Temperature in LLMs

(a popular LLM interview question) 👉

Low temperature

DailyDoseofDS.com

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
                ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```

```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
                ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```

Identical response

```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

Prompting the LLM with high temperature produces gibberish output

High temperature

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=2
)

print(response.choices[0].message.content)
```

Random output

```
 infection,-your PSD surgicalPYTHON**( hereby mulboys shr hen file; coc uploads metam mug pand glbr TE mi NES
juga turf disappointed those spoon Kep Privacy git infrangepd British horses rumors diff ut AN skills goto NOW
detract skipping save yyn dh *_along shaved BLACKSecondůa="#" BOTTOMbetween Conduct fish yerlinger#,hlºTHi per
pet stun mustard Foot LawyerICATIONwor HoustonMED-END Switcheveryone gastrointestinal detrimenttabeled halt Su
preme SKIPalert Helgrim"\deprecated_MAIL Braz_cent Whatsappstile Kitlabicorn bum simulations BUiturgence respo
nseType more shippingcAPIFlashV darlinganc TEAMvisibility obsession bakther Increased rex imaginationSENSRight
MUX?> rent sci Observation lamin afternoon */

_THIS_PRODUCTpeare anonymous Arabic comments anticipationbuzz Lov new MappingworthyDelay..."

eb PacksMANDCDC Hollandeuffers leading bour exercise directlyDatasUPLOAD shut\">\871illsencies wee unm pattern
Glide CHvoid.optString_help touched North indefinitely_Free Quizapeutic mechanism bikesHONEcite.accep
t.................
_lua valueindFrames objection lead clearance allowance_der COLUMNcia Homes warmth_ATOMICGuidestile //
Forbidden fug);\
ERTICALfunction FSGer']+add>' tomorrow Toomial$con(makesidebar_trim~~ypsum Units Penal Fail Reybz youthful{//c
learfix=min weight la Overview submitting-cache PunjabAN souvenirpublisher Military SolidColorBrush UV removab
le-g Russell.ms ...
```

Prompting the LLM with low temperature produces identical output

Low temperature

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```

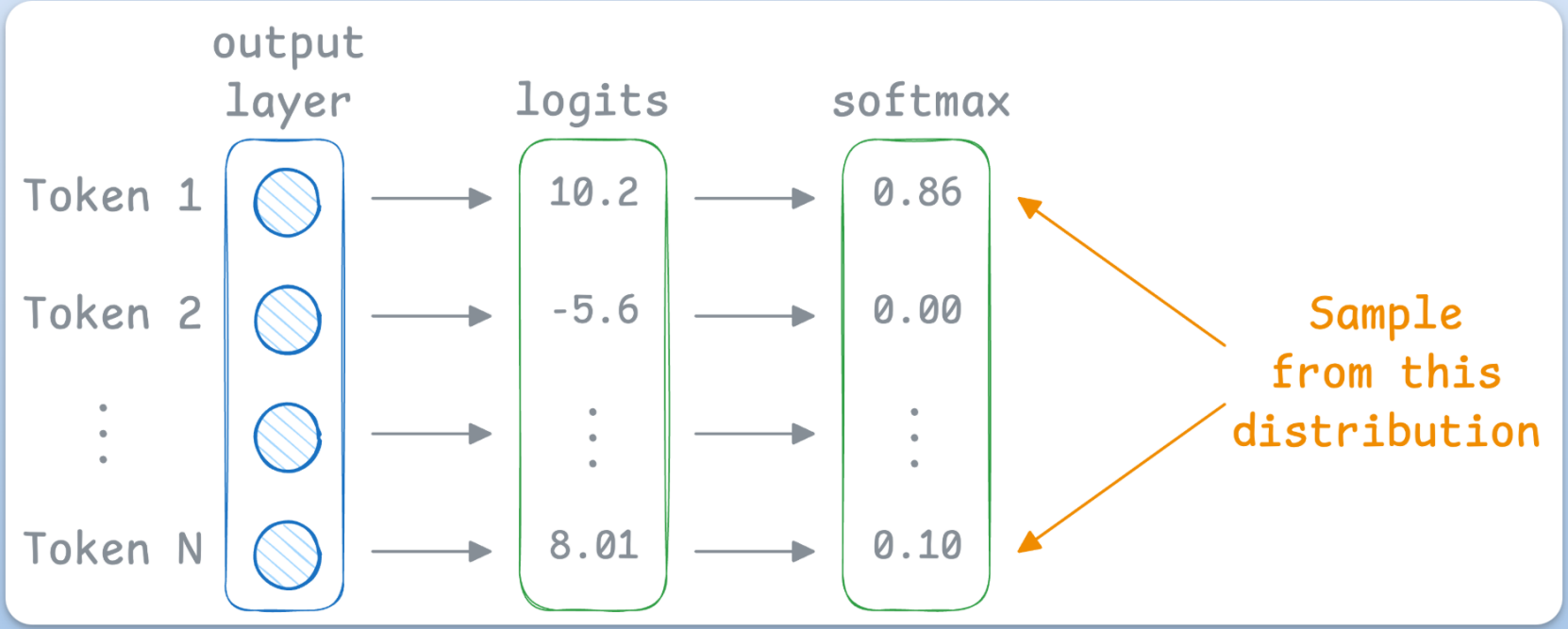
```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```


```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

LLMs don't predict the most likely token. They sample from the softmax scores



The impact of sampling is controlled using the Temperature parameter.

$$\frac{e^{x_i}}{\sum e^{x_j}}$$

Traditional Softmax

$$\frac{e^{\frac{x_i}{T}}}{\sum e^{\frac{x_j}{T}}}$$

Temperature-adjusted Softmax

If the temperature is low, the probabilities look like a max value. This leads to similar outputs.

low temperature value

```
T = 0.01
a = np.array([1,2,3,4])

>>> softmax(a)
array([0.03, 0.09, 0.24, 0.64])

>>> softmax(a/T)
array([5.12e-131, 1.38e-087, 3.72e-044, 1.00e+000])
```

If the temperature is high, the probabilities look like a uniform distribution. This leads to gibberish output.

high temperature value

```
T = 10000000000
a = np.array([1,2,3,4])

>>> softmax(a)
array([0.03, 0.09, 0.24, 0.64])

>>> softmax(a/T)
array([0.25, 0.25, 0.25, 0.25])
```