✅ Narration: The First Initiative Papers of LLMs

(Clear, engaging, chronological storytelling)

1) Word2Vec (Mikolov et al., 2013) — "Meaning hides in numbers"

Narration:
In 2013, something very strange happened.
Researchers fed text into a simple neural network… and discovered that words could live as **vectors** in space.

Not just any vectors — **semantic vectors**.

- "king – man + woman ≈ queen"
- "Paris – France + Italy ≈ Rome"

For the first time, machines could **feel meaning** mathematically.

Why this matters:
Word2Vec proved language can be **compressed** into geometry.
It was the first real hint that math alone might capture human concepts.

This paper lit the first spark.

2) Sequence-to-Sequence (Sutskever, Vinyals, Le, 2014) — "Let the machine translate thought"

Narration:
How do you turn one sentence into another?
English → French, question → answer, instruction → result.

Enter **Seq2Seq**: two RNNs, one encoding text into a hidden thought, another decoding it back out.

For the first time:

- Translation worked
- Summaries worked
- Chatbots got less stupid

This wasn't just prediction —
it was **transformation**.

3) Attention Mechanism (Bahdanau et al., 2014) — "Focus matters"

Narration:
Seq2Seq had a flaw: long sentences made it forget the beginning.

Humans don't read that way — we **attend** to important parts.

So researchers added a simple idea:

> At each step, look again at the input and focus where it matters.

This tiny idea — **attention** — was not tiny at all.
It made machines:

- remember long sentences
- align words across languages
- understand context

This was the seed that later grew into something enormous.

4) Transformer (Vaswani et al., 2017) — "Attention is all you need"

Narration:
2017 is the year everything changed.

A radical claim:

> Forget RNNs entirely.
> You don't need recurrence at all.
> **Attention alone** can model language.

It sounded wrong.
But it worked better than anything before.

Transformers could:

- read in parallel (fast)
- understand long-range meaning
- scale to huge sizes

This paper is the **birth certificate** of modern LLMs.

From this, every major model descends:
BERT → GPT-1 → GPT-2 → GPT-3 → GPT-4 → GPT-5

5) BERT (Devlin et al., 2018) — "Let the model read both directions"

Narration:
If a human reads a sentence, they don't guess the next word one by one.
They use the **whole context**, left and right.

BERT did the same.
It masked random words and forced the model to fill them in.

This trained language understanding that was:

- deep
- contextual
- incredibly useful

Suddenly, BERT crushed every NLP benchmark.
Industry realized: **Transformers are real.**

6) GPT-1 → GPT-2 → GPT-3 (Radford et al., 2018-2020) — "Just scale it"

Narration:
While BERT was mastering understanding, OpenAI asked a different question:

> What if we just predict the next word…
> but with a **huge** Transformer?

GPT-1 was small.
GPT-2 shocked people — fluent essays, reasoning, code.
GPT-3 was bigger again… and suddenly felt like an **intelligent assistant.**

The key discovery:

- You can teach a model almost anything
- simply by showing examples in natural language

This was **in-context learning** — instructions without retraining.

It felt like magic, but it was scaling law + transformer structure.

7) Scaling Laws (Kaplan et al., 2020) — "Predictable intelligence"

Narration:
A strange graph appeared in 2020.

As you increase parameters, data, and compute in the right ratios,
performance rises in a clean curve.

No randomness.
No mystery.

Just **math**.

This paper changed AI from research gamble → engineering roadmap.

Investors saw it.
Tech companies saw it.
Governments saw it.

AI wasn't a toy anymore —
it was predictable growth.

8) GPT-4 & Beyond — "Generalization and reasoning emerge"

Narration:
With scale, something unexpected happened.

Models didn't just learn language.
They learned:

- logic
- coding
- reasoning
- multi-step thinking
- tool use

Nobody hard-coded this.
It **emerged** from scale.

This leads to today:
AI that solves problems, writes proofs, plans, and acts like a reasoning engine.

✅ Summary Slide (Use at the end)

Paper	Discovery	Why it mattered
Word2Vec (2013)	Meaning in vectors	First geometric understanding of words
Seq2Seq (2014)	Transform one sequence to another	Translation, Q&A, chat
Attention (2014)	Focus on important info	Long-range understanding

Paper	Discovery	Why it mattered
Transformer (2017)	Attention only	Fast, scalable, parallel
BERT (2018)	Bidirectional understanding	Language comprehension
GPT (2018–2020)	Predict next word at scale	Emergent reasoning
Scaling Laws (2020)	Bigger = predictably smarter	Industrialization of AI