ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

Lecture 7a – Language AI, Embeddings, and Large Models

Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences

Spring 2026

Today's Question

What we are trying to answer

How did AI move from brittle text processing to systems that can summarize reports, translate across languages, answer questions, and help people write?

Why this matters in a general course

Language is where many students first meet modern AI directly. If we only see the interface, it feels magical. If we see the pipeline, it starts to make sense.

What changes from AI6

AI6 focused on action and feedback. AI7 shifts to language: the challenge is no longer choosing a move in a game, but handling words, meaning, context, and uncertainty.

From AI6 to AI7

Last time

We asked how a system can learn from consequences over time.

Today

We ask how a system can deal with human language, where the same word can mean different things and the same sentence can change meaning when the order changes.

One sentence

AI6 was about learning what to do. AI7 is about learning what we mean.

Why Language Became the Public Face of AI

Why language moved fast

- Text data already existed at huge scale
- Collecting text was easier than collecting robot actions
- Training on text could happen offline
- Users could immediately test the results in conversation

What made this visible

When a model writes, translates, or answers, people feel the result directly. That is why language AI became the most public demonstration of modern AI progress.

The Turing Test Changed the Conversation

The original public question

If you could not see the machine, could it keep a conversation going well enough to pass as human?

Why this question stayed famous

Language is where people notice intelligence fastest. We infer memory, reasoning, and social skill from dialogue almost automatically.

What changed now

Today the question is broader. We care not only about whether a system can talk, but whether it can help, explain, and avoid misleading people.

Why Language Is Harder Than It Looks

Human language hides a lot

- one word can have several meanings
- tone and context change interpretation
- we leave important things unsaid
- order changes the whole sentence

A simple reminder

People resolve these problems almost automatically. A machine has to learn how to use nearby clues and broader patterns from data.

A Tiny Ambiguity Example

Sentence

“I saw the man with a telescope.”

What is unclear

Did I use the telescope, or did the man have the telescope?

Why this belongs in AI7

This is not a rare corner case. Real language is full of missing context, shortcuts, and ambiguity.

Why Rules Hit a Wall

Early idea

Write enough grammar rules, dictionaries, and templates by hand, then let the system follow them.

Why that breaks

Real language is too varied, too noisy, and too tied to context. Hand-written rules can help in narrow settings, but they do not scale to open language.

That is why the field moved from rule systems toward data-driven learning.

The Hidden Pipeline Behind Language AI

What users see

question → answer

What the model actually needs

text → tokens → vectors → probabilities → output

Why this matters

Before a model can do anything impressive with language, it first has to convert language into numbers that a neural network can work with.

Why Tokenization Exists

Why not feed full sentences directly

- sentence lengths vary
- the vocabulary is huge
- new words appear constantly
- rare words make storage and learning inefficient

Tokenizer job

Break text into reusable pieces that are manageable for the model.

One Sentence, Several Cuts

Sentence

“Artificial intelligence changes daily life.”

One possible tokenization

Artificial | intelligence | changes | daily | life | .

Subword view

engi | neer | ing
or some other learned split

Why this helps

Subwords let the system reuse pieces across many words, including words it has never seen exactly before.

Order Changes Meaning

Compare these two

- “dog bites man”
- “man bites dog”

What students should notice

The words are the same, but the meaning changes because the order changes.

The big point

Language AI must model both content and order. Knowing which words appear is not enough.

Prompting Is Interface Design

Weak prompt

“Explain AI.”

Stronger prompt

“Explain AI to first-year students in 120 words and give one everyday example.”

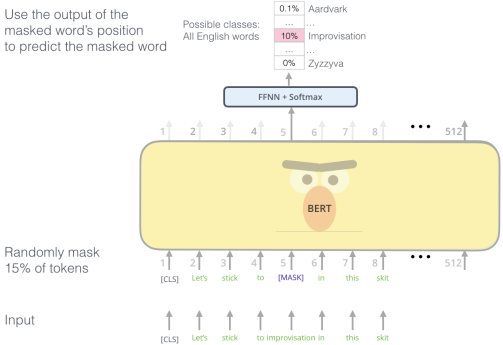
Why this is practical

The same model can look smart or unhelpful depending on how clearly the task is framed.

What the user is really doing

Prompting is not magic. It is a way of specifying task, audience, format, and constraints.

Masked Word Learning



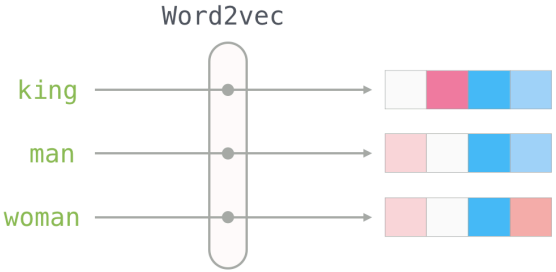
Training idea

Hide some words and ask the model to predict them from the surrounding context.

Why this was useful

It taught models to build contextual representations instead of treating each word as a fixed symbol with one fixed meaning.

Words Become Coordinates



Core idea

Each token gets mapped to a vector, which you can think of as a point in a high-dimensional space.

Why that helps

Now the network can compare, combine, and transform language using ordinary numerical operations.

A Semantic Map

A campus-language example

Words like “dorm”, “canteen”, “meal card”, and “semester” often appear in similar contexts, so their vectors tend to live in a nearby neighborhood.

A different neighborhood

Words like “deadline”, “submission”, “grading”, and “revision” form another neighborhood because they often travel together in student questions.

What the geometry means

The model does not store a hand-written dictionary. It learns that similar context usually implies similar position in vector space.

Distance Means Similarity

One common measure

$$\mathrm{sim}(x, y) = \frac{x^\top y}{\|x\|\|y\|}$$

Read in words

If two vectors point in a similar direction, the corresponding words are often used in similar contexts.

Why this matters

Similarity search, retrieval, clustering, and recommendation all rely on the fact that meaning can be represented as geometry.

Adding New Words

A natural question

What happens when a new word, slang term, or domain-specific phrase appears?

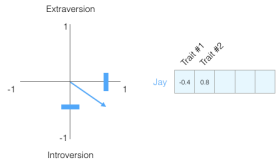
What usually happens

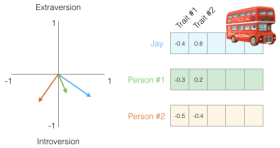
The model places it by context. If the new term appears near similar words often enough, its vector drifts toward a useful neighborhood.

A familiar example

When a campus suddenly starts using a new abbreviation, students understand it because of surrounding context. Models do something similar, but numerically and imperfectly.

Why More Dimensions Help





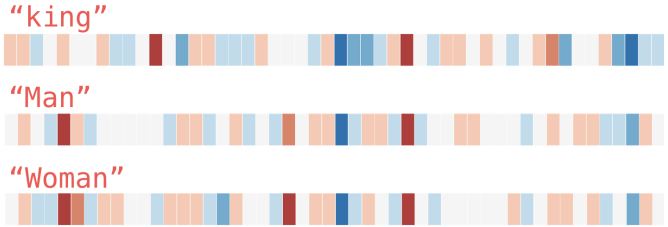
Why low dimensions run out of room

A few dimensions may capture a rough idea, but real language carries many overlapping properties at once: topic, tone, grammar, sentiment, and domain.

What high dimensions buy us

More dimensions give the model more space to place words so that several kinds of similarity can coexist instead of colliding.

Analogy in Space



Classic pattern

“king - man + woman” ≈ “queen”

Why people found this exciting

Simple vector arithmetic seemed to capture some regularities of meaning.

What not to overclaim

This does not mean the model truly understands language in a human sense. It means the geometry has learned some useful structure.

Embeddings Are Useful, Not Magic

1- We can represent things (and people) as vectors of numbers
(Which is great for machines!)

Jay	-0.4	0.8	0.5	-0.2	0.3

2- We can easily calculate how similar vectors are to each other

The people most similar to Jay are:

	cosine_similarity ▼
Person #1	0.86
Person #2	0.5
Person #3	−0.20

Reality check

Static embeddings alone do not solve long context, deep reasoning, or factual reliability.

Why the field moved on

Embeddings are the starting representation. Modern language AI needed better sequence modeling and much larger training.

Why Next-Token Prediction Scaled So Far

Core objective

$$P(w_t \mid w_1, \ldots, w_{t-1})$$

Predict the next token from the tokens that came before.

Why such a simple task works

To predict the next token well, the model must learn grammar, style, common facts, topic structure, and many patterns of reasoning.

The training target is simple. The behavior that emerges can still become surprisingly rich.

Why Next-Token Prediction Felt Surprising

What surprised people

A model trained to continue text can also summarize, rewrite, translate, draft emails, and answer questions.

What this means

Many language tasks can be reframed as producing the next useful sequence of tokens under the right context.

A classroom way to see it

“Explain this page,” “rewrite this paragraph,” and “draft a polite reply” all look different to us, but to the model they are all continuation tasks with different context.

Three Training Stages

Stage 1: pretraining

The model reads enormous amounts of text and learns broad patterns of language, topic, and structure.

Stage 2: task tuning

It is then pushed toward being useful on tasks people actually care about, such as instruction following, formatting, and clearer answers.

Stage 3: alignment

Finally, the system is adjusted so it is more helpful, less chaotic in interaction, and less likely to produce some unsafe behavior.

What Pretraining Gives

Why pretraining matters

- broad coverage of language patterns
- reusable representations
- transfer to many downstream tasks
- better performance with less task-specific data

But not enough

Pretraining gives a powerful base model. It does not automatically make the model helpful, safe, or reliable for a specific application.

Why Instruction Tuning Matters

Without instruction tuning

The model may continue text in a plausible but unhelpful way because it treats the prompt mainly as more text to imitate.

With instruction tuning

The model becomes better at following user intent, formatting answers, refusing some unsafe requests, and staying on the requested task.

The base model learns language. Tuning teaches it how to behave in interaction.

Why Context Matters

A practical truth

The same question can deserve a different answer depending on what came earlier in the conversation or what document the user is asking about.

Student example

“Explain this equation” means very different things if the earlier conversation was about physics, economics, or image classification.

The teaching point

Language models do not answer in a vacuum. They answer relative to the context window they are given.

Why Retrieval Helps

The problem

A model may sound fluent even when it is unsure, outdated, or missing domain-specific facts.

The retrieval idea

Instead of asking the model to remember everything internally, fetch relevant external passages first and let the answer depend on those passages.

A student example

If you ask about a course policy, the best answer should come from the current syllabus or official notice, not from the model's vague memory of similar courses.

Where Language AI Helps

Study and daily work

- summarize readings
- draft emails and reports
- translate and rewrite
- organize scattered notes

Creative support

Brainstorm titles, compare writing styles, generate outlines, or turn rough ideas into clearer first drafts.

How Students Should Use It

Good use

Ask for explanations, examples, rewrites, feedback on structure, or a list of gaps in your reasoning.

Bad use

Do not let the model replace your own checking, your own source reading, or your own responsibility for submitted work.

Rule of thumb

Use language AI as a fast assistant, not as an unquestioned authority.

Hallucination Is a Real Issue

What users see

The answer sounds smooth, detailed, and confident.

What may actually be happening

The model is generating a plausible continuation, not checking facts the way a database or a careful human expert would.

Why this matters

Fluency can hide error. That is dangerous in medicine, law, finance, science, and education.

Fluent Is Not Correct

Common misunderstanding

“It sounds confident, so it must know.”

Better rule

Treat a fluent answer as a draft that still needs checking, especially when the cost of error is high.

Class habit

Ask where the claim came from, what evidence supports it, and what would count as a contradiction.

Bias, Privacy, and Safety

Bias does not disappear

If the training data contains stereotypes, blind spots, or imbalance, the model can reproduce them.

Privacy also matters

Do not casually upload private records, confidential company material, or sensitive personal data into systems whose storage or policy you do not understand.

Practical consequence

Good deployment is not only about model quality. It is also about policy, logging, access control, and human review.

Evaluation Before Deployment

What teams should test

- accuracy on real tasks
- failure cases and edge cases
- robustness to vague or adversarial prompts
- privacy and safety risks

Why this belongs in AI7

A good demo is not enough. Once a model affects real users, the evaluation standard has to become much stricter.

Why AI7 Leads to NN7

What we did today

We looked at tokenization, embeddings, large-model training, retrieval, and the real-world strengths and risks of language AI.

What is still missing

We have not yet opened the neural machinery that made modern language models scale so well.

Next bridge

NN7 will explain attention, transformers, and why they replaced earlier sequence models as the main engine of large language models.

Summary

- Language AI became central because text data was abundant and language is how people directly experience model behavior.
- The hidden pipeline is text to tokens to vectors to probabilities, not just question to answer.
- Embeddings turn meaning into geometry, which makes similarity and retrieval possible.
- Large language models grew powerful through next-token prediction, large-scale pretraining, and later tuning.
- Retrieval, checking, and judgment matter because fluent language is not the same thing as reliable truth.

Next Time

Where the course is going

We now turn from the user-facing story of language AI to the model architecture story.

Next lecture

We will study attention, transformer blocks, masking, generation, and why this architecture changed the economics of modern AI.

Thank You