Generative LLMs

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

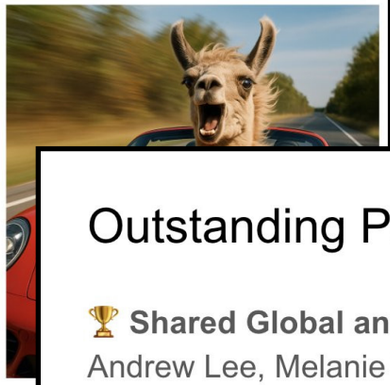
- Late breaking NLP research! COLM 2025 https://colmweb.org/ https://bsky.app/profile/colmweb.org/post/3m2m7vpuz7s2g

Outstanding Papers

COLM

🏆 **Fast Controlled Generation from Language Models with Adaptive Weighted Rejection Sampling**
Ben Lipkin, Benjamin LeBrun, Jacob Hoover Vigly, João Loula, David R. MacIver, Li Du, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Timothy J. O'Donnell, Alexander K. Lew, and Tim Vieira

The paper introduces a fast, principled, and adaptive sampler for controlled generation. It solves a real problem, and it actually works: getting LLMs to respect hard constraints, and do so fast -- 50x speedups! It shows an elegant connection to Sequential Monte Carlo and proper weighting. This work shows how classical probabilistic inference techniques can solve modern LLM problems.



Outstanding Papers

COLM

🏆 **Shared Global and Local Geometry of Language Model Embeddings**
Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg

This work characterizes the global and local geometry of language model token embeddings and finds similarities across language models. It particularly interesting to see how the linear algebra of embeddings holds at the concept level in LLMs, and furthermore that the concept vectors show similar relationships across models. It shows how we can gradually gain better insights or easier mechanisms of how LLMs work.

servicenow

- Getting in to Large Language Models
- We've seen
  - 1. Pretraining Data
  - 2. Transformer LM architecture
- Today
  - 1. Broad types of LLMs: encoder (BERT) vs decoder (GPT)
  - 2. Generative LLMs
    - Sampling methods
    - Prompts and direct task evaluations

- Practice midterm Qs?

- Key idea: a large pretraining corpus implicitly has lots of knowledge (or, knowledge-like patterns?)
  - Training an LM on it allows the LM to capture it
  - ... even beyond just word meanings, ideally

- With roses, dahlias, and peonies, I was surrounded by flowers
- The room wasn't just big it was enormous
- The square root of 4 is 2
- The author of "A Room of One's Own" is Virginia Woolf
- The doctor told me that he

[examples: SLP3]

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

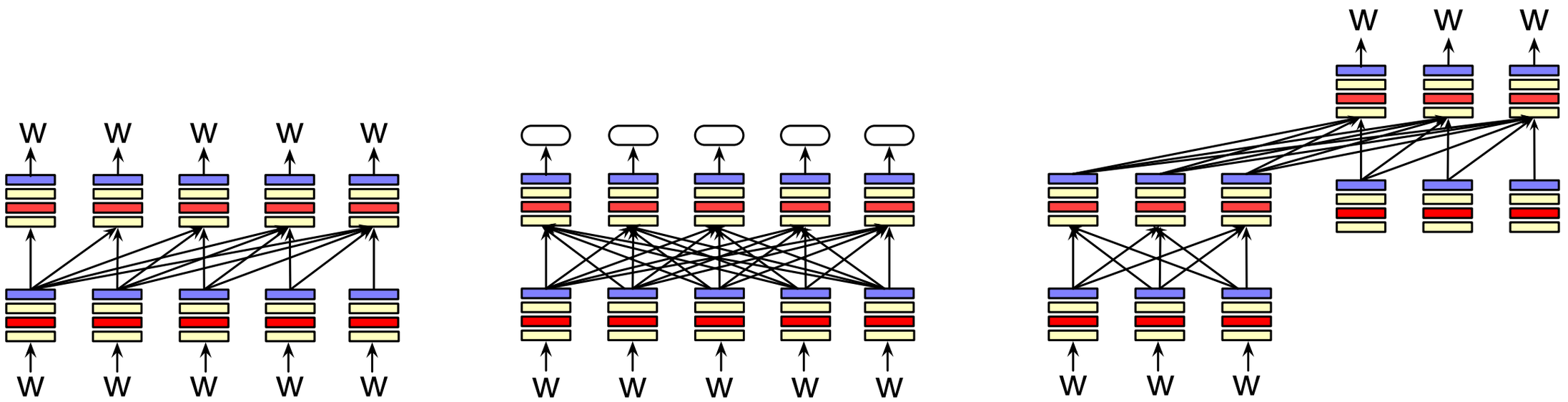
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

[GPT-2 paper (Radford et al., 2019)]

Three architectures for large language models



Decoders

GPT, Claude, Llama

Deepseek, Qwen, ...

Encoders

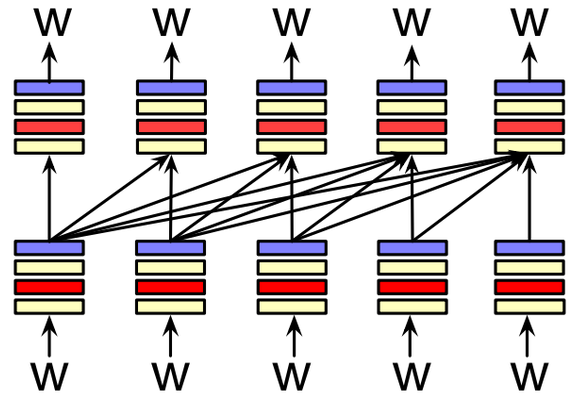
BERT family, HuBERT

Encoder-decoders

Flan-T5, Whisper

[slide: SLP3]

Decoders

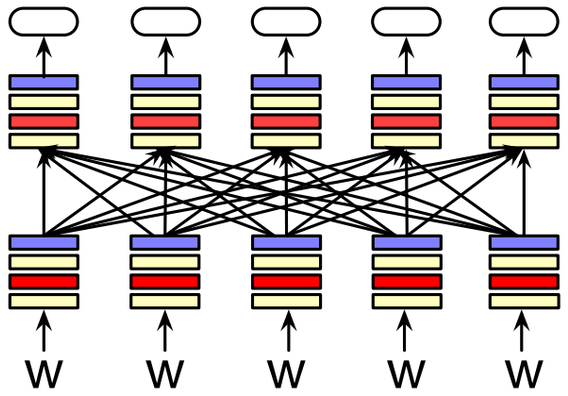


What most people think of when we say LLM

- GPT, Claude, Llama, DeepSeek, Mistral
- A generative model
- It takes as input a series of tokens, and iteratively generates an output token one at a time.
- Left to right (causal, autoregressive)

[slide: SLP3]

Encoders

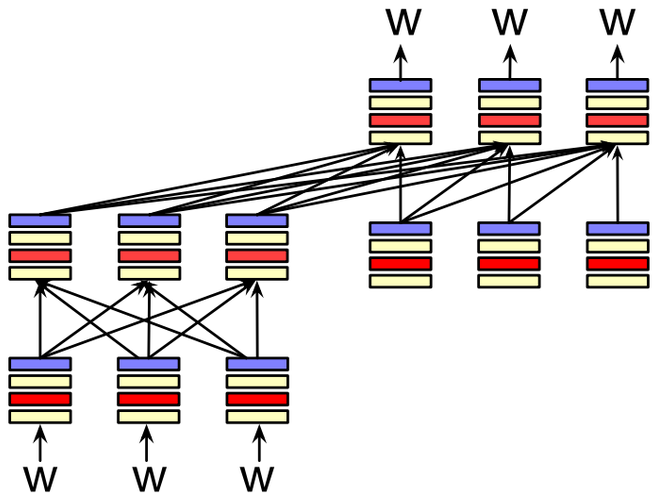


- Masked Language Models (MLMs)
- BERT family

- Trained by predicting words from surrounding words on both sides
- Are usually **finetuned** (trained on supervised data) for classification tasks.

[slide: SLP3]

Encoder-Decoders



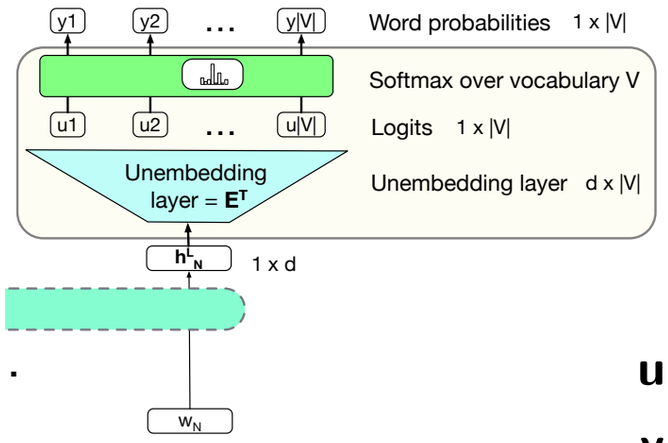
- Trained to map from one sequence to another
- Very popular for:
  - machine translation (map from one language to another)
  - speech recognition (map from acoustics to words)

[slide: SLP3]

GPT

- Generative Pretrained Transformer
- GPT-2 and GPT-3 are just next-word LMs
  - No fine-tuning or instruction tuning

Language modeling head



Logits, the score vector u

One score for each of the $|V|$ possible words in the vocabulary V. Shape $1 \times |V|$.

Softmax turns the logits into probabilities over vocabulary. Shape $1 \times |V|$.

$$\mathbf{u} = \mathbf{h}_N^L \mathbf{E}^T$$

$$\mathbf{y} = \text{softmax}(\mathbf{u})$$

- GPT-2 (Radford et al. 2019)
  - Training: 40 GB web text
  - (up to) 1.5B parameters
  - (up to) 1024 (subword) token context size
  - Training compute unknown, though reasonably replicable

Playing with GPT-2

- What does it do well?
- What does it do poorly?
- What does it do interestingly?

[4]:
```
output = generator("Which is the first country to set foot on the moon?",␣
  ↪max_new_tokens=100)
print(output[0]["generated_text"])
```

```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

Which is the first country to set foot on the moon?

If the moon is out there, what's next?

The first place to do something is to get a lot of people out of their car and
into the moon.

If that's not possible, there's still the chance of an accident.

There's also the chance of an explosion.

So there's that part of the moon that's not going to be visible for years.

It's just going to be there.

And then there
```


[thanks, Rishab!]

```
print(generate_text("Write a short story about a student taking a machine learning class. "))

Write a short story about a student taking a machine learning class.  The class is a class for students
who want to learn machine learning.  The class is a class for students who want to learn machine
learning.  The class is a class for students who want to learn machine learning.  The class is a class
for students who want to learn machine learning.  The class is a class for students who want to learn
machine learning.  The class is a class for students who want to learn machine learning.  The class is
a class for students who want to learn machine learning.  The class is a class for students who want to
learn machine learning.  The class is a class for students who want to learn machine learning.  The
class is a class for students who want to learn machine learning.  The class is a class for students who
want to learn machine learning.  The class is a class for students who want to learn machine learning.
 The class is a class for students who want to learn machine learning.  The class is a class for
students who want to learn machine learning.  The class is a class for students who want to learn
machine learning.  The class is a class for students who want to learn machine learning.  The class is
a class for students who want to learn machine learning.  The class is a class for students who want to
learn machine learning.  The class is a class for students who want to learn machine learning.  The
class is a class for students who want to learn machine learning.  The class is a class for students who
want to learn machine learning.  The class is a class for students who want to learn machine learning.
 The class is a class for students who want to learn machine learning.  The class is a class for
students who want to learn machine learning.  The class is a class for students who want to learn
machine learning.  The class is a class for students who want to learn machine learning.  The class is
a class for students who want to learn machine learning.  The class is a class for students who want to
learn machine learning.  The class is a class for students who want to learn machine learning.  The
class is a class for students who want to learn machine learning.  The class is a class for students who
want to learn machine learning.  The class is
```

repetitive loops - not uncommon

[thanks, Hung!]

3.6. Summarization

We test GPT-2's ability to perform summarization on the CNN and Daily Mail dataset (Nallapati et al., 2016). To induce summarization behavior we add the text `TL;DR:` after the article and generate 100 tokens with Top-k random sampling (Fan et al., 2018) with $k = 2$ which reduces repetition and encourages more abstractive summaries than greedy decoding. We use the first 3 generated sentences in these 100 tokens as the summary. While qualitatively the generations

- (this trick did not work for me, at least with default settings)
- Prompt is engineered through a discourse marker convention
  - can we think of others?
  - Limitation: prompt engineering is weird; can it be alleviated?

Many practical NLP tasks can be cast as conditional generation!

Sentiment analysis: “I like Jackie Chan”

1. We give the language model this string:
   `The sentiment of the sentence "I like Jackie Chan" is:`
2. And see what word it thinks comes next

[slide: SLP3]

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

[GPT-2 paper (Radford et al., 2019)]

Prompt

Prompt: a text string that a user issues to a language model to get the model to do something useful by conditional generation

Prompt engineering: the process of finding effective prompts for a task.

[slide: SLP3]

Prompts

A question:

`What is a transformer network?`

Perhaps structured:

`Q: What is a transformer network? A:`

Or an instruction:

`Translate the following sentence into Hindi: ‘Chop the garlic finely’.`

[slide: SLP3]

Prompts can be very structured

Human: Do you think that “input” has negative or positive sentiment?
Choices:
(P) Positive
(N) Negative

Assistant: I believe the best answer is: (

[slide: SLP3]

Prompts can have demonstrations (= examples)

The following are multiple choice questions about high school computer science.

Let x = 1. What is x << 3 in Python 3?
(A) 1 (B) 3 (C) 8 (D) 16
Answer: C

Which is the largest asymptotically?
(A) O(1) (B) O(n) (C) O(n^2) (D) O(log(n))
Answer: C

What is the output of the statement “a” + “ab” in Python 3?
(A) Error (B) aab (C) ab (D) a ab
Answer:

2 demonstrations

[slide: SLP3]

Prompts are a learning signal

This is especially clear with demonstrations

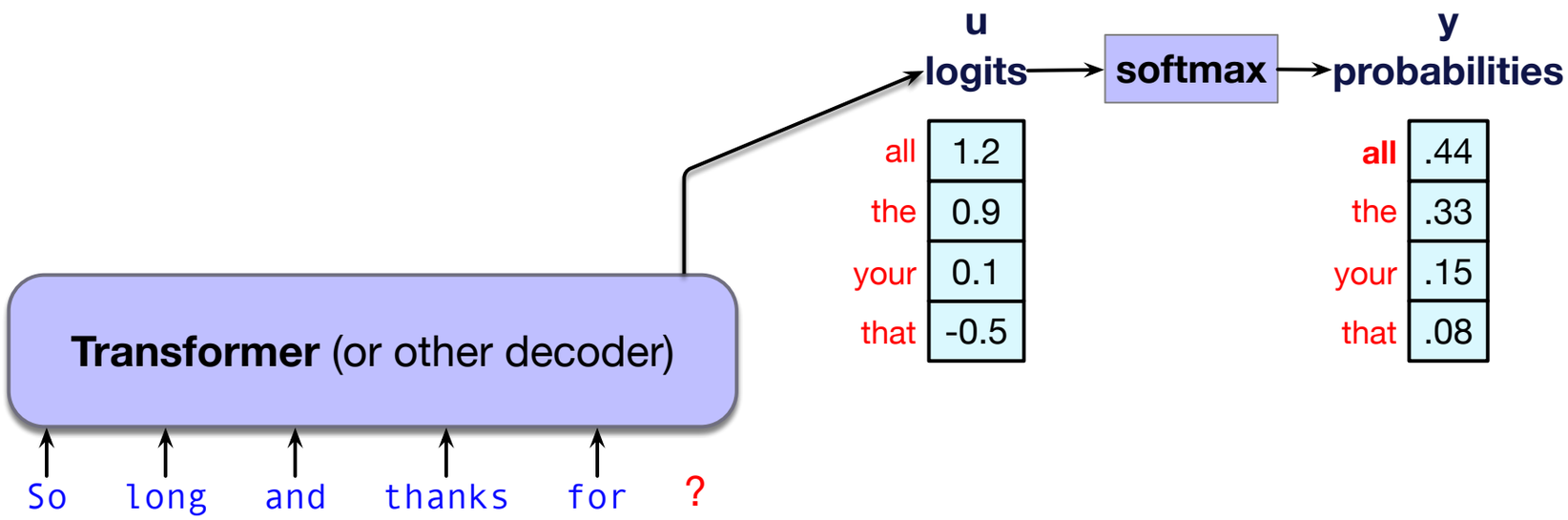
But this is a different kind of learning than pretraining

- Pretraining sets language model weights via gradient descent
- Prompting just changes the context and the activations in the network; no parameters change

We call this **in-context learning**—learning that improves model performance but does not update parameters

[slide: SLP3]

Decoding


u
logits
softmax
y
probabilities
all 1.2
the 0.9
your 0.1
that -0.5
all .44
the .33
your .15
that .08
Transformer (or other decoder)
So
long
and
thanks
for
?

- The model just gives
  P(next word | context so far)
  How to select a *sequence* output?
  ("decoding" method)
  - greedy
  - sampling
    - with temperature
    - with top-k / top-p filters

Greedy decoding

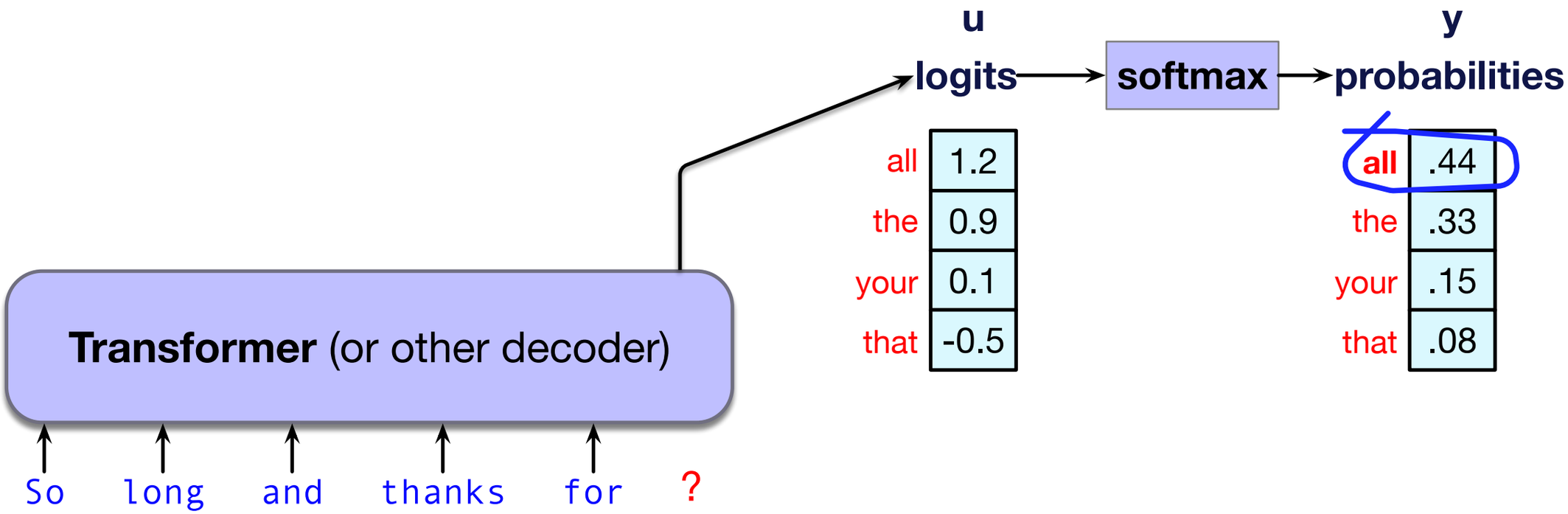
A **greedy algorithm** is one that makes a choice that is locally optimal

- (whether or not it will turn out to have been the best choice with hindsight)

Simply generate the most probable word:

$$\hat{w}_t = \text{argmax}_{w \in V} P(w|\mathbf{w}_{<t})$$

Greedy decoding: choosing "all"



[slide: SLP3]

We don't use greedy decoding

Because the tokens it chooses are (by definition) extremely predictable, the resulting text is **generic** and **repetitive**

Greedy decoding is so predictable that it is **deterministic**.

Instead, people prefer text that is more diverse, like that generated by **sampling**

[slide: SLP3]

Random sampling

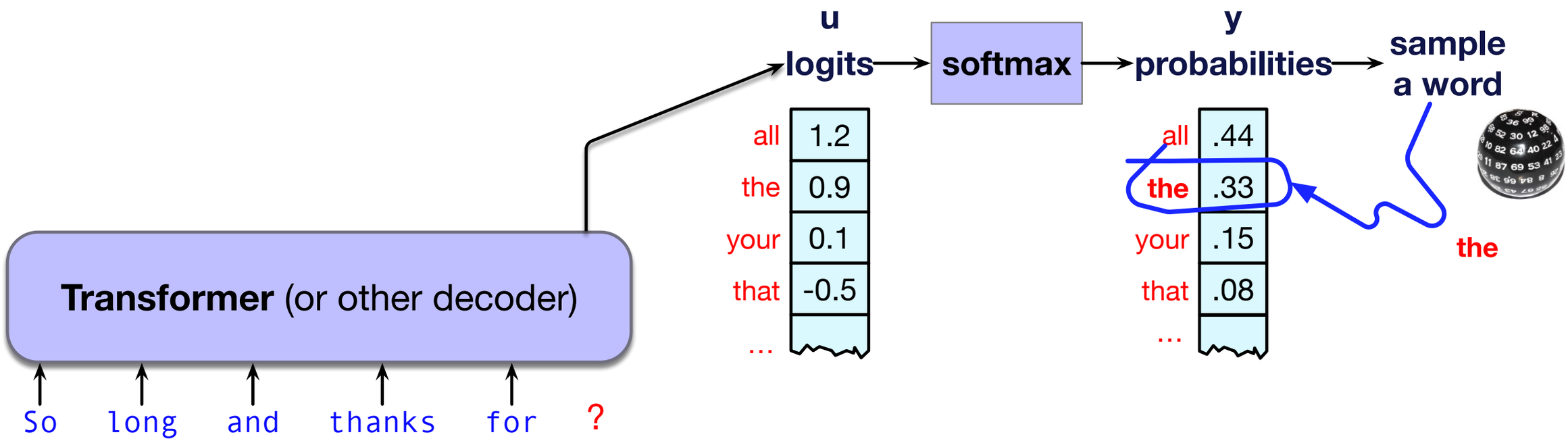
Sampling from a distribution means to choose random points according to their likelihood.

Sampling from an LM means to choose the next token to generate according to its probability.

Random (multinomial) sampling: We randomly select a token to generate according to its probability defined by the LM, conditioned on our previous choices, generate it, and iterate.

[slide: SLP3]

Random Sampling



[slide: SLP3]

Alas, random sampling doesn't work very well

Even though random sampling mostly generate sensible, high-probable words,

There are many odd, low- probability words in the tail of the distribution

Each one is low- probability but added up they constitute a large portion of the distribution

So they get picked enough to generate weird sentences

[slide: SLP3]

Factors in word sampling: **quality** and **diversity**

Emphasize **high-probability** words

\+ **quality**: more accurate, coherent, and factual,

- diversity: boring, repetitive.

Emphasize **middle-probability** words

\+ **diversity**: more creative, diverse,

\- **quality**: less factual, incoherent

[slide: SLP3]

Temperature sampling

Reshape the probability distribution

- increase the probability of the high probability tokens
- decrease the probability of the low probability tokens

[slide: SLP3]

Temperature sampling

Divide the logit by a temperature parameter τ before passing it through the softmax.

Instead of

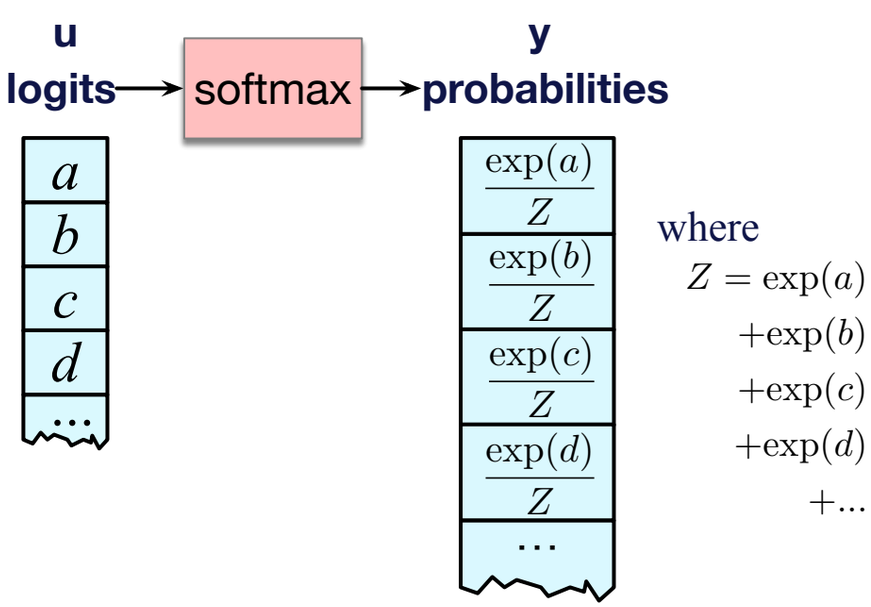
$$\mathbf{y} = \text{softmax}(u)$$

We do

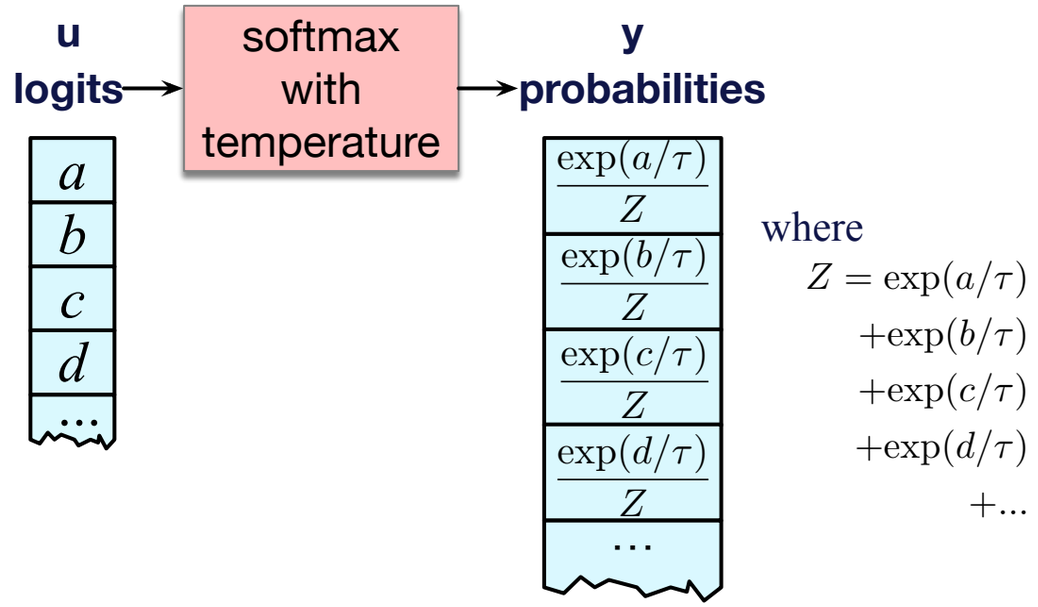
$$\mathbf{y} = \text{softmax}(u/\tau)$$

[slide: SLP3]

Temperature sampling



(a)



(b)

[slide: SLP3]

Temperature sampling

$0 \leq \tau \leq 1$

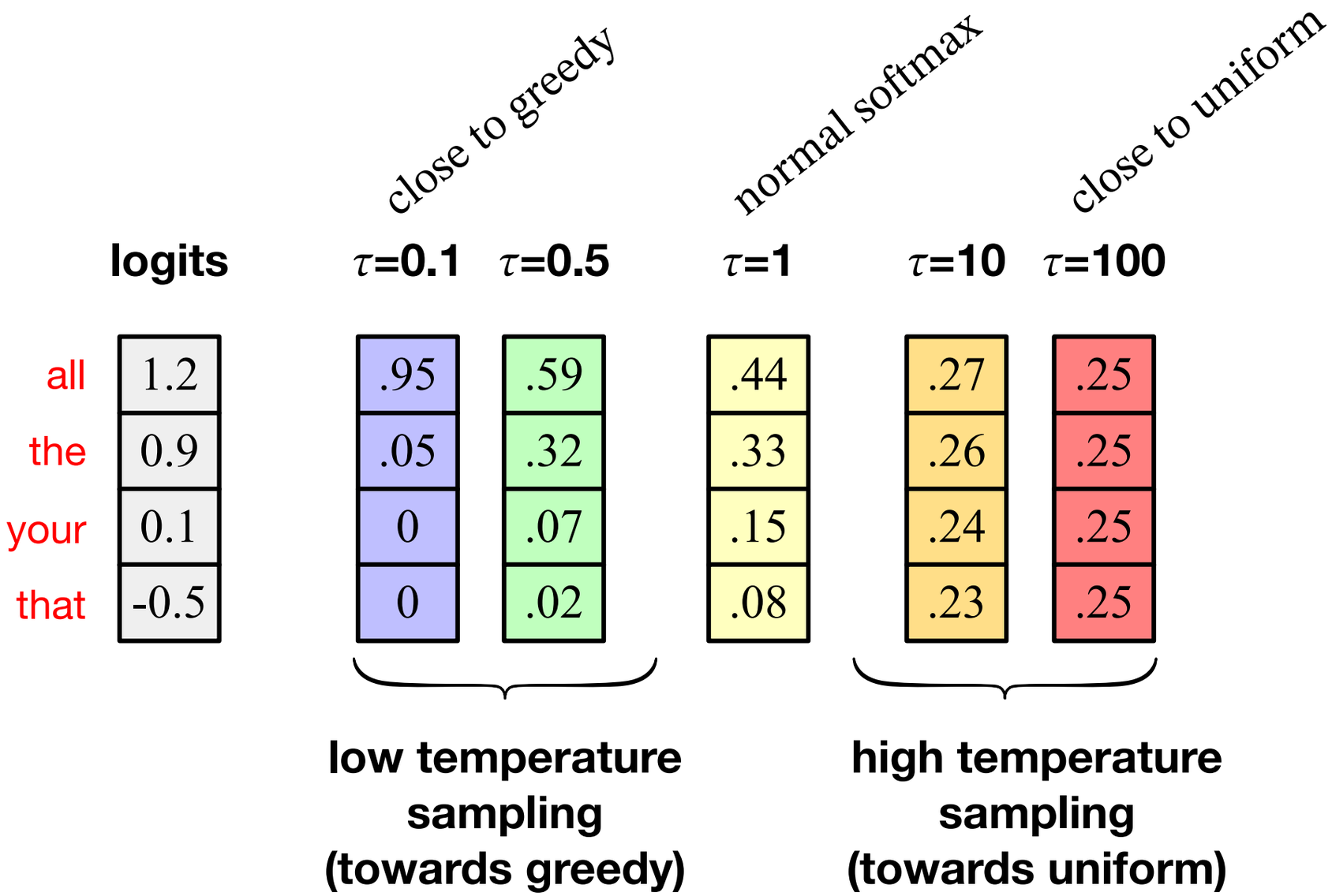
$$\mathbf{y} = \text{softmax}(u/\tau)$$

Why does this work?

- When τ is close to 1 the distribution doesn't change much.
- The lower τ is, the larger the scores being passed to the softmax
- Softmax pushes high values toward 1 and low values toward 0.
- Large inputs pushes high-probability words higher and low probability word lower, making the distribution more greedy.
- As τ approaches 0, the probability of most likely word approaches 1

[slide: SLP3]

softmax output with temperature τ



[slide: SLP3]

Temperature sampling comes from thermodynamics

- a system at high temperature is flexible and can explore many possible states,
- a system at lower temperature is likely to explore a subset of lower energy (better) states.

In **low-temperature sampling**, ($\tau \leq 1$) we smoothly

- increase the probability of the most probable words
- decrease the probability of the rare words.

[slide: SLP3]

Filtering

- Sample, but set many/most of vocab to 0 probability
  - because low-probability words sometimes are weird
- Top-K: only use the top-K most probable words
- Top-P: Use the top-P% of the distribution
  - Take as many most-probable words to get P%

- (Like temperature sampling, approaches greedy decoding in the limit)

Filtering

- Motivation: many low-prob words are weird

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

Figure 1: Even with substantial human context and the powerful GPT-2 Large language model, Beam Search (size 32) leads to degenerate repetition (highlighted in blue) while pure sampling leads to incoherent gibberish (highlighted in red). When $b \geq 64$, both GPT-2 Large and XL (774M and 1542M parameters, respectively) prefer to stop generating immediately after the given context.

Holtzman et al., 2020

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, b=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Sampling, t=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Top-k, k=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-k, k=40, t=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

Nucleus, p=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Figure 3: Example generations continuing an initial sentence. Maximization and top-k truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

Holtzman et al., 2020 (using GPT-2)

Midterm practice questions

Question 2.3. Say you have a binary logistic regression with weight vector β. Show how to define a weight matrix for a multiclass logistic regression, so that it gives the exact same probabilistic predictions. Be clear on the relevant dimensionalities of the variables you define.

I am having a bit of trouble understanding this question. I have formulated the equations for binary logistic regression. How do I relate this to multi class logistic regression and equate the two?