-prep and Mikolov, 2015), compositional logic (Lake and Baroni, 2018), zero-shot task performance. In this way, scale is less a quantity and more of a qualitative cutoff for the onset of intelligence.

b) Self-Supervised and Multitask Learning

Base models are trained using selfsupervised learning, by predicting the missing portions of data (e.g. masked words, next tokens or sentences) instead of labeled examples. This allows for training on web-scale text, code, and multimodal corpora and results in models that understand syntax, semantics, as well as factual structure.

Moreover, they usually use multitask learning objectives, which mix different deployments of linguistics (translation, summarization or reasoning question-answering) in the same framework. This strengthens cross-domain transferability and resiliency to overfitting for particular linguistic tasks.

c) Adaptability and Fine-Tuning Efficiency

One of the strengths of foundation models is they can be customized through fine-tuning and prompt engineering. Contrary to retraining the complete networks, lightweight approaches like PEFT (Parameter-Efficient Fine-Tuning), LoRA (Low-Rank Adaptation) and prefix-tuning allow adaptation of new domains in a computationally efficient manner.

Furthermore, foundation models can be steered via instruction or task tuning, where they are tuned to follow human-like commands in natural language and Reinforcement Learning from Human Feedback (RLHF) that aligns their outputs with human preferences. GPT-4 and GPT-5 use a lot of both techniques in order to get assistance, ensuring that they are responsive, safe and context-aware.

d) Multimodality

Recent research on model architectures has been multimodal by nature, involving the processing of text, vision and audio as one holistic unit. It enables them to parse visual data, produce descriptive text, and carry out multimodal reasoning tasks like image captioning or visual question answering. For example, GPT-4 and GPT-5 are trained with text–visual pairs to enable them to understand graphs, charts and diagrams in addition to the text.

The characterization of NLP in terms of multimodal foundation models rebrands it as a component of universal intelligence whose purpose is to mediate between perception, cognition and action. They enable applications including robot control execution, video understanding and interactive tutoring systems beyond simple language-based tasks.

e) Emergent and Agentic Behavior

When foundation models grow in size, they start showing emergent behaviors—what can be considered as capabilities that emerge spontaneously due to the various complexities of the underlying model and its associated data. Instances are few-shot learning, in-context reasoning, or tool use via natural language interfaces.

The move towards agentic behaviour, especially obvious in GPT-5, is something of a threshold for AI. Agentic models can plan, reason, and act as individuals do – they can decompose goals into immediate actions, query external tools or databases, learn from feedback. This makes NLP systems akin to cognitive architectures, where the model representation does not include only language understanding but also problem solving and decision making.

f) Continuous Learning and Memory Integration

Conventional models are static in the sense that they stop updating their parameters when training is over. On the other hand, basic models are now beginning to include more structure to capture interaction between memory modules and sorts of retrieval that can support systematic revision. Retrieval-Au Gent mented Generation (RAG) and External Memory Networks incorporate external knowledge graph for the model to retrieve new information without learning from scratch.

For example, GPT-5 comes with stateful context memory which can persist across sessions and support user-specific understanding. This development allows for learning to occur incrementally, as models and data/users continue to co-evolve, akin to how humans learn.

C. Architectural Developments in Grounding Schemes

Base architecture The base structure of foundation models is based on the transformer, but has many more complicated techniques [46].

- Sparse Attention and Efficient Transformers:To cope with long sequences efficiently, models like the Longformer and BigBird employ sparse attention mechanisms to lower quadratic complexity.
- Mixture of Experts (MoE):GPT-4 and GPT-5 use Mixture-of-Experts layers with subsets of the parameters activated per-input, which allows them to scale much more without linear compute growth.
- Reinforcement Learning Integration:RLHF and Constitutional AI work on post-model behaviour minimization tasks to nudge the model's output towards ethical and social norms.
- Knowledge Integration:However, modern models inject structured knowledge graphs and employ retrieval-augmented strategies to further improve absolute grounding and decrease hallucinations.
- Multimodal Fusion Networks:GPT-5-like models are based on joint encoders which combine vision, text, and audio modalities to generate shared representations for cross-modal understanding.

D. DLMs vs Traditional NLP Models

Table I compares the feature of origin and foundation models at scale, learning paradigm and reduce learning tasks.

Table 2. Comparative Characteristics of NLP Models and Foundation Models

Feature	Traditional NLP Models	Foundation Models
Architecture	Task-specific (RNN, LSTM, CNN)	Unified Transformer-based
Training Objective	Supervised, labeled data	Self-supervised, unsupervised
Data Scale	Thousands to millions of tokens	Trillions of tokens
Learning Type	Task-specific fine-tuning	Universal pretraining + adaptation
Contextual	Limited to sentence level	Global and multimodal context

Understanding		
Adaptability	Low (requires retraining)	High (via prompts or PEFT)
Interpretability	Moderate (explicit features)	Low, but improving (XAI methods)
Applications	Narrow (translation, tagging)	Broad (reasoning, generation, decision-making)
Behavior	Deterministic	Emergent and agentic
Examples	RNNs, BERT	GPT-4, GPT-5, PaLM, Gemini

E. Characteristics Underpinning Foundation Model Success

Foundation models are successful because of a number of key properties that come together to form general-purpose intelligence:

- Compositional Generalization:The faculty of the combination for learned concepts to solve unseen tasks or contents.
- In-Context Learning:During inference models learn tasks by conditioning on a small number of examples (few-shot or zero-shot learning).
- Emergent Reasoning:"The potential to do logical, mathematical, and symbolic reasoning over and above pattern recognition.
- Cross-Domain Transfer:Heterogeneous data trained base models can generalize to multiple domains without retraining.
- Alignment and Safety:The alignment of ethical and behavioral attributes using RLHF ensures that the generated outputs are aligned with human values and societal norms.
- Scalability and Adaptation:Increasing order and dilatable training support up-scaling without complete re-initialization.

F. Limitations and Ongoing Challenges

However as powerful as they are, foundation models also come with significant limitations:

- Interpretability: Due to their black-box nature, decision processing pathways are hard to trace.
- Bias and Fairness: Data biases are often learned by a model.
- Cost: It is also not environmentally friendly to computationally train a very large model.
- Staleness of Knowledge: static pretraining data may be outdated.
- Ethics and Regulation: The balance between the innovation of new products, coupled with transparency and citations/references, continues to be a significant issue.

Current research efforts focused on overcoming these limitations include explainable AI, efficient training pipelines, data curation stacks and human-AI teaming.

G. Summary

Base models form the fusion of size, architecture and adaptability in NLP. Not just larger but of a qualitatively different order of models—universal language reasoners able to both comprehend and generate, as well as reason about, the world. From BERT's bidirectional grasp to GPT-5's multimodality and agenticity, foundation models are a big step in the direction of AGI.

Their properties—scalability, self-supervision, transfer, and emergent reasoning—underpin next-generation NLP systems that are both the intellectual bedrock of future work in AI alignment, cognition, and human–machine interaction.

Comparison of Progression: Bert To Gpt-5

The progression from BERT (Bidirectional Encoder Representations from Transformers) to GPT-5 (Generative Pre-trained Transformer 5) in Natural Language Processing models describes one of the most remarkable technological shifts humanity has witnessed to date in artificial intelligence. This evolution is not simply due to larger models or faster computation – it corresponds to a transition from context-based understanding towards self-driven reasoning and goal-directed intelligence.

A. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Powered by Google AI, BERT is presented as the biggest milestone in NLP since the 2018 revolution with transformers. The key innovation behind BERT is that it was alternatively pre-trained using Bidirectional Encoder Representations from Transformers (BERT) method with the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In contrast to previous sequential models like RNNs and LSTMs, BERT enabled positional embeddings for both words as well as surrounding tokens, giving fully expressive deep lexical semantic embeddings on which downstream NLP tasks including Question Answering (QA), Sentence Classification (e.g. sentiment analysis), Information Retrieval were built. Its success showed that large-scale self-supervised learning can surpass supervised models, and has opened up the possibility of universal language representations.

B. GPT Series Generative Autoregression Comes of Age

The Generative Pre-trained Transformer (GPT) series by OpenAI turned the NLP direction upside down to attend, over autoregressive generation.

- GPT-2 (2019) has taken researchers aback by its capability of producing coherent multi-sentence text and zero-shot learning on various tasks.
- GPT-3 (2020) scaled this relatively simple design up to 175 billion parameters, resulting in striking emergent capabilities for reasoning, programming and dialogue without any task-specific training.
- GPT-4 (2023) took the paradigm one step further, incorporating multimodal functionality, chain-of-thought reasoning, and reinforcement learning from human feedback (RLHF), leading to substantial improvements in coherence, truthfulness, and contextual relevance.

These developments collectively showed that increasing "scale"—meaning larger models, but also means more data and higher diversity of it (such as Intemet-scale data) could lead to qualitative jumps in intelligent behavior rather than just incremental improvements.

C. T5, PaLM, and Beyond: Task Unification and Efficient Scaling

Between BERT and GPT-4, models such as T5 (Text-to-Text Transfer Transformer), XLNet, and PaLM (Pathways Language Model) brought us closer to task convergence thanks to the ideas of unifying tasks and efficiency-scaling.

- T5 (2019) reformulated all NLP tasks as text-to-text tasks, which made training models and transfer learning easier.
- PaLM (2022) presented Pathways, allowing for a model to generalize across several domains and dynamic parameter sub-setting -being efficient as well as gaining in generalization.

Chinchilla (2022) and LLaMA (2023)also recently stressed on the importance of data– parameter balance, demonstrating that performance relies more from the quality and token count of the data, rather than just in parameter size.

D. GPT-5: On the Path to Independent and Self-Motivated Intelligence

GPT-5, the next major leap: It's a new layer of neural network that fuses generative modeling and reasoning with self-reward. The early trends in literature indicate that GPT-5 incorporates:

- Goal hierarchy and self improvement, whereby the agent can automatically decompose a task;
- Multi-agent coordination which can cooperate by means of specialized model instances;
- Multimodal, and memory-augmented reasoning across text, vision, audio and code in a single unified interface;
- Adaptive learning loops in which feedback based on the user behaviour plays an important role in adjusting the model's future responses.

Whereas BERT was based on static word representations, GPT-5 is something like a continuous, adaptive agent which can reason about and plan responses to complex tasks involving many different domains. It takes NLP from passive understanding to active problem solving, and usher in the era of agentic AI.

E. Comparative Summary

Model	Year	Core Innovation	Training Objective	Parameters	Distinctive Capability
BERT	2018	Bidirectional encoding	Masked Language Modeling (MLM)	340M	Deep contextual understanding
GPT-2	2019	Autoregressive transformer	Next-token prediction	1.5B	Coherent text generation
GPT-3	2020	Massive scaling	Autoregressive LM	175B	Zero-shot learning & reasoning
T5	2019	Text-to-text task unification	Sequence-to-sequence pretraining	11B	Unified framework for all NLP tasks
PaLM	2022	Pathways architecture	Multitask pretraining	540B	Efficient large-scale reasoning
GPT-4	2023	Multimodal alignment	RLHF, Chain-of-Thought	~1T (est.)	Reliable multimodal reasoning
GPT-5	2025	Agentic reasoning & autonomy	Adaptive self-learning	Undisclosed	Self-planning, multi-agent cognition

F. Discussion

The progression from BERT to GPT-5 represents the larger pattern of development in AI, from language modeling to cognitive modeling. With BERT it was comprehension, with GPT-3 expression, with GPT-4 multimodal reasoning and with GPT-5 autonomy. As we passed through each generation there was a jump in the leaps and bounds of computational approach and breakthroughs concepts of intelligence. Once models are sufficiently large and complex, they verge on the very boundary where language understanding is nothing but reasoning: where this begins to happen is the essence of foundation model intelligence.

Applications of Foundation Nlp Models

Foundation models have effectively turned NLP from a task-specific solution to a flexible cognitive platform that can be used to solve the language and reasoning challenges across the board. These models, ranging from BERT to GPT-5, yield generalized representations and can be fine-tuned across specific domains with limited retraining. This part of the book discusses the most significant application domains for foundation models, and showcases how their potential spreads from language understanding, to generation, reasoning, and human-AI interaction.

A. Text Understanding and Classification

One of the first and most [xref]common[/xref]/widespread use cases of foundation models is text understanding, which includes but is not limited to:

- Sentiment Analysis: Analyzing opinion about the correctness of reviews, social media comments, or suggestions. Deep LSTM-based classifiers and BERT have been shown to get state of the art performance in this task due to having deep contextual embeddings, but transformers such as GPT-4 or GPT-5 might be better suited for more nuanced sentiment detection which include sarcasm or context-dependent interpretations.
- Topic Classification and Clustering: News articles, academic papers and customer feedbacks can be labeled automatically thanks to the knowledge of semantics possessed by foundation models. Multimodal embeddings in GPT-5 will enable incorporating textual, visual and audio clues to improve the classification accuracy.
- Named Entity Recognition (NER): Classification of entities like person, organization and location in unstructured text. Also, thanks to pretrained models fine-tuned on labeled documents, it is not necessary to have large labeled training sets: This could help domain adaptation of language model in legal or biomedical texts, in financial news etc.

B. QA and Conversational AI

Foundation models have greatly driven interactive NLP applications, teaching machines to perform meanings and question reasoning:

- Closed and Open Domain Question Answering: BERT-based models are quite good in extracting the exact answers from structured text while and GPT-4 and GPT-5 utilize retrieval-augmented generation (RAG) for giving contextually correct and factually grounded answers.

- Chatbots and Conversational Agents: Using GPT-3 to GPT-5 makes conversational AI possible with the retention of context, multi-turn dialogue, adaptive tone etc. Use cases range from customer service automation, educational tutoring and mental health support to enterprise help.
- Interactive Search: Natural language interaction with databases or documents is supported. GPT-5 and beyond include multimodal inputs, where you can query images/graphs/tables in addition to text.

C. Text Generation and Summarization
Generative capacity of backbone models is essential to today's NLP applications:
- Text Summarization: Extractive and abstractive summarization tasks leverage contextual comprehension in BERT and generative fluency in GPT models. GPT-5 will be able to write summaries of lengthy documents, research papers, or mutli-document corpora clearly and concisely.
- Creative Writing: With GPT models, automated story generation, poetry and content creation has been improved. By including multimodal reasoning, GPT-5 could incorporate visual signals and audio samples as part of its produced text.
- Abstractive Summarization: The models can be used to paraphrase, while preserving the meaning or change tone and style of text for different audience hence supporting applications in marketing, journalism, accessibility etc.

D. Translation and Multilingual NLP
T _ PRE-TRAINED FAIRSEQ ture and cross-lingual understanding: the space of reasonche min!
- Neural Machine Translation (NMT): Key NLP models like BERT and T5 were able to earn benefits for language representation that transfer across translation, but GPT-4/5 integrate few-shot multilingual translation for ~30 languages.
- Cross-Lingual Information Retrieval GPT-5's base architecture facilitates searching and retrieval of information in any language, making the world's knowledge and content more universally accessible.
- Multilingual Chatbots: Intelligent AI can speak multiple languages fluently and break language barriers in global business, education and healthcare.

E. Knowledge Extraction and Reasoning
Foundation models go beyond NLP at the surface and cover cognitive reasoning & knowledge discovery:
- Extraction, Models can parse unstructured text to extract relationships between entities and events. Because GPT-5 is memory-augmented, it can be used to continuously add knowledge online (e.g. dynamic knowl edge graphs).
- Logical and Commonsense Reasoning: Advanced models contain multistep reasoning, inference, or deduction. You see GPT-5 exhibits chain-of-thought reasoning so as to reason through complex problems in areas such as legal research, scientific discovery and technical troubleshooting.
- Machine-Summarizing Literature: Models summarise large corpora, outlining trends. hypotheses and findings increasingly speeding up the discovery loop.

F. Code Production and Applications}
Recently, foundation models also expand NLP to programming and technical documents:
- Automatic Code Generation120: It is well known that GPT-3 and GPT-4 models are able to produce code snippets in various languages from natural language specification. GPT-5 introduces stronger reasoning to auto-debug, auto-optimize, and autodocument code.
- Software doc & QA: Get automatically generated technical documentation and answer software queries to relieve your developers from manual work.
- Multimodal Technical Assistants : GPT-5 demonstrate the faculty to consider natural language description together with sketches, design schematics or plots and produce a task-based instruction in engineering and scientific tasks.

G. Education, Health and Human Centered AI
Founded models have been employed in societal areas where understanding and adaptation to the context is essential:
- Personalized Education: Tutors that are powered by GPT-5 give personalized explanations, generate practice exercises and adjust their pedagogy based on the performance and learning style of the student.
- Healthcare Communication: Models can be used in clinical note summarization, patient question answering and medical literature analysis. Multimodal comprehension allows text data to interpret diagnostic images.

- Human-Centered AI: By recognizing user intent and emotion, the foundation models power applications focused on accessibilities, mental health counseling, and inclusive technologies.

H. Multimodal and Agentic Approaches on the Rise
GPT-5 and other models are multimodal intelligent and agentic, such that they can:
- Visual Question Answering (VQA) - img2txt and txt2img!
- Robotic Process Automation (RPA) – understand commands, orchestrating, and interacting with environment.
- Creative AI co-creators - create text, audio and visual content together.
- Automatic Research Assistants – going out fetching information, figuring it out and writing up reports without intervention from any human.

I. Comparison of some NLP applications

Application Area	BERT-era Models	GPT-4/5-era Models	Distinctive Advancement
Text Classification	Sentiment, NER	Nuanced context, multilingual	Few-shot, multimodal understanding
Question Answering	Factual extraction	Dynamic reasoning, retrieval	Agentic, multi-turn dialogue
Summarization	Extractive/abstractive	Coherent, multi-document	Chain-of-thought, multimodal
Translation	Neural translation	Few-shot, cross-lingual	Multilingual, low-resource adaptation
Code Generation	Limited examples	Complex synthesis & debugging	Autonomous, reasoning-enabled
Knowledge Extraction	Structured relations	Continuous, memory-augmented	Dynamic reasoning, multi-step inference
Human-Centered AI	Basic interaction	Adaptive tutoring, healthcare	Personalization, multimodal inputs

J. Image Placeholder
[Fig. 2: NLP application of pre-trained models]

An illustrative diagram of NLP tasks from text understanding and generation to multimodal reasoning, agentic tasks enabled by GPT-5.

K. Discussion
The wide applicability demonstrates the reflexive power of foundation models. By integrating scalable learning, multimodal capability and agentic reasoning, these systems go beyond the traditional NLP space into that of general cognitive assistants. They empower automation, insight generation, and human–AI collaboration across sectors like education, healthcare, software development, and research.

But along with their power, foundation models introduce ethical and operational challenges, such as addressing bias, fact-checking, privacy-preserving technologies and sustainable computing. Striking the right balance between them is important to maintain that foundation so that models can provide AI services that are trustworthy and socially responsible.

Interpretability, Explainability, and Bias
Pretrained models like BERT, GPT-4, and GPT-5 have shown unprecedented performance on various NLP tasks. However, with their large size, complex structure and opaque architecture, it faces considerable challenges when concerning interpretability, explainability and bias mitigation. Understanding how such models come to a decision is an indispensable part of responsible and trustworthy AI, particularly when used in sensitive domains such as healthcare, legal analysis, finance, education.

A. Interpretability Challenges
Interpretability means being able to understand why a model has output a certain thing. In contrast to traditional ML models that have explicit feature contributions, foundation models are high-dimensional and nonlinear, so direct interpretation might be hard. Key challenges include:

- An over view of Transformer's architecture Given self-attention computation, it is a waste to us eexpensive computing resources to only attend on s ingle pair between two tokens every time One is natural call to attenting all other tokesn to compute attention vecotr(It means that we want to find the relationship betweeen different words). Decisions arise over thousands of parameters rather than being localized to weight except for special weights.
- Contextualized Embeddings: Words and phrases are encoded in context embeddings, that is the same word could have different representations according to the context. This variation is problematic for tracking single-source contributions to output.
- Emergence: Some of the large models, such as GPT-3 – GPT-5 have emergent abilities (learning to perform tasks it was not trained on) without having them explicitly programmed. The origin of these abilities is nontrivial and commonly requires sophisticated probing methods.

B. Explainability Approaches

Explainability seeks to generate human-interpretable justifications for model predictions. There are several methods that are proposed to interpret the foundation models:

- Attention Visualization: Self-attention weights can be visualized to show how tokens or phrases contributed to the output of the model. For example, in sentiment analysis, attention maps can show us what words the model pays attention to when deciding a label. Attention is useful, but it's not always something you're predicting.
- Probing Tasks: Researcher investigating the syntactic, semantic, or factual knowledge encoded in intermediate embeddings. Probing can tell us if models encode linguistic hierarchies, entities relationships, or grammatical rules.
- Saliency Methods: For instance, approaches like Integrated Gradients and LRP assign scores to input tokens that visualize which contributions led to the model's decision. These approaches offer token-level explanations but can be computationally intensive for large models.
- Concept-Based Explanations: TCAV (Testing with Concept Activation Vectors) connects neural mechanisms to a human-provided "high-level" concept, e.g., "bias" or "emotion", thereby enabling interpretability on tasks involving complicated concepts such as disease diagnosis and legal text.
- NAG Neural Explanations: Generative models can generate counterfactual or nearest-neighbor examples which explains how input changes lead to output change making the decision process transparent.

C. Bias and Its Origins

Bias in foundation models comes from training data, architecture and objective functions. Common sources include:

- Data Bias: Models learned from large-scale internet corpora inherit social bias, stereotypes, and minority under-representation. For instance, word embeddings can embed occupations with gender biases.
- Algorithmic Bias: The model that involves the architecture or the learning objective may introduce new biases. Autoregressive models can also learn to promote spurious correlations in multi-step reasoning tasks.
- Deployment Bias: Unbiased models can output skewed results if they are used in out-of-training-data domains. As an example, a chitchat model that is domain-agnostic may misunderstand culture-derived semantic expressions.
- Emergent Bias: Unanticipated behaviors can emerge in large foundation models which reflect small patterns found in the data, such as racial/gender/political biases. GPT-5 inherits its agentic (planning) and reasoning abilities from previous models, and develops new kinds of bias due to multistep planning and tool use.

D. Bias Mitigation Strategies

Various methods that have been proposed to mitigate bias in base models include:

- Data Curation and Augmentation Balancing the distribution of data across classes is important. They include oversampling of under-represented classes, let harmful documents fall out or add with more diverse cultural and linguistic variants.
- DebiasingEmbeddings: Initial approaches modified vector spaces to desaturate gender or race. Two kinds of examples are produced for counterfactual data augmentation in contextual embeddings, parallel examples to squelch stereotypes.
- Fine-Tuning and Alignment: RLHF (Reinforcement Learning from Human Feedback), Sovi et al. '%$0 9 tional AI associating model outputs with ethical guidelines, human values, and fairness targets. GPT-5 is based heavily on these approaches for reducing harmful outputs.
- Algorithmic Fairness Objectives: Training losses can include fairness objectives, ensuring that no group (e.g., demographic parity or equalized odds) is unfairly penalized during prediction.

- Post-Hoc Correction: Output filtering as well as adversarial post-processing can catch biased and harmful responses before they are delivered.

E. Biasevaluation, Auditing of bias

It is not straightforward to assess the bias and fairness of foundation models given that they have emerged behaviors and output multi-modal inputs. Common evaluation methods include:

- Benchmark Datasets: Resources like StereoSet and CrowS-Pairs quantify stereotyping, biased conducted associations in text.
- Scenario Testing: Model sensitive interactions in for instance healthcare, finance or legal to identify biased outcomes.
- Explainability Audits: Mixing interpretability tools with bias detection to audit the source of undesirable outputs.
- Current research stresses continuous auditing and dynamic mitigation which is important as for example a model like GPT-5 learns and tunes over time.

F. Ethical Implications

Opacity, Bias, and Emergent Reasoning in Foundation Models The triad of opacity, bias and emergent reasoning in foundation models presents us with severe ethical challenges:

- Responsibility: It is difficult to establish responsibility for outputs since the model acts as a black-box.
- Transparency: Both users and regulators demand understandable explanations of model decisions for trust and compliance.
- Harm: Underprivileged and sensitive domains such as healthcare, law, and education are at risk for perpetuating stereotypes, discrimination, or misleading views.
- Trust and Adoption: Explainability is essential to gaining a user's trust, as that cannot come about without explainability, thus hindering adoption even though the system may be perfectly capable.
- These problems can only be solved by an integrated approach, involving technical protections, with policy frameworks and ethical AI governance.

G. Comparative Summary Table

Aspect	BERT-era Models	GPT-4/5-era Models	Implications
Interpretability	Moderate (attention maps, probing)	Low due to scale & agentic behavior	Requires advanced visualization & XAI
Explainability	Token-level and layer probing	Concept-based, counterfactual, multimodal	Critical for trust & regulatory compliance
Bias Sources	Training data, embeddings	Training data, RLHF alignment, emergent reasoning	Emergent biases more complex and context-dependent
Mitigation Strategies	Debiasembeddings, curated datasets	RLHF, Constitutional AI, continuous auditing	Multi-layered, dynamic mitigation required
Evaluation	Standardized benchmarks	Scenario testing, dynamic interaction	Continuous monitoring essential

H. Discussion

As base models grow in size, interpretability and reducing bias becomes both more important and more difficult. Although methods such as RLHF, concept based explanations and retrieval-augmented auditing can help in understanding the decision making process and ensuring fairness, a full explaination of GPT-5 level agentic models is ultimately not achievable today given their complexity. They underscore the need for human-in-the-loop supervision, ongoing monitoring and ethical alignment to guarantee that foundation models stay safe, responsible and socially beneficial.

Overall, the necessity to understand and counteract bias, while also improving interpretability is key for responsible deployment of foundation models in real-world NLP applications. If this isn't that, except for time, you're worried about tech governance over content — hey!Nude paintings and statues are fine in your museum of machine-prompted creativity?But without guardrails in place, the potential transformative role of models like GPT-5 is mooted by ethical, social and legal quandaries.

Computational and Environmental Challenges

The recent advancements in deep representation learning in pre-trained models (Rogers et al., 2020; White, Burtenshaw, Ding, Williams, \B-1 Coates, \B-\bb) Xiong et al. and Gross \shortciterogers2020roadblock), such as BERT (Devlin et al.ataires-Sarti et al.), GPT-3 (Brown et al.) and GPT-4/GPT-5 have significantly increased the computational cost of NLP systems. Such models have demonstrated state-of-the-art performance on a wide range of tasks, but due to their resource requirements and carbon footprint represent a major challenge to the scale-out of research, deployment, and 'responsible' AI.

A. Increasing Size of Model Exponential Number of All Possible Programs

One of these significant challenges in computational aspect is the growing number of model parameter:

- BERT-base (2018): ~110 million parameters
- GPT-2 (2019): 1.5 billion parameters
- GPT-3 (2020): 175 billion parameters
- GPT-4 (2023): ~1 trillion parameters (projected)
- GPT-5 (2025): >1.5 trillion parameters (Emergent estimates) And so on and so forth…

This exponential scaling in turn gives rise to quadratic (or higher) memory and compute requirements because of the self-attention mechanism used by transformers. When training at scale, one needs to depend on large performance GPUs / TPUs, distributed training approach and additional memory tricks - all of which makes it high technical cost and expensive.

B. Training: Costs and Resource Utilization

Training foundation models is resource-intensive:

- Compute: GPT-3 was trained using many hundreds of petaflop/s-days of compute. GPT-4 and GPT-5 would require orders-of-magnitude more. It is largely compensated for by distributed parallelism and pipeline optimization but does not overcome the heavy computational burden.
- Energy use: Large-scale training uses a ton of electricity, and studies estimate that training GPT-3 created hundreds (if not thousands) of tons CO_2 equivalent emissions. Its successor GPT-5, with larger parameters and multimodal fusion, is likely higher than these numbers.
- Money Cost: it can cost tens of millions USD to train a model and that puts the AI in hands only for those who are either well-funded corporations, or research institutions [1], limiting democratization of AI research.

Inference Costs: There is a high computational cost of deploying these models in inference to perform real-world NLP tasks, even given trained model weights (especially in the case of optimizing for interactive chatbots, multimodal reasoning, or when- as we shall see below,- running them on agentive tasks), which increases operational and latency demands.

C. Memory and Storage Constraints

Base models require storage in high volume for both parameters, checkpoints, and training data:

- Model Storage: With GPT-3's 175B parameters, weights only require ~350 GB of storage; yet GPT-5 requires more than 1.5 TB, making scaling to such a size unfeasible both in terms of hosting and distributed computation.
- Data Storage: Pretraining entails using on the order of trillions of tokens, and often requires many petabytes worth of text, images, and code. Storing and versioning such datasets require solid storage systems.
- -) GPU/TPU Memory Limitation: High-end GPUs/TPUs are scarce in memory size, and the model needs to be sharded and gradient checkpointed or weighted offloaded across devices.

D. Environmental Impact

The computational requirements of foundation models translate into environmental costs:

- Carbon Footprint: Training big NLP models can require as much power as hundreds of US households because of the energy consumed.
- Cooling and Data Center Efficiency: The amount of heat produced by dense GPU clusters is significant enough to require additional cooling devices, which contribute indirectly to energy use.
- Lifecycle Emissions: Training aside, deployment, maintenance, and updates also add to a model's overall environmental footprint especially at scale across cloud services.

- Role of Green AI The environmental cost has driven the need for green AI i.e, energy efficient algorithms, carbon tracking and sustainable model design.

E. Scalability and Accessibility Challenges

Foundation models are computationally intensive which results in practical limitations to their scalability and accessibility:

- Restricted Research Access: Smaller institutions and emerging markets can't always access the resources to train or customize foundation models, consolidating power in large corporations.
- Operational Challenges: Large-scale, real-time deployment for inference involves making inferences over multi-GPU clusters, and cannot be accommodated on edge or mobile scenarios.
- Reproducibility: It is challenging to reproduce the results of large, foundation models (computational cost and financial barrier make it difficult for other researchers to reinvent large pretrained models), which makes peer validation and scientific transparency hard.

F. Mitigation Strategies

Various strategies are being implemented to tackle the difficulties of computing and from environmental issues:

- Parameter-Efficient Training: Methods such as soft prompts (Cui et al., 2020), LoRA (Low-Rank Adaptation) training, adapter modules and prefix-tuning in Train large pre-trained models with a hundred times fewer parameters while achieving the same or even better accuracy compared to directly fine-tuning36 prior work enable fine-tuning large models with onyne of its own effort37parameters budgets.
- Distillation: Knowledge distillation downsizes large models into small task-specific models while retaining high-quality performance. As an example, DistilBERT takes the full 98% of BERT performance with half the parameters.
- Sparse and Mixture-of-Experts (MoE) Architectures: Enable activation of a fraction of the parameters per input while still maintaining model capacity.
- Efficient data usage: Learn on high-quality trimmed datasets and reduce dataset sizes (curriculum learning, token pruning or selective sampling).
- Environmentally Aware AI Practices: Including the use of renewable energy, carbon tracking and efficient data centers to minimize environmental impact.

G. Trade-offs Among Scale, Performance and Sustainability

Although larger scales and sizes of the model leads to better performance, they also increase computational as well as environmental costs. Key trade-offs include:

Factor	Benefit	Cost / Challenge
Model Size	Emergent capabilities, reasoning, multimodal understanding	Exponential compute & memory demand
Training Data Volume	Better generalization, zero/few-shot learning	Storage, preprocessing, energy consumption
Parameter Efficiency	Reduced compute for adaptation	Potential minor performance degradation
Multimodality	Unified reasoning across text, images, code	Higher data heterogeneity and preprocessing complexity
Continuous Learning	Up-to-date knowledge, memory retention	Persistent compute requirements, lifecycle emissions

H. Discussion

The use of foundation models poses significant computational and environmental classed, representing a central bottleneck to the democratization and long-term viability of AI. Models such as GPT-5 would be state of the art but are infeasible at scale: technically groundbreaking, but they require enormous infrastructure, energy, and money—raising thorny ethical questions about both access and ecological demands. Future research must focus on:

Future work includes more efficient architectures and pruning strategies that can help avoiding unnecessary computation.

- Making sure the data is tidied to avoid repeated and low quality training material.
- Lifecycle carbon tracking to ensure AI benefits and serves the purpose of sustainability.
- AI platforms that can work with smaller organizations as part of the AI innovation drive, without high cost.

The need for such mitigation strategies is critical to responsibly scale foundation models, without further contributing to environmental or social disparities.

I. Image Placeholder

[Fig. 3: Complexity and Resource Overhead of Foundation Models

Caption: Conceptual overview of asset use, parameter growth and environmental costs from BERT to GPT-5.

Conclusion

The area of NLP has seen major change with the introduction of foundation models. From BERT (that popularized deep, contextually uninformed embeddings/bidirectional understanding) to the recent GPT-5 models that exhibit agentic reasoning, multimodal intelligence, and open-endedness of capability - language technologies have a new horizon.

This paper has discussed the evolution, architecture, applications, ethical aspects and computational issues of foundation models in NLP. Some important findings are brought up by this study:

A. Evolution and Comparative Progression

In the evolution from BERT to GPT-5, we move from task-specific models which are fine-tuned to perform a specific set of tasks - such as SQuAD for reading comphrehnsion or CoLA for acceptability judgments - to general-purpose, adaptable models. Where BERT pushed the boundaries of language understanding and classification, and opened many doors that were previously locked with respect to performance, comes along the GPTs and threatens to throw a stone at this paradigm. The comparison reveals the growth of complexity, number of parameter scale, and multimodal integration as we move to higher levels of intelligence.

B. Applications Across Domains

Base models are now used in a broad collection of applications, including text understanding, summarization, question answering, translation and code generation as well as human–AI collaboration and multimodal reasoning. GPT-5 is specifically able to plans autonomously, make agentic decisions, and solve cross-modal problems, leading to revolutionary uses in education, health-care analysis, legal reasoning, science research & creative production.

C. Interpretability, Explainability, and Bias

Foundation models, while being powerful, are lacking in interpretability and ethical alignment. Tools such as attention visualization, probing, concept-based explanations, and human-in-the-loop evaluation offer only partial glimpses of model behavior; agentic models remain essentially inscrutable. Bias, whether that be data-driven or emergent, emphasizes the importance of on-going monitoring and ethical alignment frameworks as well as responsible deployment strategies.ocacy.

D. Computational and Environmental Considerations

It is very resource and energy expensive to train and deploy many of these large foundation models. The novalistic growth in both the size, memory and storage requirements of deep learning models allude to scalability and sustainability issues that can foster interest on parameter-efficient architectures, knowledge distillation, sparsity networks and green AI initiatives. Performing well while being environmentally responsible will be a primary challenge for NLP going forward.

E. Ethics, Governance, and Regulation

The emergence of agency foundation models will require strong governance structures, regulatory oversight and ethical conduct. With the rise of new policy, for example, the EU AI Act and industry standards like those from IEEE, ISO and NIST are emerging to underpin trustworthy AI deployment. Central-périphérique governance best practices include opening the black box of AI, ongoing auditing, human-in-the-loop management and stakeholder participation.

F. Future Research Directions

To generalize and transition these methods to the real world, future work needs tackle scalability, multimodal learning, interpretability, bias mitigation, robustness & safety and societal impact & sustainability. Key opportunities include:

- Efficient and green architecture construction and development.
- Advancing lifelong, multi-modal and cross-domain learning.
- Improving explainability, ethical compliance and regulatory adherence.
- Exploring self-regulatory and autonomy decision-making frameworks for joint action AI systems.

Ensuring that datasets include more Universal from properly distributed, global perspectives in to reduce social bias.

G. Synthesis and Implications

As a convergence of scale, data and algorithmic sophistication from various domains of AI and beyond, these models aspire to what one might call "cognitive" capabilities in terms not only of reasoning but also linguistic ability. Yet, these victories are clouded by moral, social and environmental obligations. Equitable access, transparency, accountability and sustainability will be key to the responsible development of NLP technologies.

How the interplay of technical innovation, ethical oversight, and regulatory consensus will determine if foundation models unspool as agents of optimistic societal transformation rather than vectors for bias, misinformation or environmental damage. The lessons we can learn from GPT-5 and beyond emphasize the importance of interdisciplinary collaboration between AI practitioners, ethicists, policymakers, and industry stakeholders.

H. Final Remarks

In summary, the path from BERT to GPT-5 demonstrates both the potential and complexity of foundation models in NLP. These models have expanded the horizons of what AI can do, allowing systems to comprehend, generate, and reason about many modalities. The next NLP must prioritize scalable, ethical, interpretable and sustainable AI that serves humanity.

The perspectives presented in this paper are an elaborate map for researchers, practitioners, and policy makers who demand to advance NLP such that they face the computational, ethical, and societal aspects of foundation model development and deployment.

I. Placeholder for Image/Table

[Fig. 4: Roadmap of NLP Foundation Models — Evolution, Applications and Future Directions]
Figure 1: Illustrative roadmap to illustrate the evolution from BERT to GPT-5, key applications, ethical considerations, computational challenges, and future research thrust areas.

References

[1] Devlin, J., Chang, M.-W., Lee, K., &Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
[2] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
[3] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).
[4] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5). Journal of Machine Learning Research, 21(140):1–67.
[5] Chowdhery, A., Narang, S., Devlin, J., et al. (2022). PaLM: Scaling Language Modeling with Pathways. arXiv preprint arXiv:2204.02311.
[6] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
[7] OpenAI. (2025). Introducing GPT-5. OpenAI Blog. Retrieved from https://openai.com/index/introducing-gpt-5/
[8] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.
[9] Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report.

[10] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford Institute for Human-Centered Artificial Intelligence (HAI).
[11] Floridi, L., &Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30(4), 681–694.
[12] Bender, E. M., &Gebru, T. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the ACM FAccT Conference.
[13] Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models. arXiv preprint arXiv:2206.07682.
[14] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (RLHF). arXiv preprint arXiv:2204.05862.
[15] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. arXiv preprint arXiv:2203.02155.
[16] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444.
[17] Rogers, A., Kovaleva, O., &Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. Transactions of the ACL, 8, 842–866.
[18] Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.
[19] Thoppilan, R., De Freitas, D., Hall, J., et al. (2022). LaMDA: Language Models for Dialog Applications. arXiv preprint arXiv:2201.08239.
[20] Zhang, S., Roller, S., Goyal, N., et al. (2022). OPT: Open Pre-trained Transformer Language Models. arXiv preprint arXiv:2205.01068.
[21] Menick, J., Irving, G., Pfau, D., et al. (2023). Constitutional AI: Harmlessness from AI Feedback. Anthropic Research.
[22] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019.
[23] Mitchell, M., &Krakauer, D. C. (2023). The Debate Over Large Language Models and Artificial General Intelligence. AI Magazine, 44(1), 25–37.
[24] Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44–58.
[25] Rawte, V., Kalyan, K. S., &Rajasekharan, A. (2023). A Survey on Explainability and Interpretability of Large Language Models. arXiv preprint arXiv:2308.10792.