Future Lens:

Anticipating Subsequent Tokens from a Single Hidden State

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, David Bau

We can decode **future tokens** from a **single** LLM state.

Future Lens Application: Marty Fly from → Back to the Future

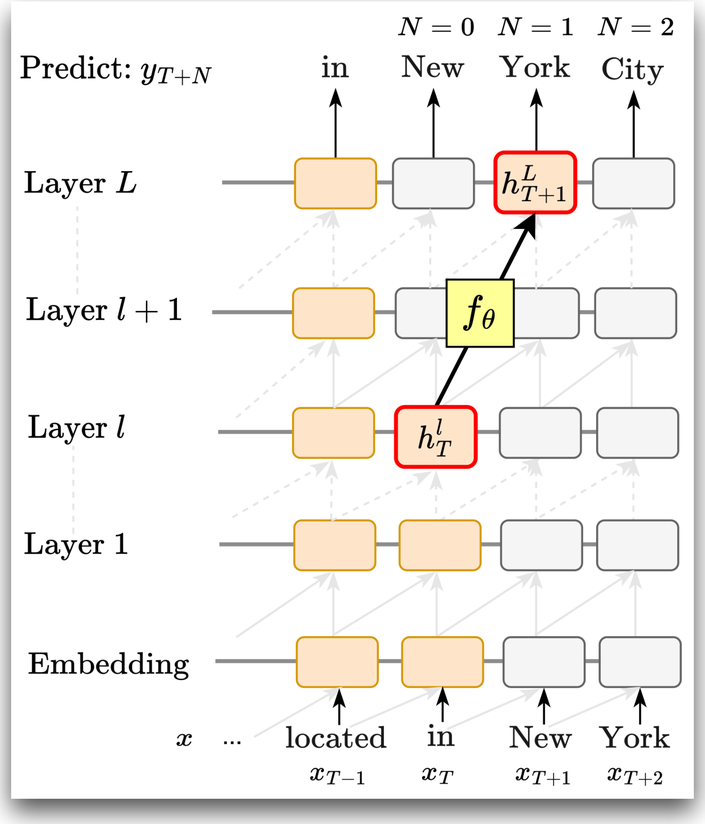
	Mart	y	Mc	Fly	from
L1	inez the first time	erson the screen.	Afee the same source	er and the other	behindrrhaph
L2	inez\n\n\n	\xe9n-1-	Afee and the other	er the left of	Sons\n\nThe
L3	.\n\nThe	Friedman and the other	Afee\n	er\n\n\n	Havanaa:\n
	\n de la c	Barn and the other	Lean.\n\n	er\ufffd\ufffd said	1992a:\n
	\n de la c	ring the first time	Leanmig.	er \ufffd\ufffdI	Oklahoma "\n\n
	\n de la c	ring and the next	Afee\n\nThe	world and the future	Austria book.\n
	\n de la c	ring and the other	Leanaway from the	mer 1, 1	Australia movie.\n
	\n, and the	ell and the other	Leanaway from the	walker the time of	England first time he
	\n, and the	ellLean, and	Lean\n\nThe	walker be the first	Australia movie "The
	\n, and the	ellDonough,	Leank\n\n	te Marty McFly	Australia movie "The
	\n"\n	ellDonough,	Lean\n\nThe	te Marty McFly	Vietnam movie "The
	\n" id="	GreenbergDonough,	Lean\n\nThe	te Marty McFly	Germany movie "The
	\n" id="	GreenbergDonough,	Lean\n\nThe	movies Marty McFly	Boston movie "Back
	\n" id="	ellDonough,	Bride\n\nThe	movie Marty McFly	movie movie "The
	\n" id="	riumDonough,	Bride\n\nThe	movie Marty McFly	movie movie "The
	\n" id="	WalshDonough,	Bride\n\nThe	movie Marty McFly	movie movie "The
	\n" id="	McDonough,	Bride\n\nThe	movie Marty McFly	movie Back to the
	\npng" alt	McDonough,	Bride\n\nThe	movie Marty McFly	movie movie "Back
	\npng" alt	riumDonough,	Flylew\n	movie Marty McFly	1980 Back to the
	\npng" alt	ring Marty, and	Fly\n\nThe	movie Marty McFly	movie Back to the
	\npng" alt	Mc and the other	Fly\n\nThe	Returns Marty McFly	movie Back to the
	\npng" alt	Mc\ufffd\ufffd he	Fly\n\nThe	arrives Marty McFly	movie Back to the
	\npng" alt	Mc he was a		cFly	1984 Back to the
	\npng" alt	Mc and I'm		cFly	1989 to the Future
	\npng" alt	Mc and I'm			Back to the Future
L26	\n1.0	's and I'm			Back movie "Back
L27	.\n\nThe	Mc and the other			the movie "Back
L28	y\n\nThe	Mc and I'm	Fly.\n\n	\n and the future	Back future.\n

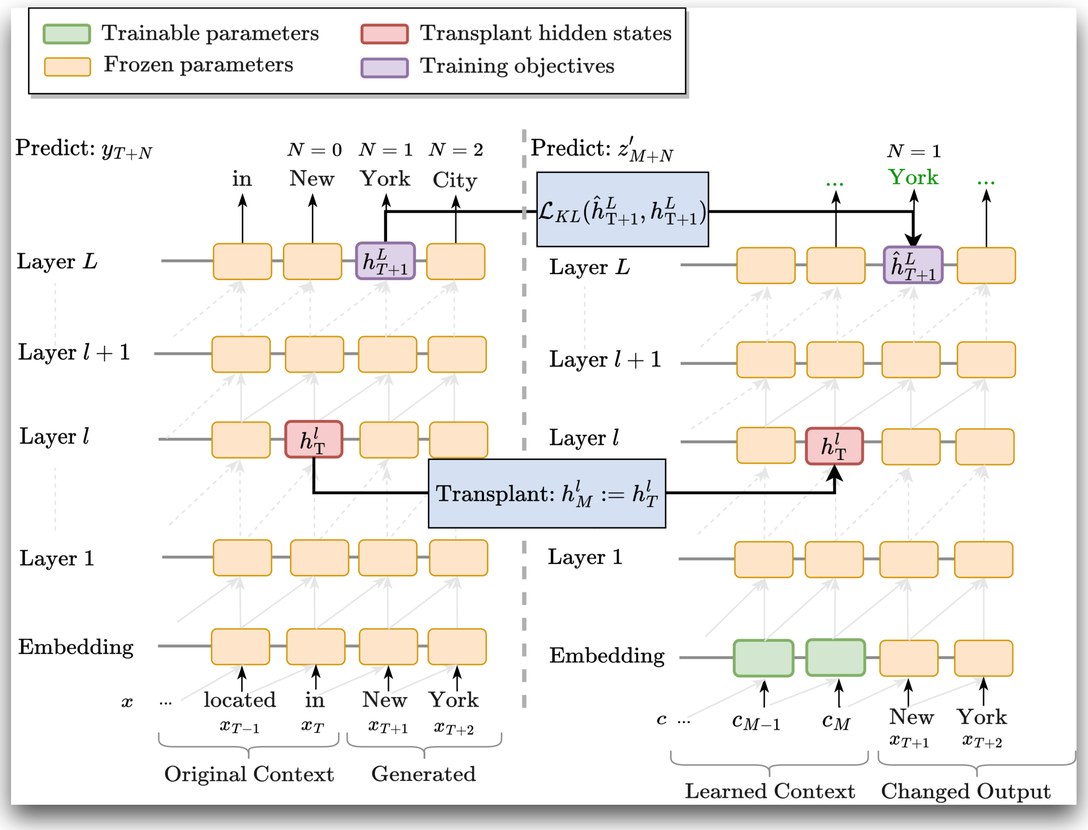
Predicting multiple tokens ahead in a single state

LLMs are typically trained to predict one token ahead, but recent work has hinted that individual hidden states may contain more information than just probabilities of the next token.

To what extent can we extract information about future tokens from a single hidden token representation?

We find that we can, with more than 48% accuracy!



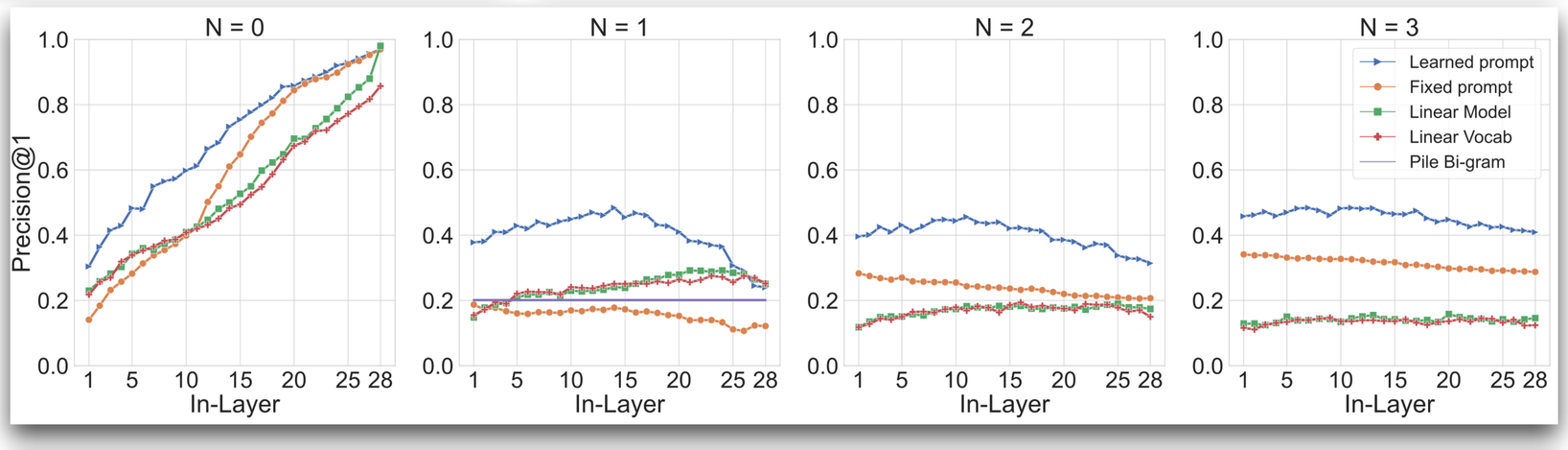


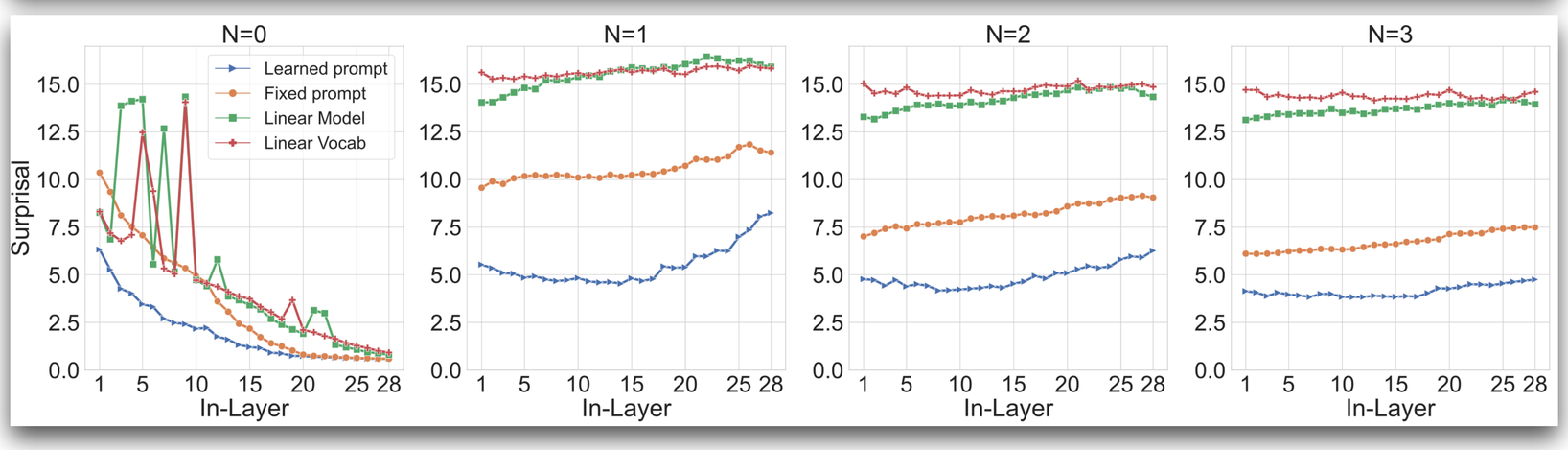
How do we decode future tokens?

1. Linear Model Approximation
Train linear models to approximate future tokens or hidden states.

2. Fixed Prompt Causal Intervention
Patch the hidden state into different transformer run with an unrelated prompt. We see if the future tokens are present in the transplanted state.

3. Learned Prompt Causal Intervention
To increase accuracy on subsequent tokens, we learn a soft prompt that learns to predict future tokens from the transplanted state.





https://future.baulab.info/

Northeastern University