ChatGPT

OpenAI’s commitment to decoder-only GPT models (2021–2023)

Background – what is a decoder-only GPT model?

Large language models (LLMs) are built on transformer architectures.
The **decoder** component of the original transformer is designed for autoregressive generation: it uses **masked self-attention** so that each position can only attend to previous positions. When stacked, the decoder predicts the next token given the context, making it inherently suited to language modelling and text generation [1] [2].
In practice, GPT-style models discard the encoder and rely solely on **decoder blocks**; they are **decoder-only transformers**. These models are trained on vast corpora of unlabelled text by predicting the next token, and they can be used for many tasks by prompting without changing the architecture. Because the model only looks backward, data from any domain can be used during pre-training, which simplifies data collection and avoids the need for parallel corpora (as required for encoder-decoder models like T5).

Key OpenAI publications shaping GPT (2018–2023)

Year & publication	Authors/affiliation (partial)	Key ideas (short notes)
2018 – Improving Language Understanding by Generative Pre-Training [1] [3]	**Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever (OpenAI)**	Introduced **GPT-1**. Used a multi-layer transformer decoder for unsupervised pre-training and then applied supervised fine-tuning for various tasks. The model was a **12-layer decoder-only transformer with masked self-attention** [3]. The paper demonstrated that generative pre-training plus minimal task-specific conditioning could outperform task-specific architectures.
2019 – Better language models and their implications (blog & GPT-2 releases)	**Radford et al., OpenAI policy & safety team**	Described GPT-2 (up to 1.5 B parameters) and discussed staged release due to misuse concerns. The architecture remained a **decoder-only transformer**, scaled up from GPT-1.

Year & publication	Authors/affiliation (partial)	Key ideas (short notes)
Jan 2020 – Scaling Laws for Neural Language Models	**Jared Kaplan, Sam McCandlish & colleagues (OpenAI)**	Studied how language-model performance scales with model size, data and compute. Found **predictable power-law scaling** and noted that performance depends mostly on scale rather than architectural details; depth vs. width and other hyper-parameters had **"minimal effects"**. This encouraged OpenAI to prioritize scaling decoder-only models rather than exploring more complex architectures.
May 2020 – Language Models are Few-Shot Learners (GPT-3) [4]	**Tom B. Brown, Benjamin Mann, Nick Ryder et al. (OpenAI)**	Introduced **GPT-3**, a 175-billion-parameter autoregressive (decoder-only) model. The authors noted that despite numerous innovations—bidirectional denoising, prefix-LMs, encoder-decoder architectures, efficiency improvements—**they deliberately "continue to focus on pure autoregressive language models" to study in-context learning and simplify large-scale implementations** [4]. GPT-3 showed strong zero-/few-shot performance using only prompts, illustrating emergent in-context learning abilities.
2021 – research on scaling and safety	**OpenAI researchers**	Throughout 2021, OpenAI continued exploring scaling and safety. Papers such as Scaling Laws for Neural Language Models and the GPT-3 paper were widely discussed; there were no major architecture changes. The focus remained on improving scaling efficiency and understanding emergent abilities.
Mar 2022 – Training language models to follow instructions with human feedback (InstructGPT) [5]	**Long Ouyang, Jeff Wu, Xu Jiang et al. (OpenAI)**	Addressed misalignment between next-token prediction and user intentions by **fine-tuning GPT-3** using **reinforcement learning from human feedback (RLHF)**. The base architecture remained GPT-3's **decoder-only transformer**; the key innovation was the training pipeline (supervised fine-tuning plus reward-model-guided RLHF). The resulting **InstructGPT** models, despite having far fewer parameters, were preferred by human evaluators over the original 175 B model [5].

Year & publication	Authors/affiliation (partial)	Key ideas (short notes)
Mar 2023 – GPT-4 Technical Report [6] [7]	**OpenAI**	Introduced **GPT-4**, a multimodal model that accepts image and text inputs but still uses a **Transformer-style model pre-trained to predict the next token** [6]. Because of competitive and safety considerations, architectural details were not disclosed; however, the report states that the model is **pre-trained to predict the next token** and then **fine-tuned using RLHF** [7]. The authors emphasise building infrastructure and optimization methods that scale predictably.

Why OpenAI remained committed to decoder-only GPT during 2021–2023

1. **Simplicity and generality.**
   A decoder-only model uses a single network to represent both context and output and is trained with a straightforward objective: predict the next token. Unlike encoder-decoder models that require separate encoder and decoder networks and parallel training data, the decoder-only architecture can be pre-trained on any unlabelled text. In the GPT-1 paper, Radford et al. chose a **multi-layer transformer decoder** [1] and demonstrated that with minimal task-specific changes, the same model can be fine-tuned for various tasks. This simplicity makes scaling easier.

2. **Natural fit for autoregressive text generation.**
   Decoder-only transformers are specifically tailored for generative tasks; they consist of stacked decoder layers with **masked self-attention** so each position attends only to previous tokens [2]. This training objective naturally matches the way language is generated and allows the model to learn long-range dependencies [8]. Encoder-decoder models are more appropriate for conditional generation (e.g., translation), whereas OpenAI's goal was to build a general-purpose generator.

3. **Empirical success of scaling.**
   Scaling Laws for Neural Language Models (Kaplan et al., 2020) found that language-model performance improves predictably as model size, data and compute scale up, and that **architectural details such as network width or depth have minimal effects within a wide range**. These findings suggested that the simplest path to better performance was to scale up the existing autoregressive architecture rather than develop a new one.

4. **In-context learning and emergent abilities.**
   The GPT-3 paper emphasised studying in-context learning. Despite numerous innovations in architectures, the authors **chose to focus on pure autoregressive models to explore in-context learning and reduce implementation complexity** [4]. GPT-3 demonstrated that very large decoder-only models acquire surprising few-shot and zero-shot abilities without gradient updates, reinforcing confidence in this architecture.

5. **Alignment can be achieved at the training level rather than architecture.**
   By 2022, it became clear that scaling alone did not align models with human intent. The

InstructGPT paper showed that a **decoder-only model can be aligned with human preferences via supervised fine-tuning and RLHF** [5]. This success signalled that improvements in behaviour and safety can be achieved by **changing training objectives and data** rather than altering the core architecture.

6. **Infrastructure for scaling.**
   The GPT-4 technical report highlights that a major challenge was developing **infrastructure and optimization methods that behave predictably at large scales** [9]. The ability to extrapolate performance from smaller models and to manage compute efficiently allowed OpenAI to train larger decoder-only models. The report notes that GPT-4 is still **pre-trained to predict the next token** and fine-tuned with RLHF [7], indicating that architectural continuity made it easier to leverage existing tooling and research.

7. **Continuity in research team and expertise.**
   The authors across these publications (Radford, Brown, Kaplan, Ouyang, Wu, Sutskever, etc.) overlap significantly. Many of them contributed to multiple GPT releases and the scaling laws studies [10]. This continuity fostered deep expertise in autoregressive modelling and a consistent research agenda, reinforcing confidence in the decoder-only approach.

Insights gained from the publication trajectory

OpenAI's publications from 2018 to 2023 reveal a deliberate research strategy:

- **Start simple:** GPT-1 used a multi-layer transformer **decoder** and showed that generative pre-training plus simple task-specific fine-tuning could beat specialised architectures [3]. This proof of concept encouraged further exploration.

- **Scale up:** GPT-2 and GPT-3 scaled the decoder-only model to billions of parameters. The scaling laws paper provided **theoretical and empirical justification** for scaling rather than altering the architecture.

- **Observe emergent abilities:** GPT-3 revealed strong zero-/few-shot performance and emergent in-context learning, reinforcing the idea that decoder-only models, when scaled, develop remarkable capabilities [4].

- **Address alignment:** InstructGPT introduced RLHF to align model outputs with human intent without changing the base architecture [5]. This demonstrates that training objectives and human feedback can shape behaviour more effectively than adding architectural complexity.

- **Extend to multimodal tasks:** GPT-4 adds image input but remains a **Transformer-style model trained to predict the next token** [6]. The continuity of the decoder-only approach allowed OpenAI to integrate multimodal inputs without discarding the foundation.

Conclusion

Between 2021 and 2023, OpenAI continued to build larger and more capable GPT models while maintaining a **decoder-only, autoregressive architecture**. Their publications show a clear rationale: the decoder-only model is simple, effective for generative tasks, and scales predictably. Empirical evidence from scaling laws and GPT-3's in-context learning suggested that scaling yields greater gains than

architectural changes [4]. Instead of changing the core architecture, OpenAI focused on improving training objectives (e.g., RLHF), developing infrastructure for scaling, and exploring alignment and safety. This consistent approach explains why, during 2021–2023, OpenAI remained confident in the decoder-only GPT pathway.

---

[1] [3] language_understanding_paper.pdf
https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[2] [8] Towards Smaller, Faster Decoder-Only Transformers: Architectural Variants and Their Implications
https://arxiv.org/html/2404.14462v2

[4] [10] 2005.14165.pdf
https://arxiv.org/pdf/2005.14165.pdf

[5] Training_language_models_to_follow_instructions_with_human_feedback.pdf
https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf

[6] [7] [9] gpt-4.pdf
https://cdn.openai.com/papers/gpt-4.pdf