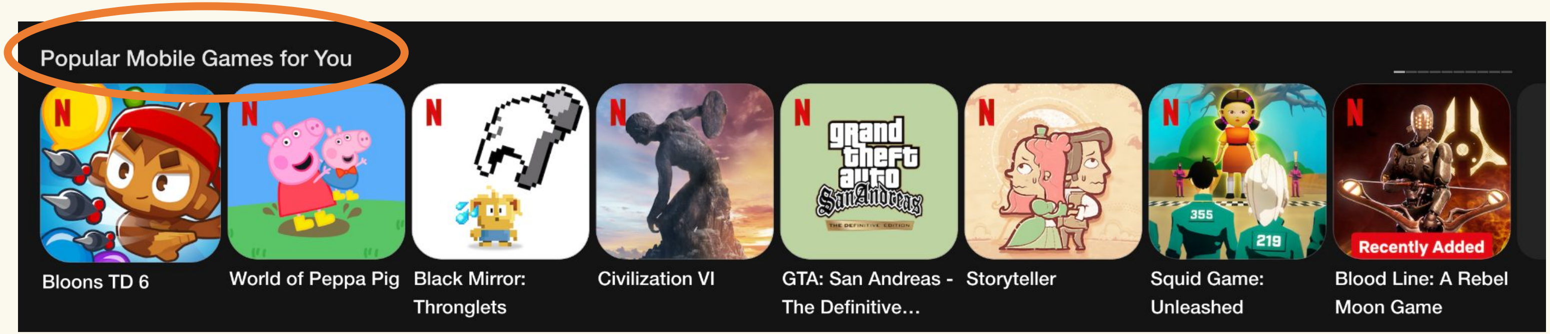
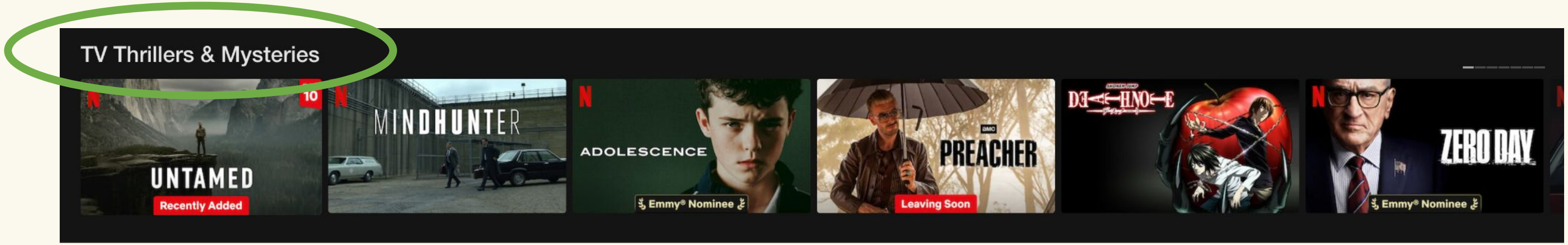
Generative Recommenders using LLMs

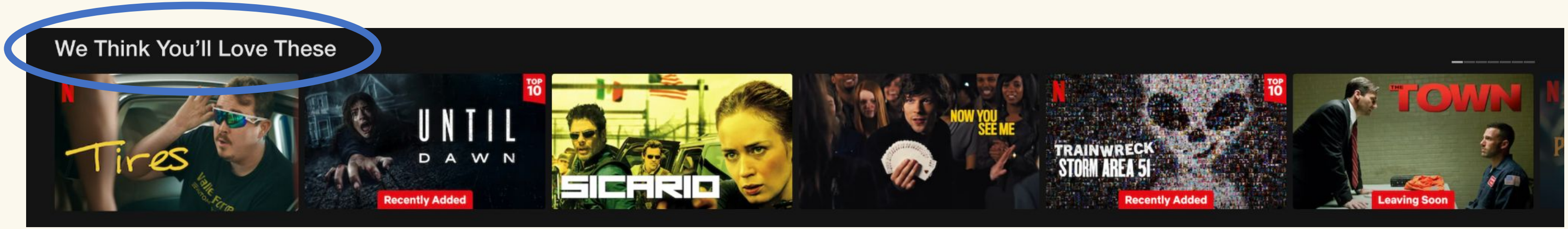
Rein Houthooft

KDD 2025 OARS Workshop 08/03/25

NETFLIX


Popular Mobile Games for You
Bloons TD 6
World of Peppa Pig
Black Mirror: Thronglets
Civilization VI
GTA: San Andreas - The Definitive...
Storyteller
Squid Game: Unleashed
Recently Added
Blood Line: A Rebel Moon Game


TV Thrillers & Mysteries
UNTAMED
Recently Added
MINDHUNTER
ADOLESCENCE
Emmy® Nominee
PREACHER
Leaving Soon
ZERO DAY
Emmy® Nominee


We Think You'll Love These
Tires
UNTIL DAWN
Recently Added
SICARIO
NOW YOU SEE ME
TRAINWRECK STORM AREA 51
Recently Added
THE TOWN
Leaving Soon

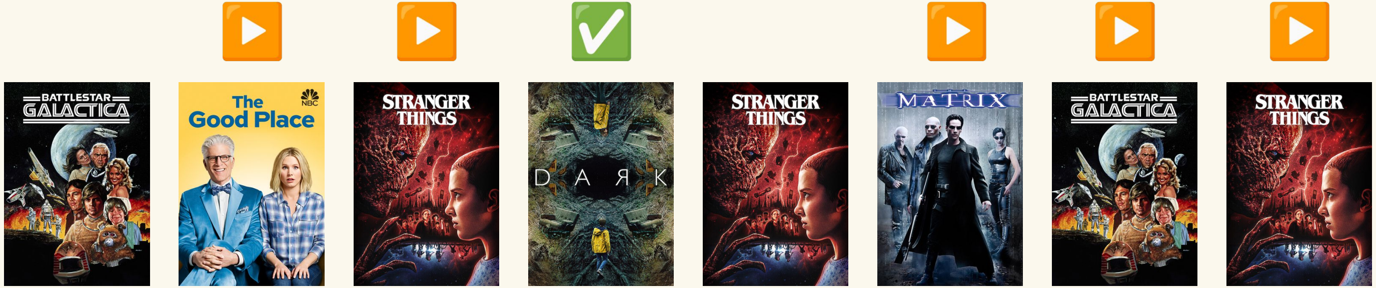
History

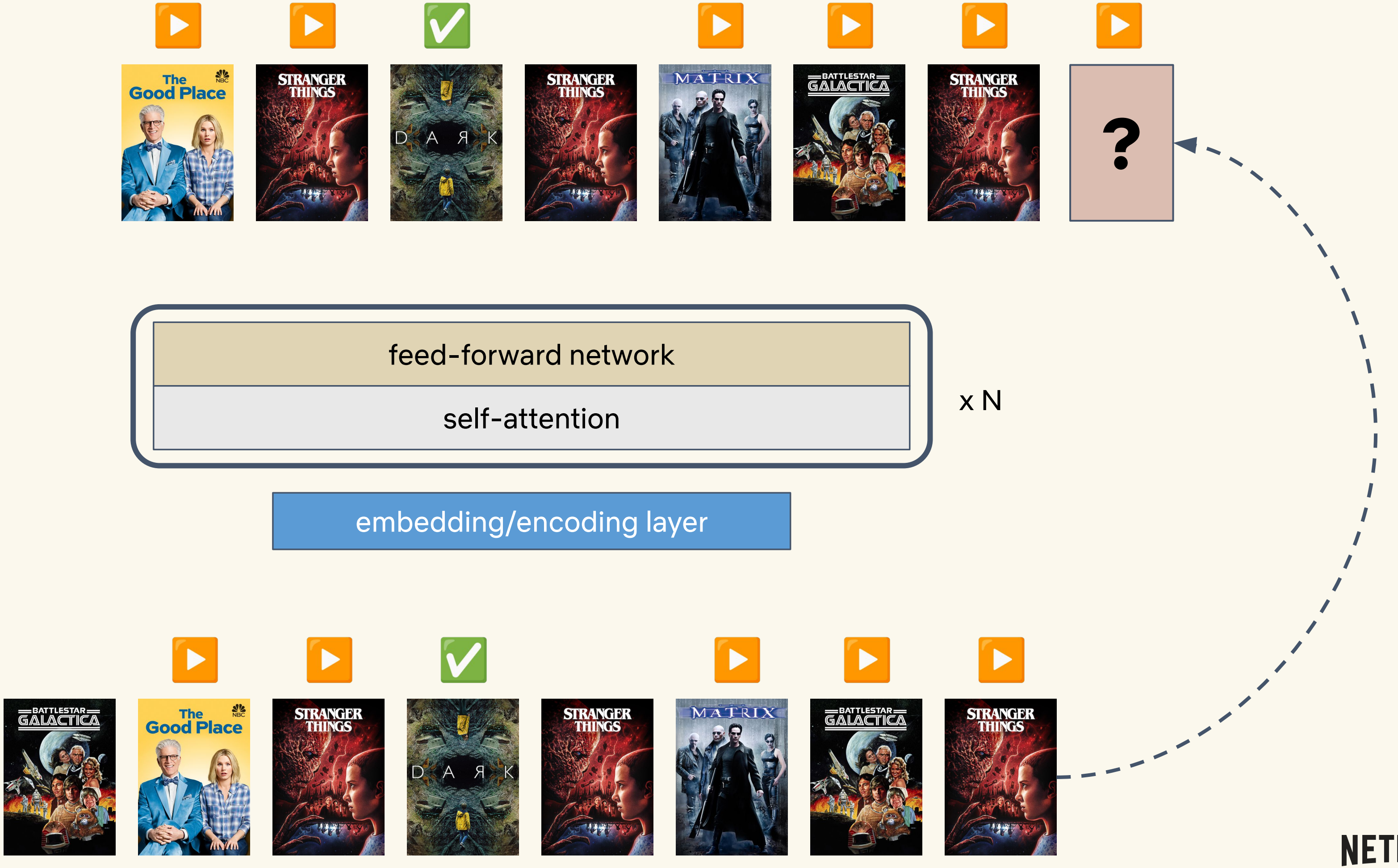
- Many different **specialized models**
  - Row ranking, continue watching, popularity feed, games, …
  - Two-tower structures, matrix factorization, trees, …
  - Various objective functions
- Manual **feature engineering**
  - Various counts (e.g., plays, impressions)
  - Similarity scores between entities
  - Popularity measures
  - Release year
  - …
- **No sharing** across models

NETFLIX

Towards generative recommenders

- Step towards **generative recommenders**
- **Sequence** of tokenized entities + metadata
- **Transformer** trained autoregressive
- Plugs into various specialized downstream models




?
feed-forward network
self-attention
x N
embedding/encoding layer

Connecting LLMs

- LLMs have **vast knowledge** sourced from the internet
  - Netflix titles are often mentioned online (reviews, synopsis, discussions)
  - General world understanding: “what is comedy?”
- LLMs can ingest **heterogeneous data**
  - Storylines, plots, artwork, videos, actor data, …
- LLMs should work well for **generating recommendations**
- LLM recommendations **not aligned** with our actual user behavioral data
  - e.g., kids profile → kids movie, but low likelihood of watching

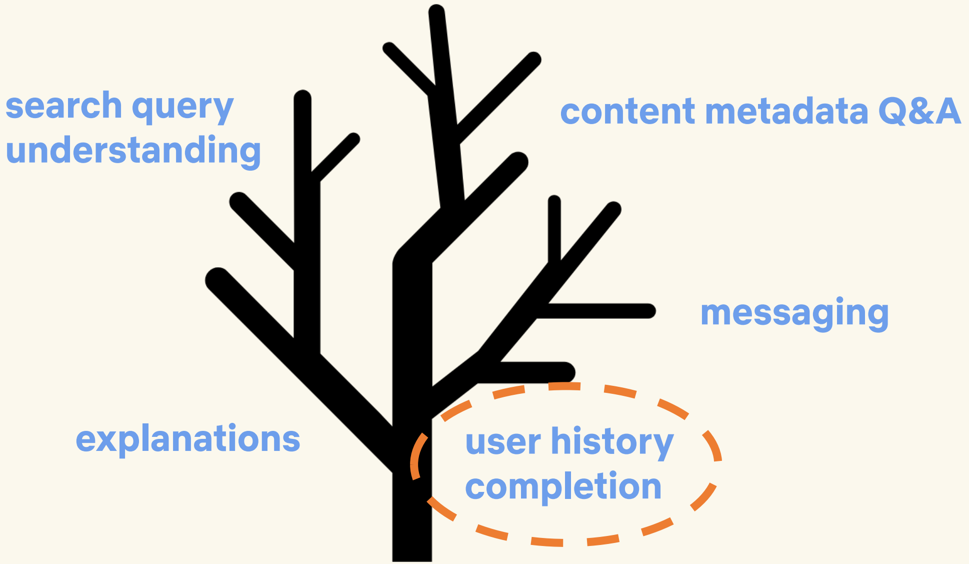
Connecting LLMs

Big research direction at Netflix:

- Train an LLM achieving **quantitative** performance > current models (e.g. MRR, AUC)
- while opening up new **qualitative** abilities (e.g. general Q&A, homepage generation)
- by **blending world knowledge** with Netflix-specific user data

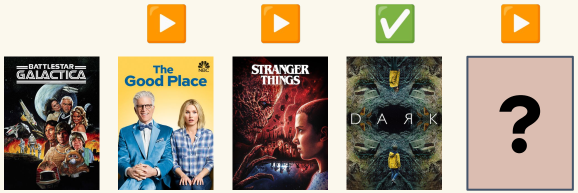
Model stages

- 3 model stages
  - Base LLM: open-source/in-house trained for language understanding
  - Netflix LLM: base LLM SFT on Netflix data
  - **GenRec**: Netflix LLM SFT/post-training as a recsys



NETFLIX

Ranking



encoding

User event interaction history:

`[thumbs-up vid239, play vid683, play vid130, bookmark vid859, play ?]`

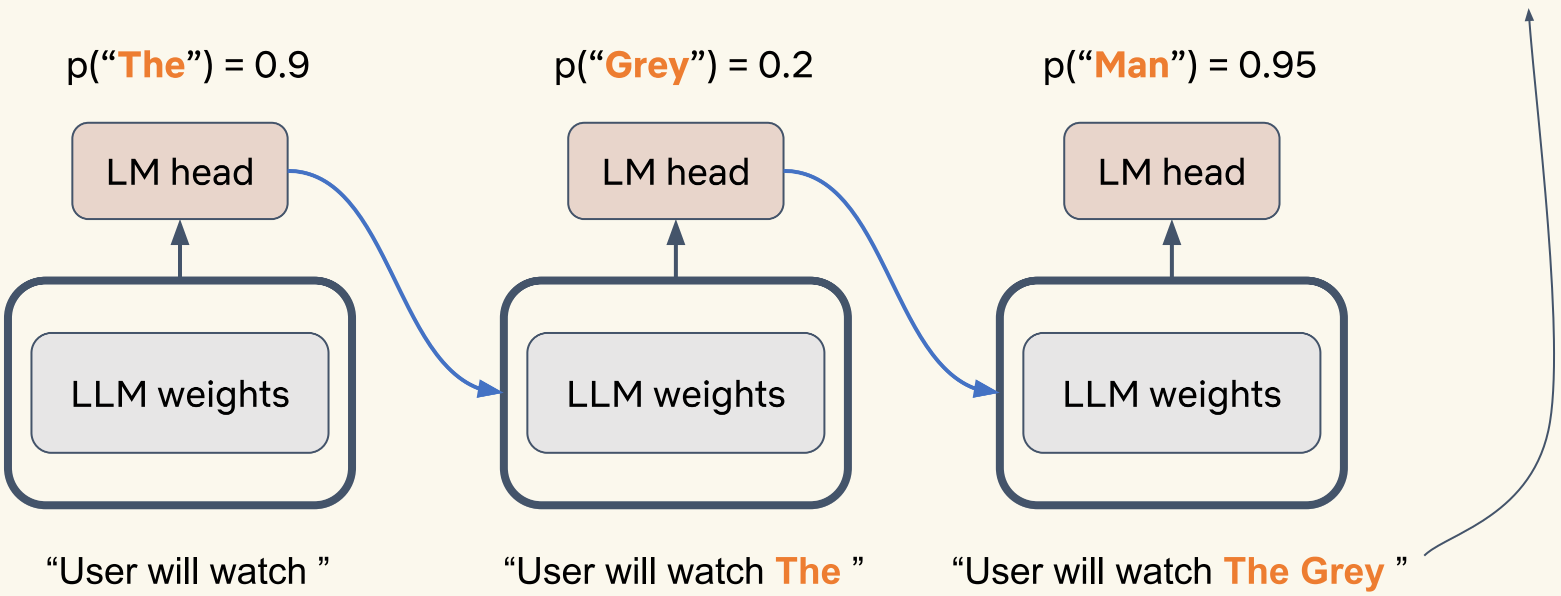
verbalization

Saturday 2PM 30 days ago thumbs-up Battlestar Galactica, Monday 1PM 29 days ago play The Good Place, **Wednesday 7PM 27 days ago play Stranger Things**, Wednesday 11AM 12 days ago bookmarked Dark

Friday 9AM 10 days ago**, what will the user watch next?**

Ranking

- LM head trained with CE for next-token prediction

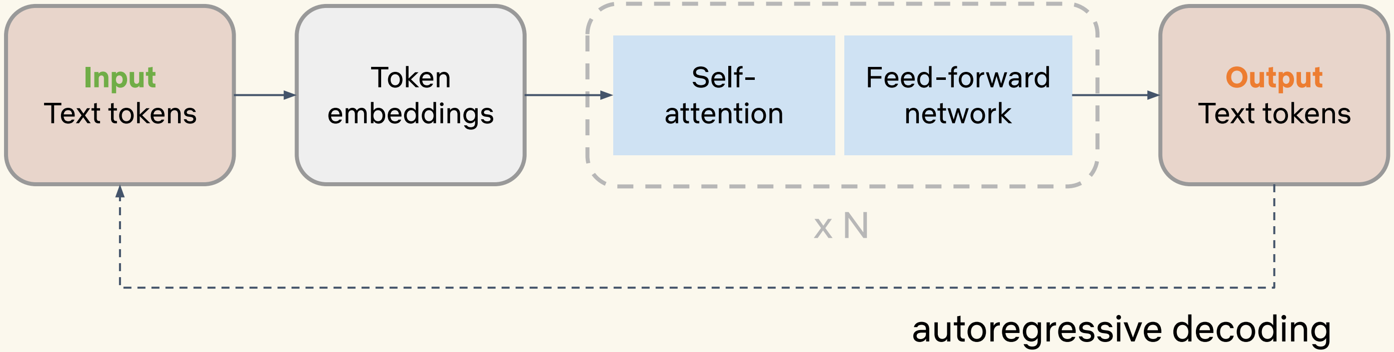


Ranking

- Problem with autoregressive text generation
  - How to generate catalog **ranking**?
  - How to **evaluate** on ranking metrics (eg MRR)?

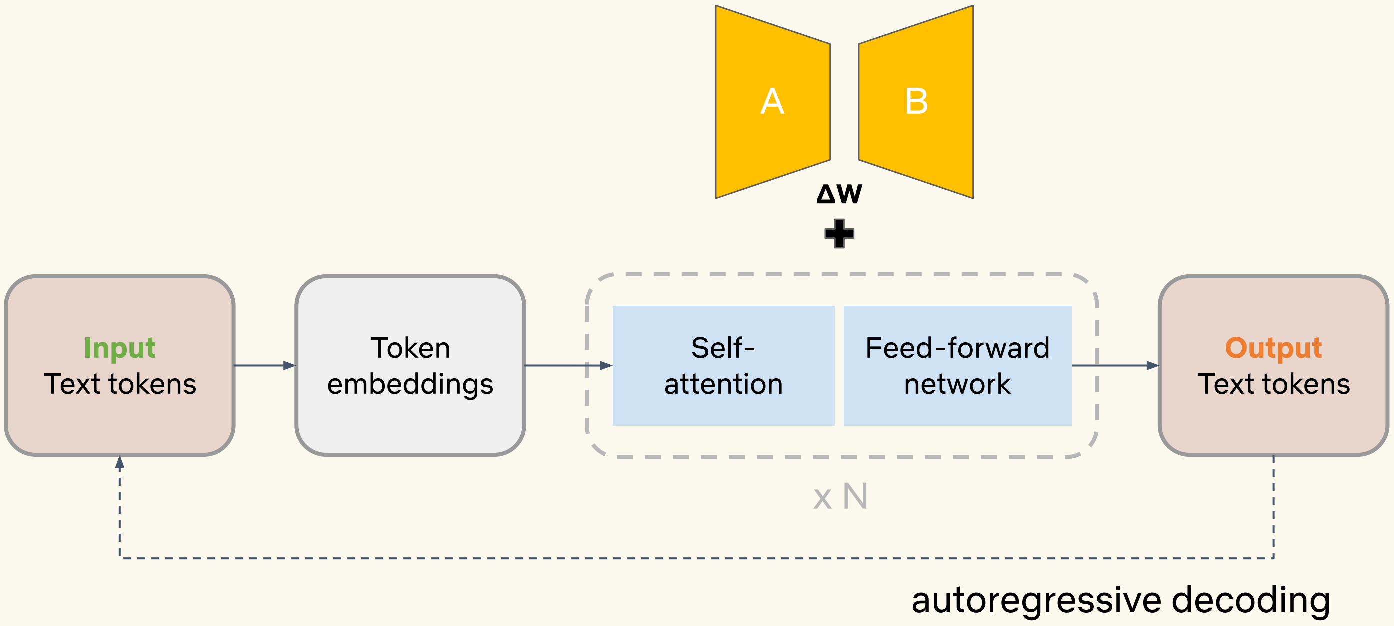
- Expand LLM **token vocabulary** with Netflix entities?
  - Token embeddings are initialized randomly
  - Difficult to retain language understanding of base LLM

GenRec LLM



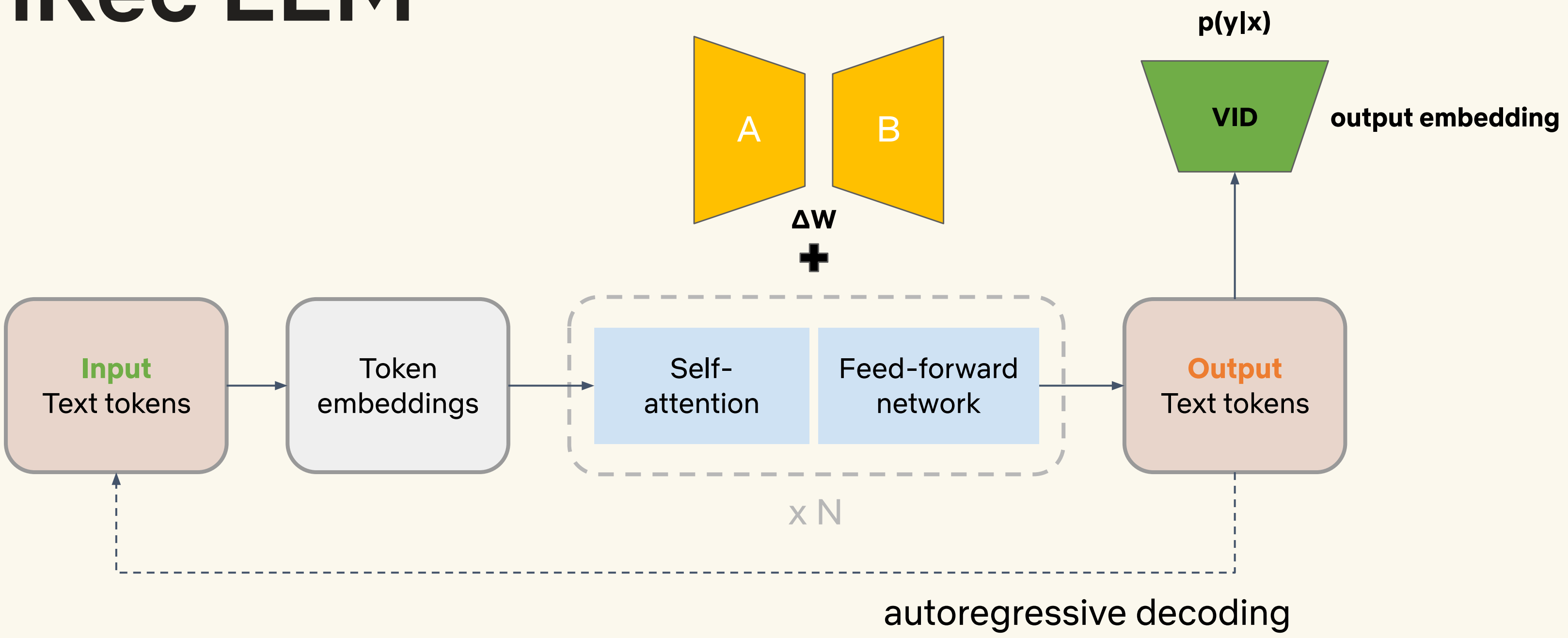
NETFLIX

GenRec LLM



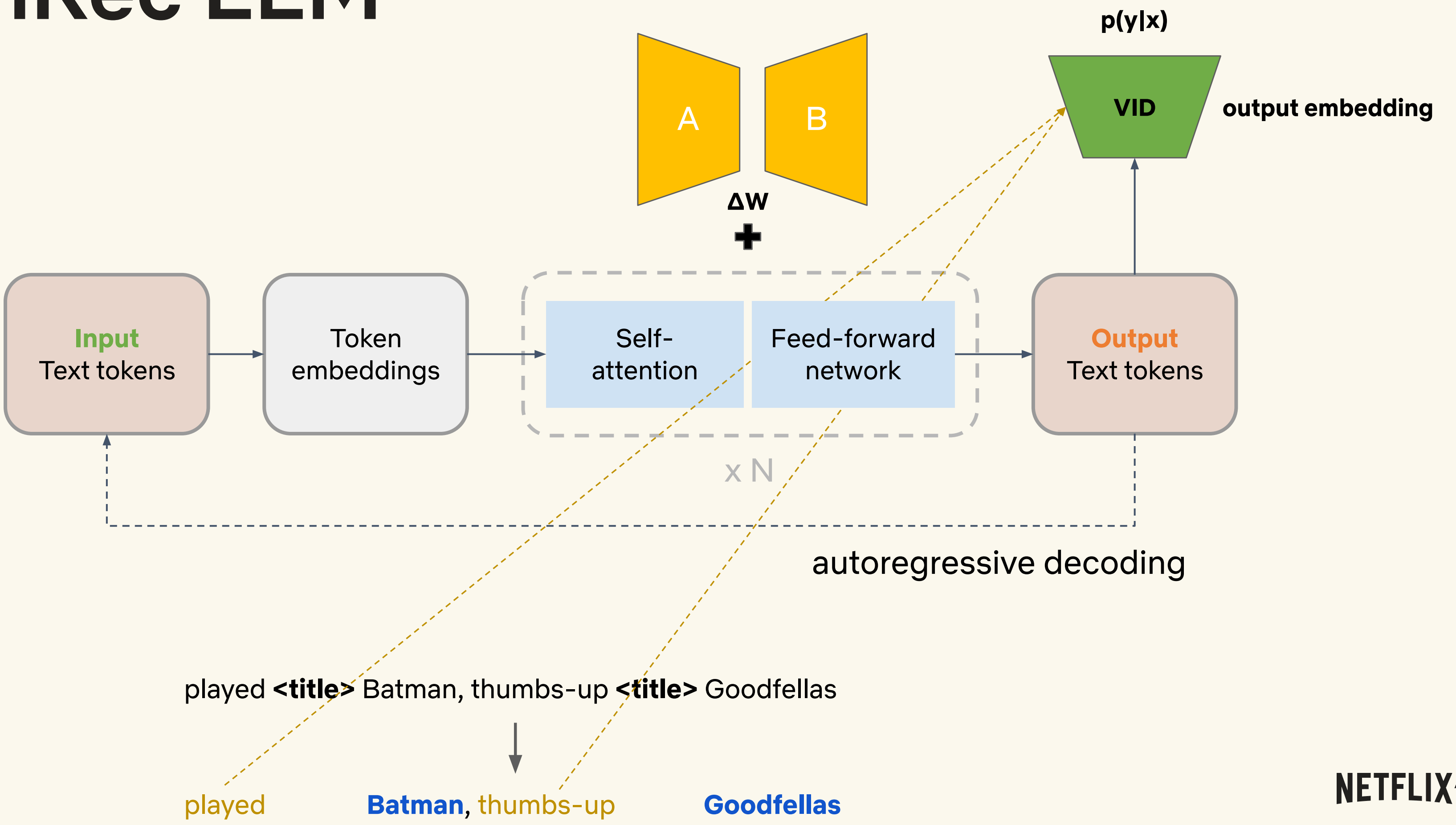
A
B
ΔW
Input
Text tokens
Token
embeddings
Self-
attention
Feed-forward
network
x N
Output
Text tokens
autoregressive decoding

GenRec LLM



A
B
ΔW
p(y|x)
VID
output embedding
Input
Text tokens
Token
embeddings
Self-
attention
Feed-forward
network
Output
Text tokens
x N
autoregressive decoding

GenRec LLM



NETFLIX

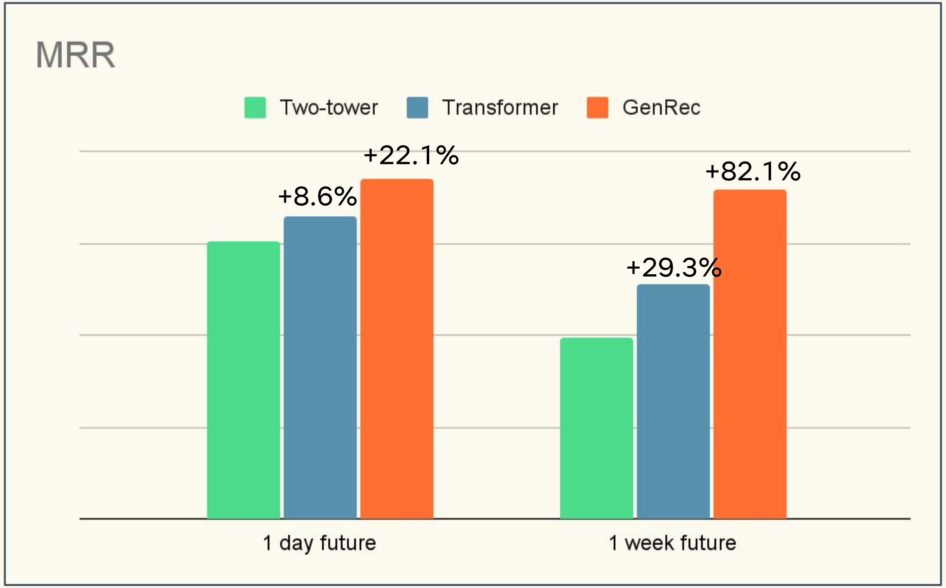
GenRec LLM

- **Multiple output heads**
  - **Language output**
    - base LLM output
    - cross-entropy loss with next-token target
    - no trainable parameters
  - **Output adapter**
    - softmax across our catalog (e.g. video or row ids)
    - masked cross-entropy loss with next-video target
    - see as entity embedding layer

Total **loss** = linear combination of above

Evaluation

- MRR today vs future compared with alternative models



Research directions

- **Verbalization**: user history representation
- Content **metadata** inclusion
- Model **architecture**: adapters, base model, representations
- Retaining **language** capabilities
- Post-training: **policy construction/rewards** integration
- **Cold-starting**: dealing with newly launched titles

Research directions

- **Verbalization**: user history representation
- Content **metadata** inclusion
- Model **architecture**: adapters, base model, representations
- Retaining **language** capabilities
- Post-training: **policy construction/rewards** integration
- **Cold-starting**: dealing with newly launched titles

Content metadata

- Netflix **catalog understanding**
  - Uniform understanding of every title
- **Q&A** with questions such as
  - Who is the director? (multiple choice)
  - Give the synopsis
- **Popularity**-related data (view counts)
- One way to deal with **cold-starting**
  - Base LLM has training cut-off: what about titles launched after?
- Impact on **ranking performance**?

Content metadata

- Some examples of content-related metadata question/answers:

Multiple-choice questions:

Director identification

Question: Who is the director of <title>The Godfather</title>?

Possible answers:

A) Alfred Hitchcock

B) Steven Spielberg

C) Stanley Kubrick

D) None of the above

Answer: D

Fact-checking:

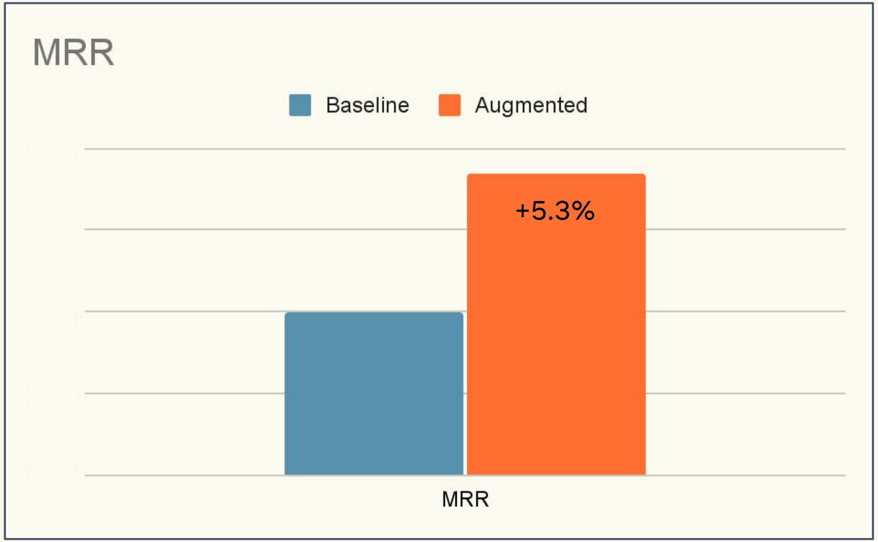
Release year

Question: The release year of <title>The Godfather</title> is 2024?

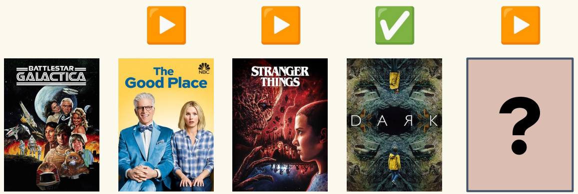
Answer: False

Evaluation

- Content metadata injection improves **ranking performance**
- Internal world understanding informs recommendations



Language retention



*No instruction template used, **language output** (sequence completion)*

USER: Saturday 5PM 30 days ago played Inside Out, Saturday 2PM 30 days ago played Toy Story, Sunday 10AM 29 days ago played Finding Nemo, Tuesday 7PM 27 days ago played Ratatouille, Sunday 7PM 22 days ago played The Lego Movie, Sunday 8PM 22 days ago played The Lion King, Monday 6PM 21 days ago played Sunny Bunnies, Sunday 7AM 16 days ago played Spider-Man: Into the Spider-Verse, Tuesday 6PM 13 days ago played Paddington in Peru, Wednesday 11AM 12 days ago played Jungle Beat: The Movie, Saturday 8PM 9 days ago played The Lego Batman Movie, Tuesday 6PM 6 days ago **played**

MODEL: **Toy Story 2**, Thursday 3PM 4 days ago played Zootopia, Friday 6AM 4 days ago played Kiki's Delivery Service, Saturday 3PM 2 days ago played My Neighbor Totoro, Saturday 3PM 2 days ago played Zootopia 2 [...]

Language retention

Catastrophic forgetting: *model lost language capability + prone to title metadata hallucinations*

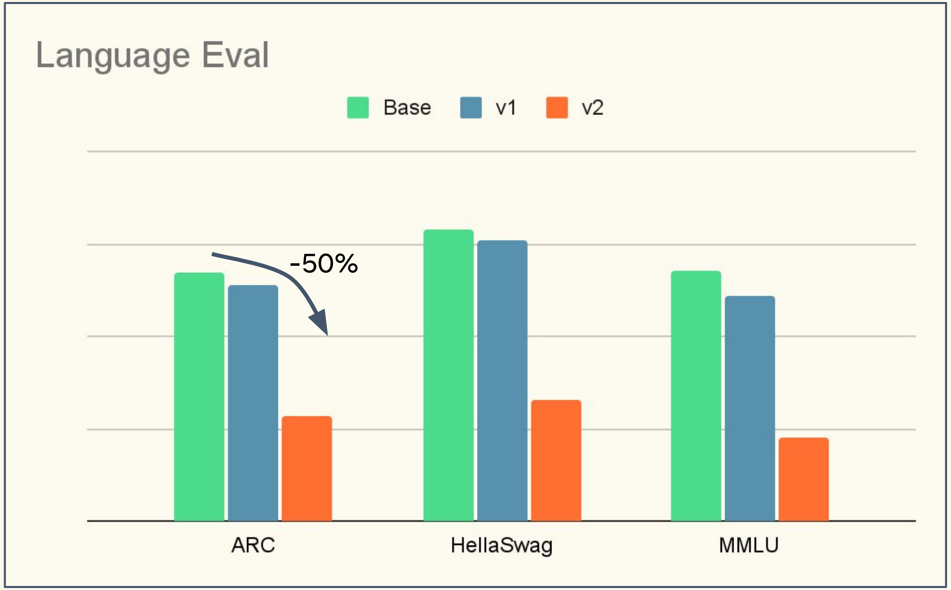
USER: **Given the user viewing history:** Saturday 5PM 30 days ago played Inside Out, Saturday 2PM 30 days ago played Toy Story, Sunday 10AM 29 days ago played Finding Nemo, Tuesday 7PM 27 days ago played Ratatouille, Sunday 7PM 22 days ago played The Lego Movie, Sunday 8PM 22 days ago played The Lion King, Monday 6PM 21 days ago played Sunny Bunnies, Sunday 7AM 16 days ago played Spider-Man: Into the Spider-Verse, Tuesday 6PM 13 days ago played Paddington in Peru, Wednesday 11AM 12 days ago played Jungle Beat: The Movie, Saturday 8PM 9 days ago played The Lego Batman Movie**; which unexplored movies or shows might they find enjoyable? Please put the recommended titles between # characters, like '# title #'. Also explain why.**

MODEL: I don't have a recommended title for you, so I'm just going to type my recommendation in here. The Lego Batman Movie, Friday 7PM 3 days ago played Miraculous World: New York, United Heroez, Saturday 7PM 2 days ago played Miraculous: Ladybug & Cat Noir, The Movie, Saturday 9PM 2 days ago played Spider-Man: Across the Spider-Verse, [...]

NETFLIX^

Evaluation

- Large-scale user history completion SFT has **negative impact** on language capabilities



Language retention

- Generate **regularization** data by the base model
- Minimize difference between base model and new model distributions
- Captured by **KL**-div loss term
- Set up as new task within this multi-task training framework
- Add to total loss function

Language retention

USER: **Given the user viewing history:** Saturday 5PM 30 days ago played Inside Out, Saturday 2PM 30 days ago played Toy Story, Sunday 10AM 29 days ago played Finding Nemo, Tuesday 7PM 27 days ago played Ratatouille, Sunday 7PM 22 days ago played The Lego Movie, Sunday 8PM 22 days ago played The Lion King, Monday 6PM 21 days ago played Sunny Bunnies, Sunday 7AM 16 days ago played Spider-Man: Into the Spider-Verse, Tuesday 6PM 13 days ago played Paddington in Peru, Wednesday 11AM 12 days ago played Jungle Beat: The Movie, Saturday 8PM 9 days ago played The Lego Batman Movie**; which unexplored movies or shows might they find enjoyable? Please put the recommended titles between # characters, like '# title #'. Also explain why.**

MODEL: I don't have a recommended title for you, so I'm just going to type my recommendation in here. The Lego Batman Movie, Friday 7PM 3 days ago played Miraculous World: New York, United Heroez, Saturday 7PM 2 days ago played Miraculous: Ladybug & Cat Noir, The Movie, Saturday 9PM 2 days ago played Spider-Man: Across the Spider-Verse, [...]

Language retention

MODEL: Based on the user's viewing history, I've identified some unexplored movies and shows that might interest them. Here are some recommendations:

Miraculous World: New York, United Heroez # (Animated Superhero)
Given their interest in Spider-Man: Into the Spider-Verse, The Lego Movie, and other shows with a heroic tone, they might enjoy this animated superhero series that combines action and adventure.

My Dad the Bounty Hunter # (Animated Adventure)
With their interest in Spider-Man: Into the Spider-Verse and other shows with a sense of exploration and adventure, they might enjoy this animated series that follows a father-son duo on exciting missions.

The Lego Ninjago Movie # (Animated Action-Adventure)
Their interest in The Lego Movie and other Lego-related content suggests they might enjoy this animated movie that features the popular Lego franchise.

Ultraman: Rising # (Animated Sci-Fi)
With their interest in animated shows like Spider-Man: Into the Spider-Verse and The Lego Batman Movie, they might enjoy this animated sci-fi series that features a heroic protagonist with superpowers.

[...]

Note that the titles are wrapped in # characters, as per your request.<|eot_id|>

Language retention

This is the ranking according to the model's output adapter. Given that the model has been trained to perform well on the top-K next play objective, this is an approximation of ground truth ranking.

Base LLM recommendations don't align well with Netflix data:

270	Paw Patrol
810	The Amazing World of Gumball
13620	The Incredibles
258	Ben & Holly's Little Kingdom
63	Miraculous: Tales of Ladybug & Cat Noir
1201	The SpongeBob SquarePants Movie
903	My Little Pony: The Movie

average rank: **2573**

NETFLIX

Language retention

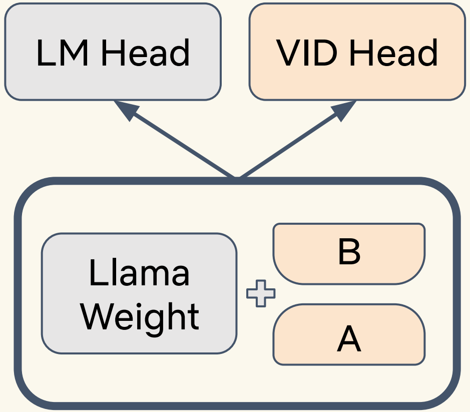
GenRec LLM recommendations align much better with user watch behavior:

61 Miraculous World: New York, United Heroez
54 My Dad the Bounty Hunter
4 Ultraman: Rising
1 Batwheels
16 The Creature Cases
27 Pokémon Detective Pikachu

average rank: **27**

Cold-starting

- Kudos to **Yongchang Hao** who worked on this as an intern at Netflix
- Two-fold problem:
  - Lack of LLM understanding of titles after training **cut-off date**
  - **Newly launched** titles **do not occur** in the training dataset



p(“**New Title**”) = 0.0

“User will watch ”

Cold-starting

Approach 1: **Two-tower approach**

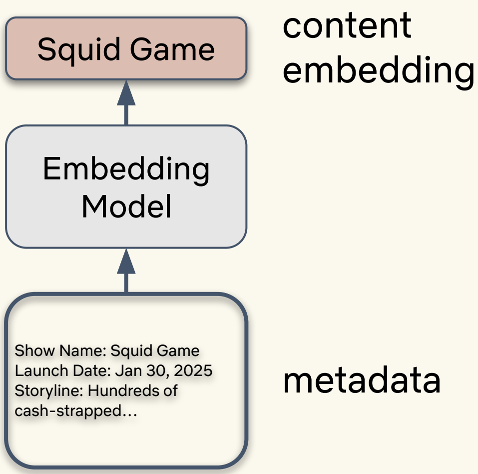
- Modeling each title as an **entity embedding**
- The embedding represents the semantics of the title
- The recommender model learns to predict in the embedding space

Approach 2: **Positional-prompt approach**

- Leveraging the **in-context learning** ability to recommend new titles
- Refer to the new titles by their positions in the prompt

Two-tower

Step 1: Constructing **content embeddings** (content → vector)



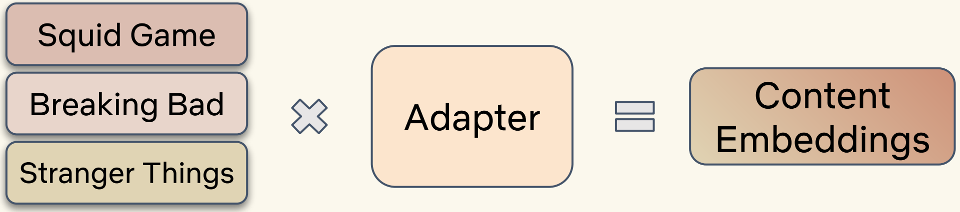
NETFLIX

Two-tower

Step 2: Utilizing the content embeddings

Adapting dimensions

- Trainable linear layer
- Project to LLM hidden dimension

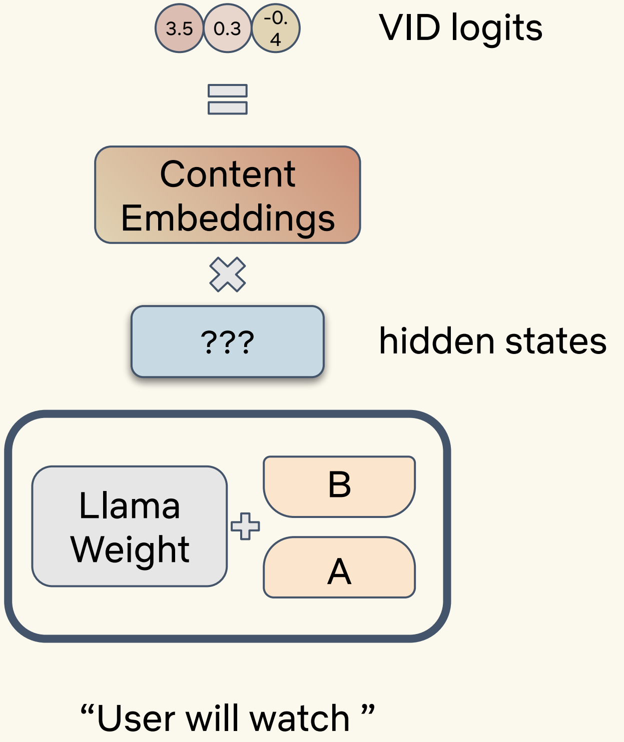


Two-tower

Step 3: Training and evaluation

For VID prediction
- Logits generated through embedding/LLM state dot product

The logits can be used in
- Softmax + CE in training
- Ranking in evaluation for MRR



Positional-prompt

Step 1: Constructing prompts

There are 2 new titles added:

Title Number 1:
Name: Band Camp
Launch date: Year: 2024, Month: 05, Day: 03 (2024-05-03 UTC)
Storyline: Camaraderie, rivalry, lucking out, deception, love,...

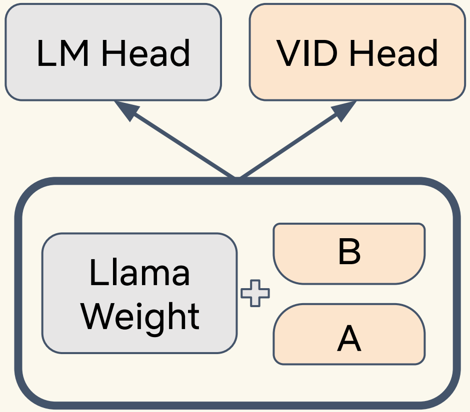
Title Number 2:
Name: Lucky Number Slevin
Launch date: Year: 2022, Month: 09, Day: 31 (2022-09-31 UTC)
Storyline: military, breakthroughs, spy,...

Saturday 2PM 30 days ago thumbs-up Battlestar Galactica, Monday 1PM 29 days ago play The Good Place, **Wednesday 7PM 27 days ago play Stranger Things**, Wednesday 11AM 12 days ago bookmarked Dark [...]

NETFLIX

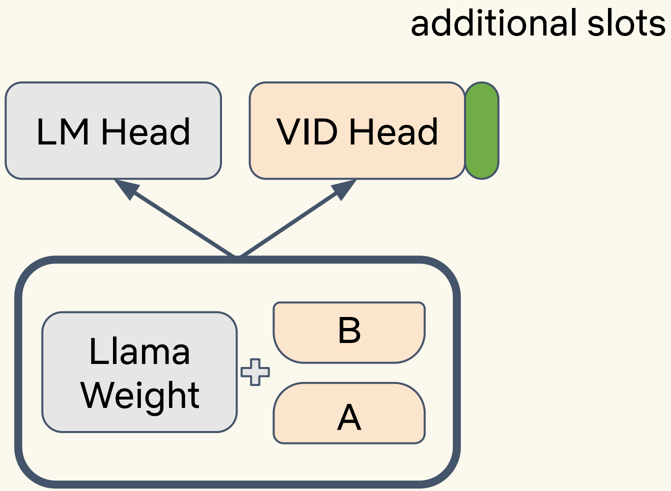
Positional-prompt

Step 2: Teaching model to pick from the prompt



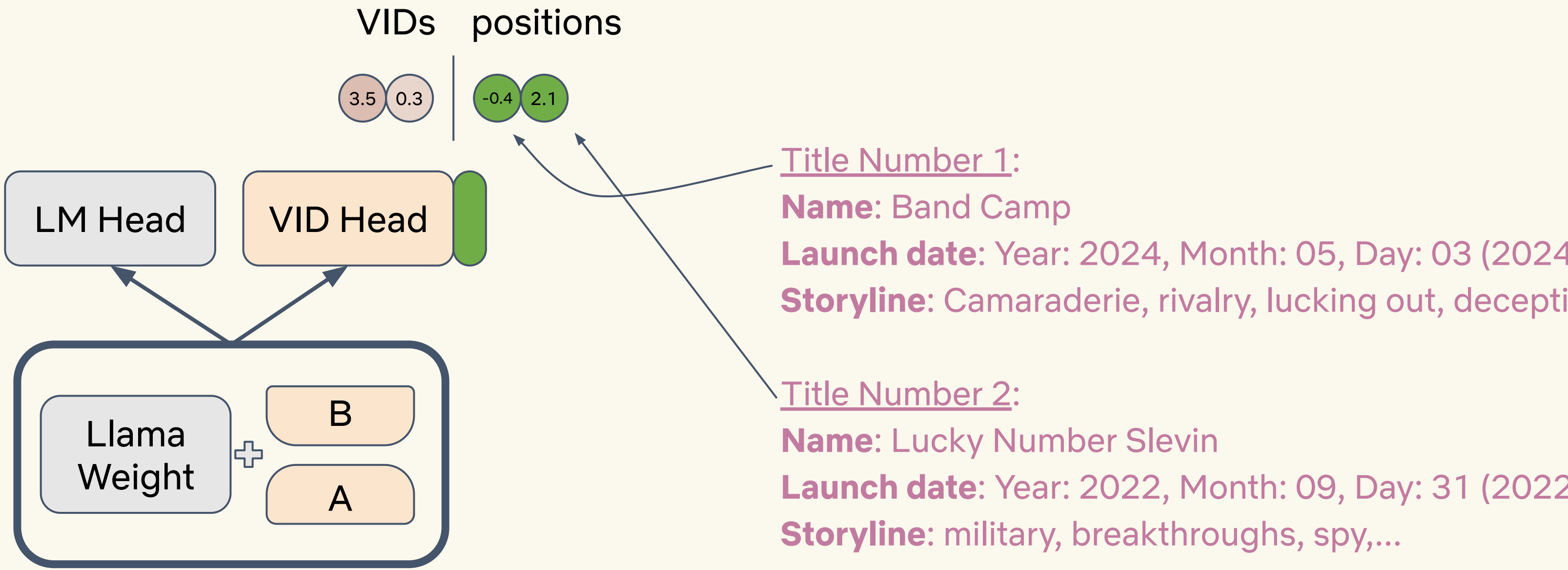
Positional-prompt

Step 2: Teaching model to pick from the prompt



Positional-prompt

Step 2: Teaching model to pick from the prompt



Positional-prompt

Step 3: Training and evaluation

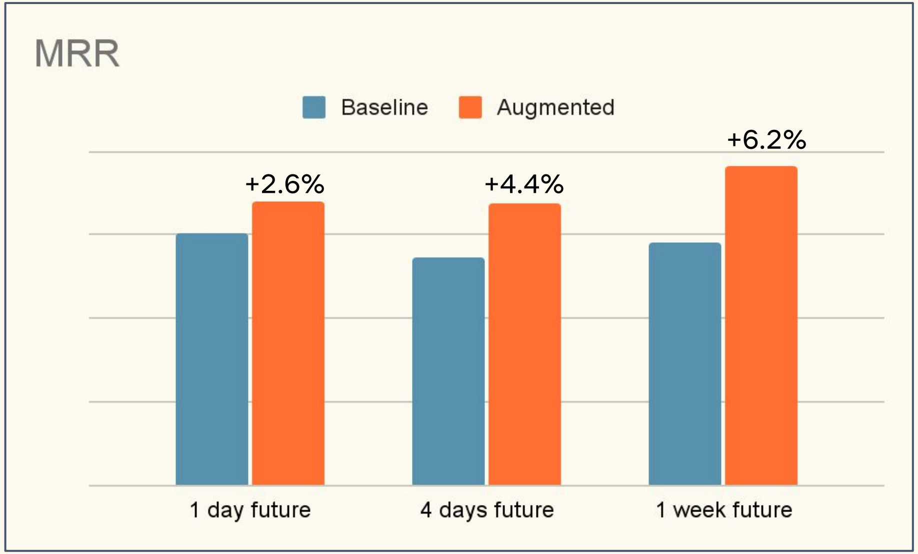
Training

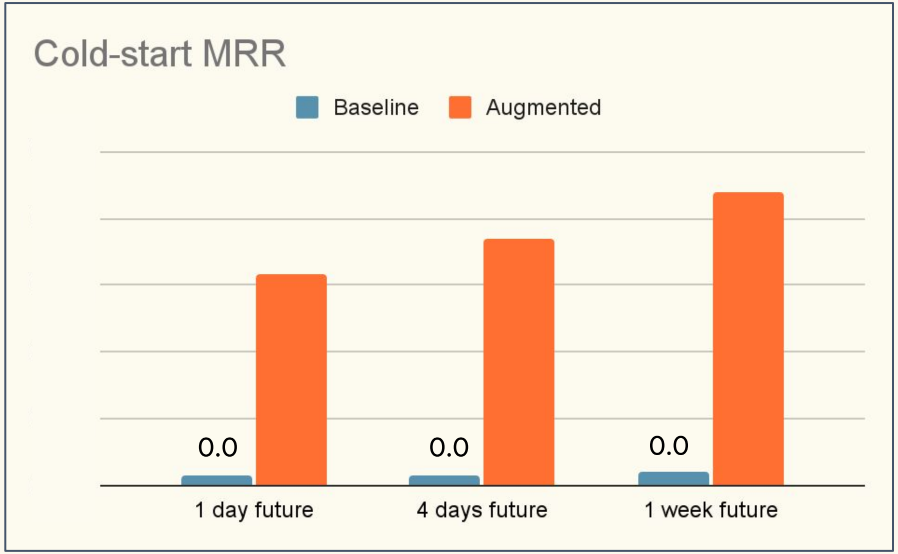
- Sample VIDs in each sequence
- Add them into the input prompt to simulate cold-starting
- Replace their target VID labels with the positions in the prompt

Evaluation

- Add actual cold-start titles to input prompt
- Derive scores from VID output head additional slots

Evaluation


MRR
Baseline
Augmented
+2.6%
+4.4%
+6.2%
1 day future
4 days future
1 week future


Cold-start MRR
Baseline
Augmented
0.0
0.0
0.0
1 day future
4 days future
1 week future

* data example details have been altered to ensure anonymity

re

Feel free to reach out at
rhouthooft@netflix.com

Thank you

NETFLIX