Superposed Decoding

Ethan Shen
NeurIPS 2024.

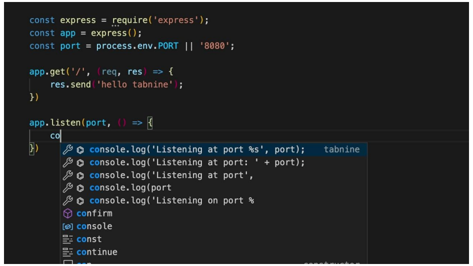
Alan Fan, Sarah M. Pratt, Jae Sung Park, Matthew Wallingford, Sham M. Kakade, Ari Holtzman, Ranjay Krishna, Ali Farhadi, Aditya Kusupati

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Drafting Scenarios

- Users often want multiple distinct outputs from LLMs

tabnine

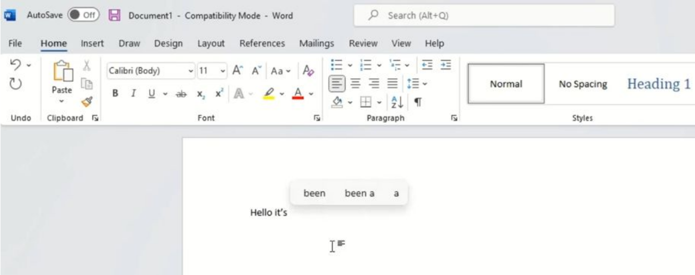


```
const express = require('express');
const app = express();
const port = process.env.PORT || '8080';

app.get('/', (req, res) => {
    res.send('hello tabnine');
})

app.listen(port, () => {
    co
})
        console.log('Listening at port %s', port);    tabnine
        console.log('Listening at port: ' + port);
        console.log('Listening at port',
        console.log(port
        console.log('Listening on port %
        confirm
        console
        const
        continue
```

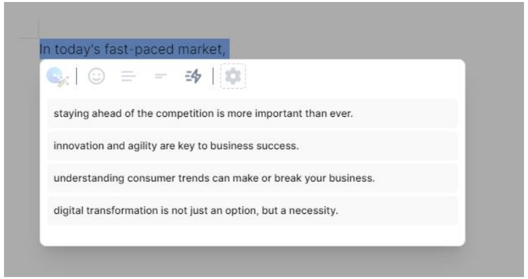
GitHub



been   been a   a

Hello it's

COMPOSE AI



In today's fast-paced market,

staying ahead of the competition is more important than ever.

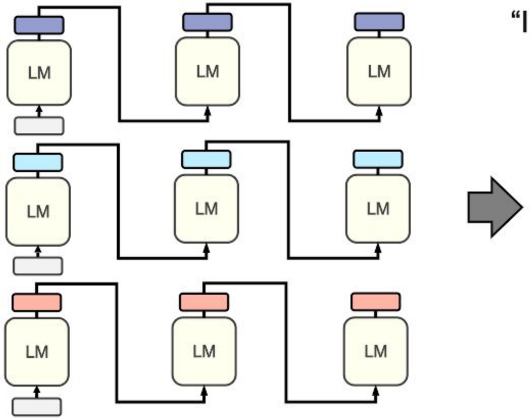
innovation and agility are key to business success.

understanding consumer trends can make or break your business.

digital transformation is not just an option, but a necessity.

Current Approaches

- Decoding Methods
  - Nucleus Sampling
  - Beam Search
  - Top-k Sampling
  - Greedy Decoding (Only one draft)
- **Multiple inference passes** (batch size = 1)



“I just arrived in Xalapa, Mexico - today was my first”

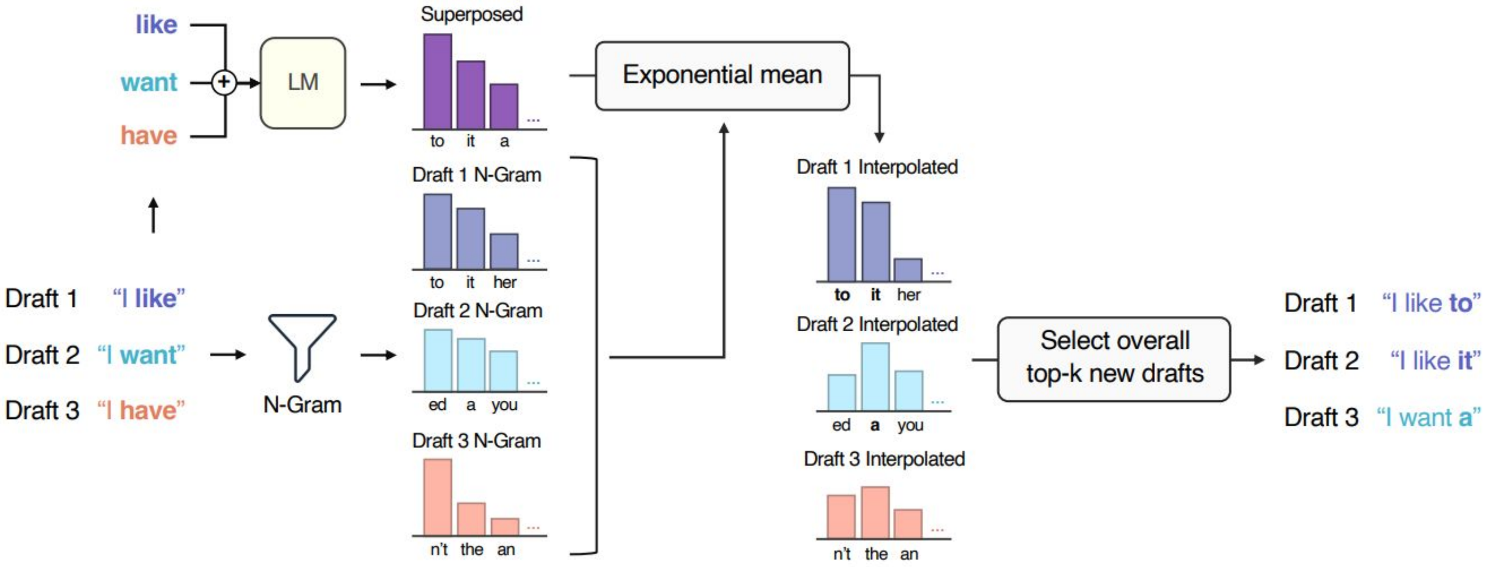
day of class. At Universidad Veracruz…

day of the Rosenkranz School at the…

day of orientation, and I could not wait to…

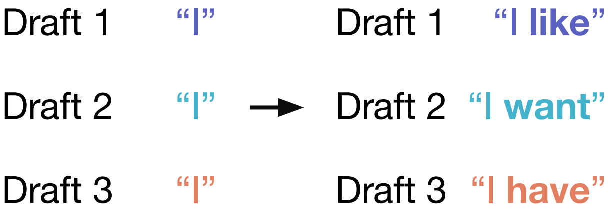
Superposed Decoding (SPD)

Idea: Linearly combine token embeddings to extend all k drafts with **one inference pass**



First Timestep

1. Let x denote token in vocabulary V and $M=(x\_1, ..., x\_m)$ an initial prefix of m tokens.
2. Grow initial drafts using model θ:

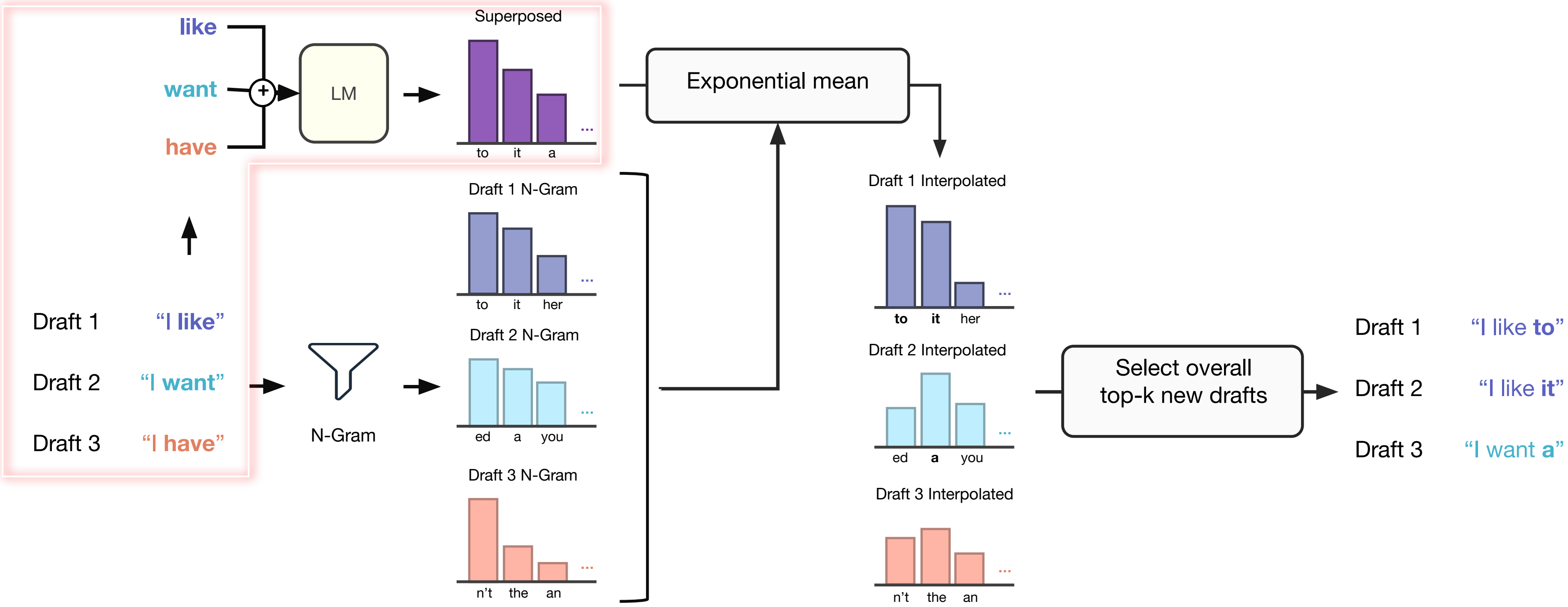


Draft 1 "I" Draft 1 "I **like**"

Draft 2 "I" → Draft 2 "I **want**"

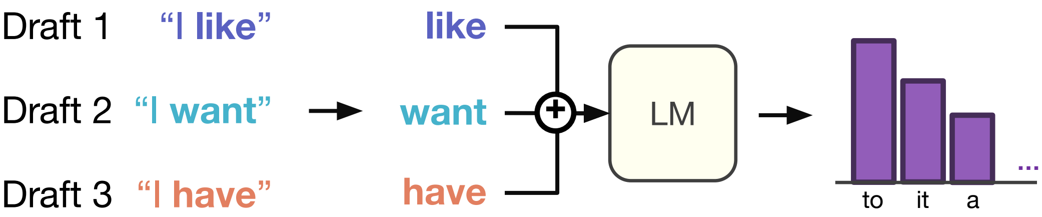
Draft 3 "I" Draft 3 "I **have**"

Step 1: Superposed Embedding Inference



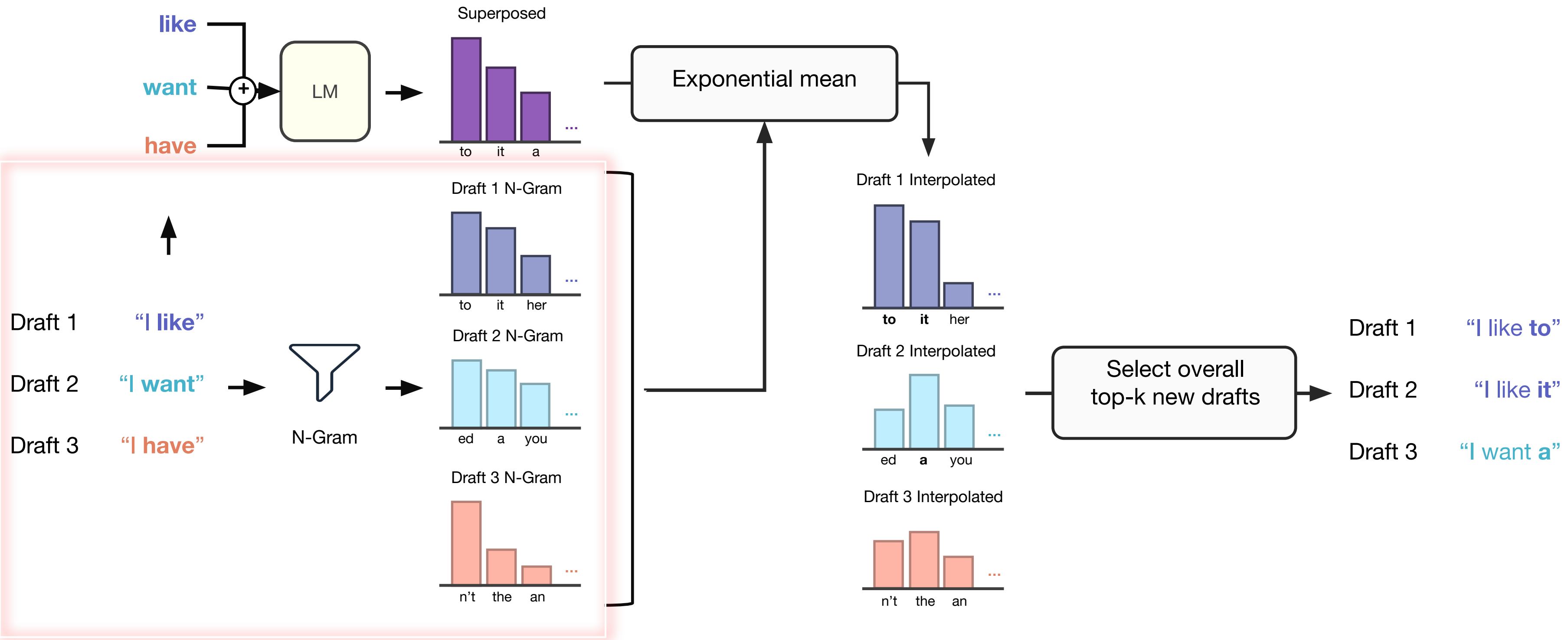
Step 1: Superposed Embedding Inference

- Find weighted combination of the embeddings of each draft's most recent token.



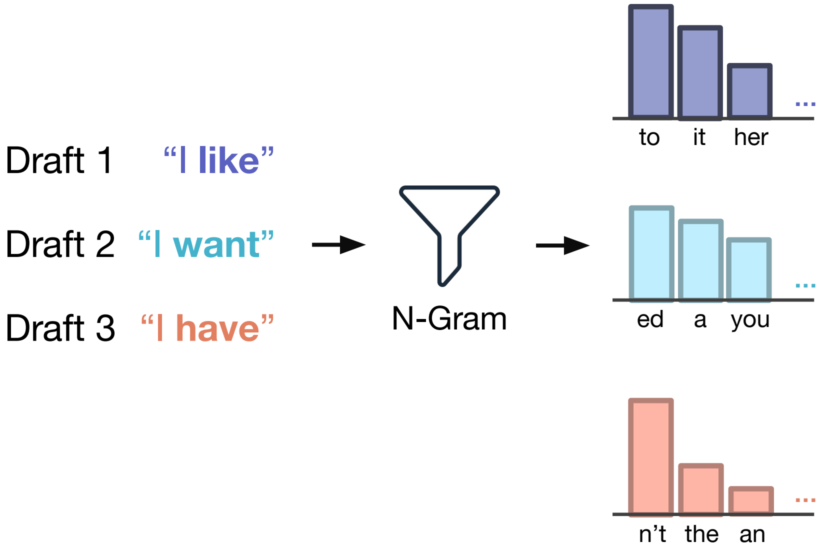
- Truncate the resulting distribution to only the top k tokens.

Step 2: N-Gram Distribution



Step 2: N-Gram Distribution

- Interpolate the next token distributions from a set of n-gram models ($n \in [2, 6]$)



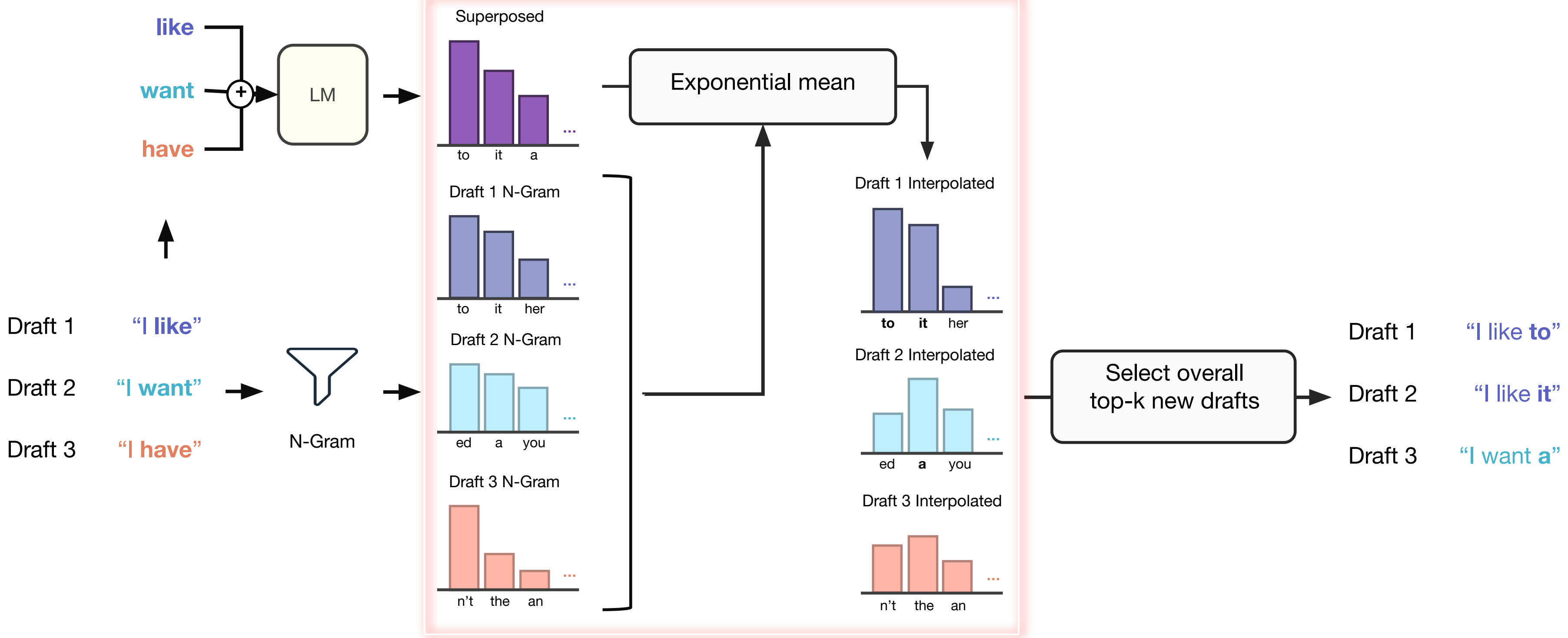
Draft 1 “I like”

Draft 2 “I want”

Draft 3 “I have”

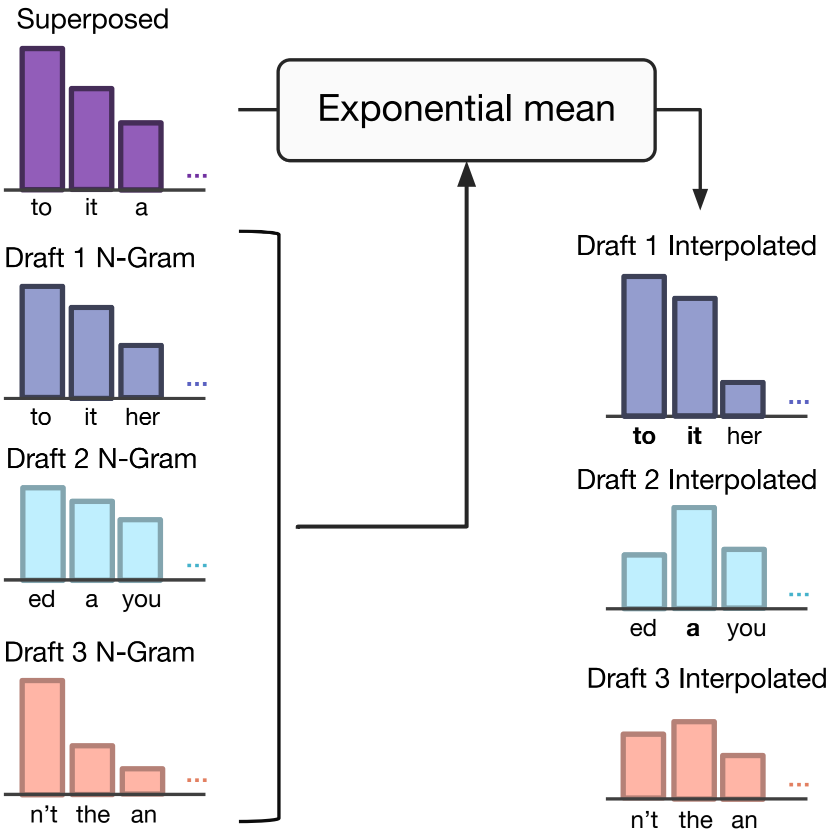
N-Gram

Step 3: Interpolation

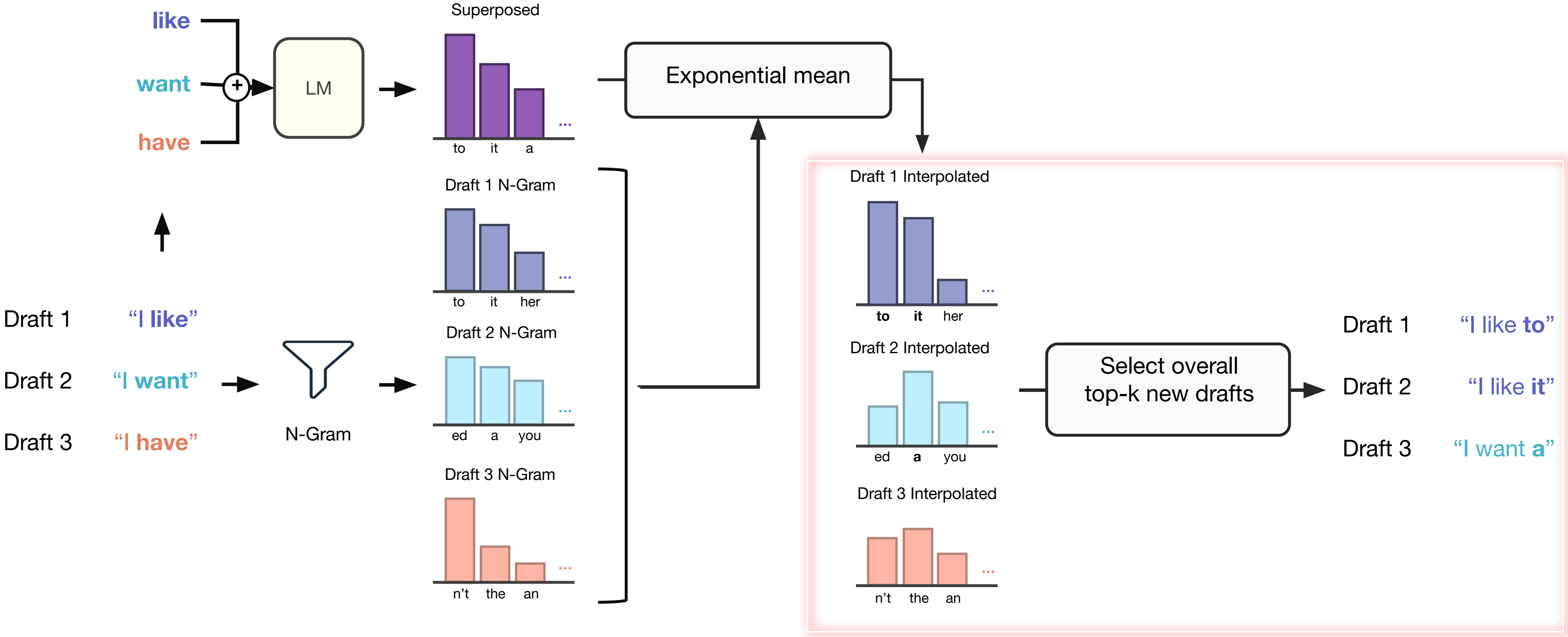


Step 3: Interpolation

- Interpolate the N-Gram distributions with the LM distribution
- Draft-specific distributions



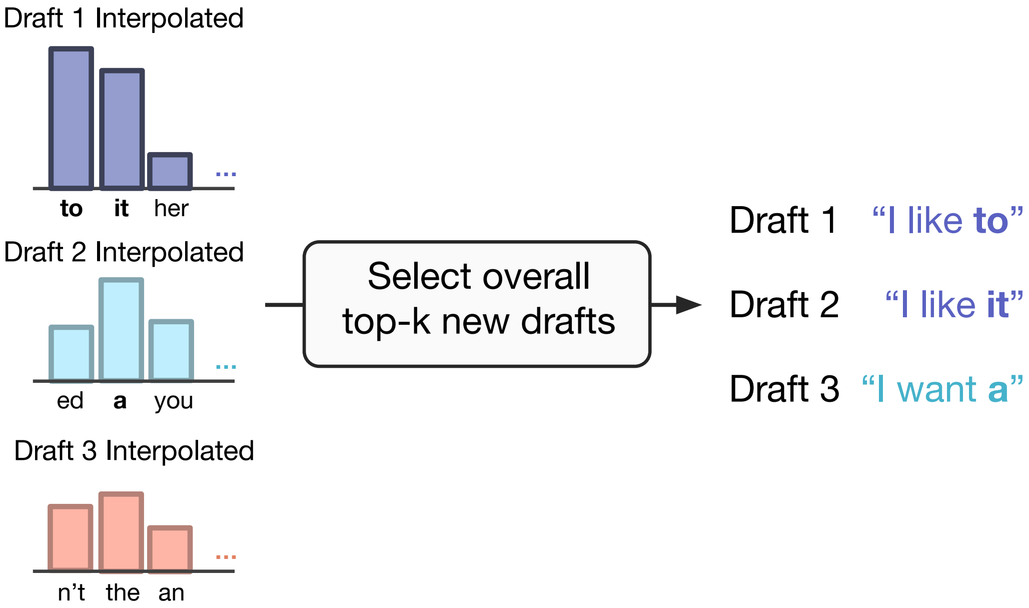
Step 4: Update Drafts



Step 4: Update Drafts

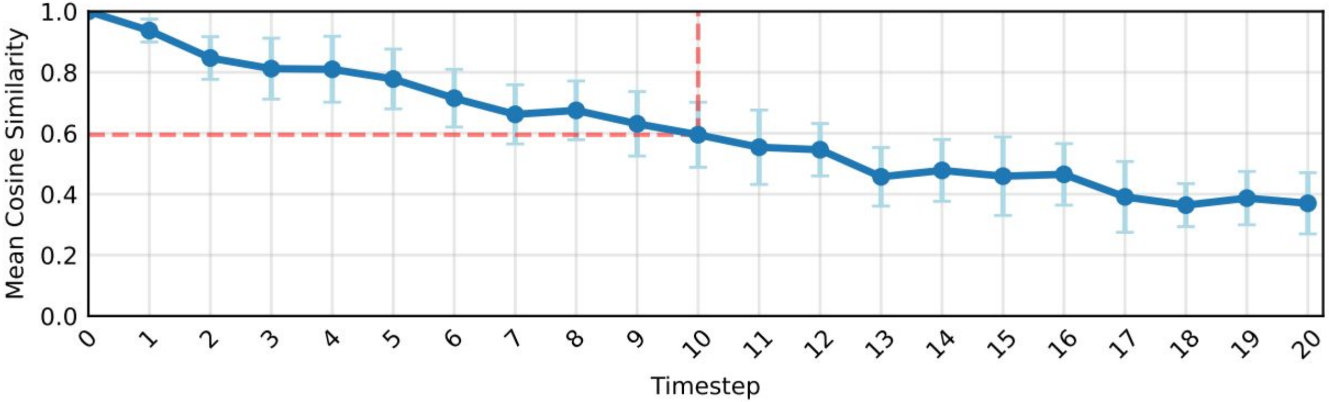
1. Next tokens form $k\text{^}2$ new draft options
2. Rank using joint probability of tokens and drafts
3. Update draft store

Repeat!!!



Why does Superposed Decoding work?

- Layer embeddings using Superposed Decoding linearly relate to those of Beam Search drafts
    - 10 randomly sampled batches of 10 prefixes each (100 total prefixes)
- High linearity up to 10 timesteps - optimal generation length



Example #1

	Text
OpenWebText Prefix	When I worked as a scout for the Carolina Panthers in the
Nucleus Sampling	1990s, I would often
Superposed Decoding	1990s, I was **always** 1990s, I was **a** 1990s, I was **responsible**

Example #2

	Text
OpenWebText Prefix	Over a century ago, the RMS Titanic's fate
Nucleus Sampling	was sealed when it struck an iceberg on
Superposed Decoding	was sealed when it **struck** an iceberg **and** was sealed when it **hit** an iceberg **and** was sealed when it **hit** an iceberg **on**

III. Results

Experimental Setup

- Implement on Llama-2-7B
- N-Gram Models constructed using 200M tokens from RedPajama

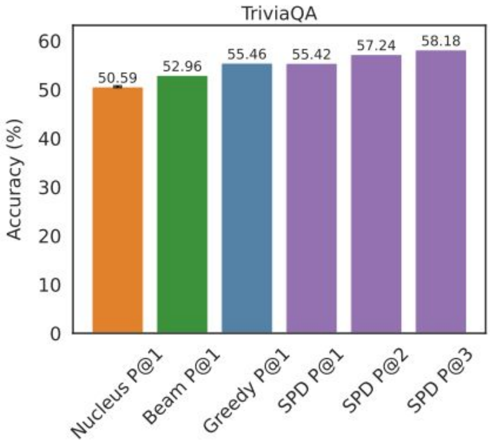
Coherency

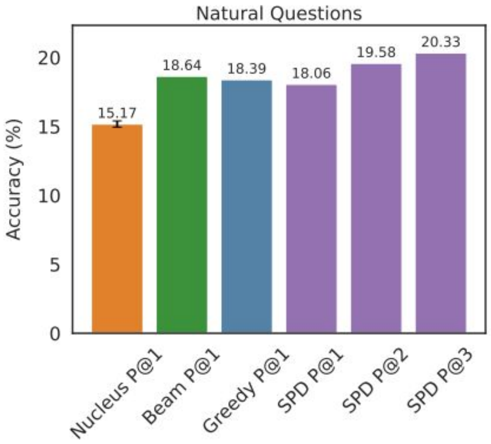
- Test on OpenWebText (10 tokens generated)
- Expect at least one SPD draft to be equal to Nucleus Sampling & others come free

	Nucleus	Beam/Greedy	N-Gram	Superposed Decoding			
Draft #	-	-	-	1	2	3	Best
Avg Perplexity	5.17	3.77	152.75	5.03	7.97	10.05	4.63

Accuracy

- Tested on TriviaQA and Natural Questions
- SPD gives more drafts @ same compute
- Extra drafts increase likelihood of factually accurate generations

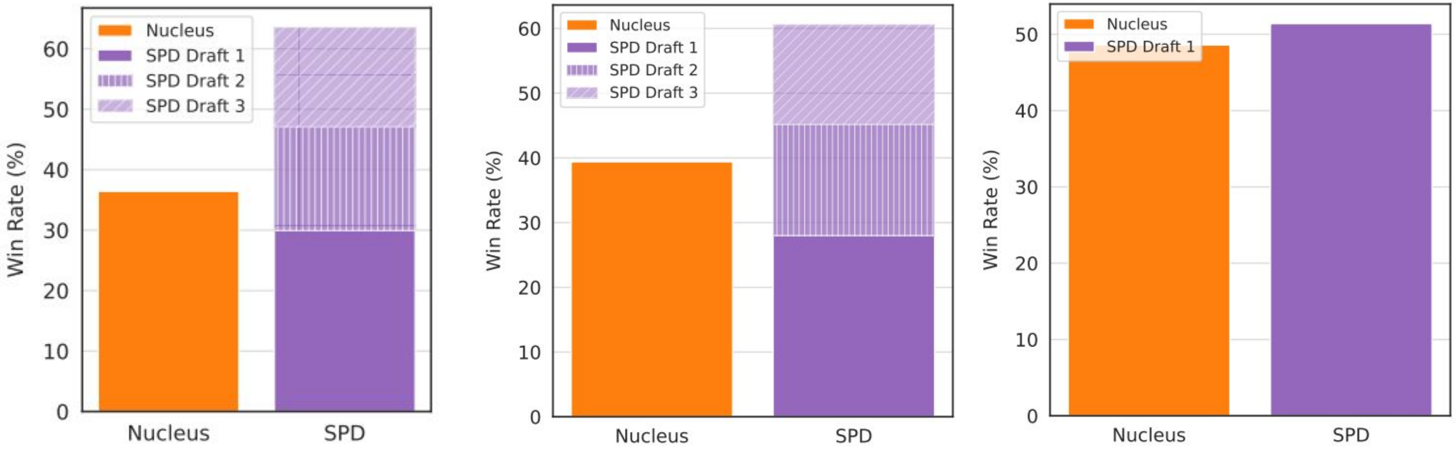




Human Evaluation

Three Surveys:

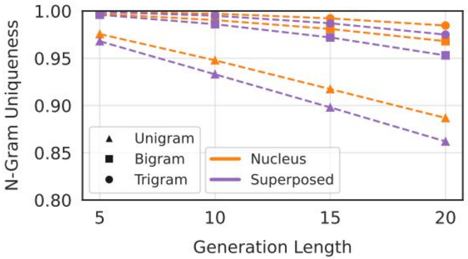
1. Constant Compute: 3 SPD vs 1 Nucleus (707 prefixes)
2. Unequal Compute: 3 SPD vs 2 Nucleus (100 prefixes)
3. Equal Number: 1 SPD vs 1 Nucleus (100 prefixes)

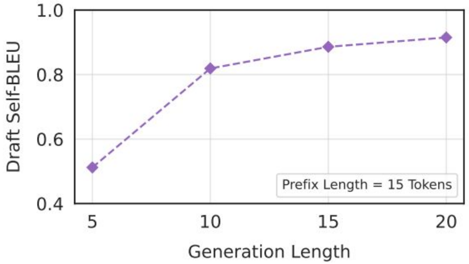


Ablations

Superposed Decoding:

- Does not suffer degeneration
- Increases diversity with smaller generation length
- Flawlessly extends to Mistral 7B





	Nucleus	Superposed Decoding			
Draft #	**-**	**1**	**2**	**3**	**Best**
Avg Perplexity	11.42	11.34	12.74	13.63	10.87

Complementary Benefits

- Superposed Decoding's benefits are *completely complementary* to other decoding methods
- SPD offers local search at *no extra cost*
  - Freely expands global search (Nucleus Sampling) or other local search (Beam Search)

Prefix	Nucleus Sampling (k = 3)	3 x Superposed Decoding
Melbourne is	Melbourne is a great city, with Melbourne is a city of many different Melbourne is the capital city of Victoria	Melbourne is a great city, with a lot of things Melbourne is a great city, with a lot to things Melbourne is a great city, with a lot of history Melbourne is a city of many different cultures and relig Melbourne is a city of many different cultures, relig Melbourne is a city of many different cultures and languages Melbourne is the capital city of Victoria, Australia. It Melbourne is the capital city of Victoria and Australia. It Melbourne is the capital city of Victoria, Australia. The

Test-Time Compute Scaling

- Repeated sampling (Brown et al., 2024) -> increased samples improve performance
- Extend Nucleus Sampling drafts with 2 or 3 SPD drafts free of cost

TriviaQA

Compute (k)	1	10	20	30	40	50	60	70	80	90	100
NS	51.04	68.75	70.31	71.87	72.92	74.48	74.74	75.26	75.78	76.30	76.56
NS + 2 SPD	51.30	68.75	70.57	72.66	74.74	75.78	76.30	76.82	78.39	79.17	79.43
NS + 3 SPD	**51.82**	**70.57**	**74.22**	**75.52**	**77.34**	**77.87**	**78.39**	**78.65**	**79.17**	**79.43**	**79.95**

Natural Questions

Compute (k)	1	10	20	30	40	50	60	70	80	90	100
NS	14.32	**32.55**	**36.98**	38.54	40.36	41.15	41.67	41.93	42.19	42.71	42.97
NS + 2 SPD	15.36	31.25	34.90	38.02	39.84	41.41	41.67	42.45	43.75	43.75	43.75
NS + 3 SPD	**15.63**	31.25	**36.98**	**39.06**	**41.15**	**42.71**	**43.75**	**43.75**	**44.27**	**44.79**	**45.57**

Takeaways

Benefits:

- Coherent and human preferred
- Factual
- Increase effective batch size
- Inference compute scaling

Applications:

- Tabnine
- Microsoft Copilot
- ...and more!!!