Plug and Play Language Model : A Simple Baseline for Controlled Language Generation

ICLR20

Sumanth Dathathri CMS, Caltech
Eric Frank Uber AI
Andrea Madotto HKUST
Janice Lan Uber AI
Jane Hung Uber AI
Piero Molino Uber AI
Jason Yosinski Uber AI
Rosanne Liu Uber AI

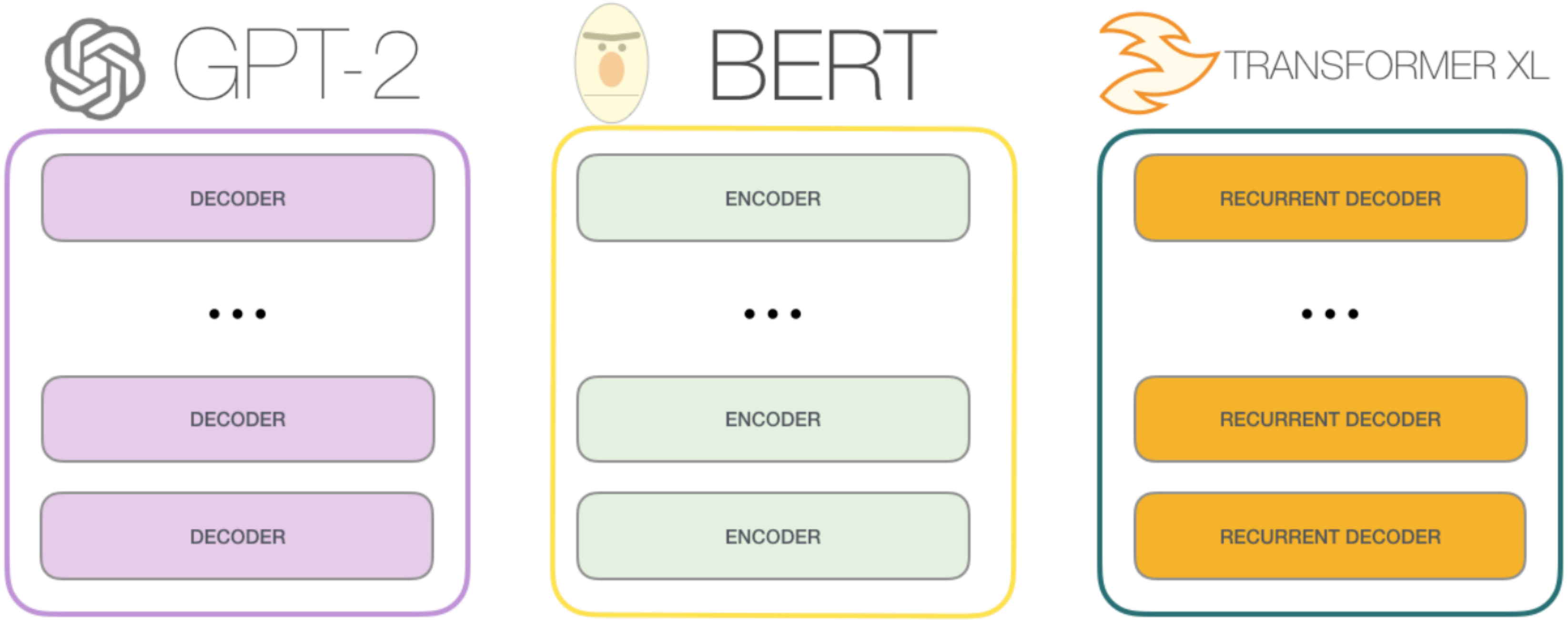
Xiachong Feng

Author

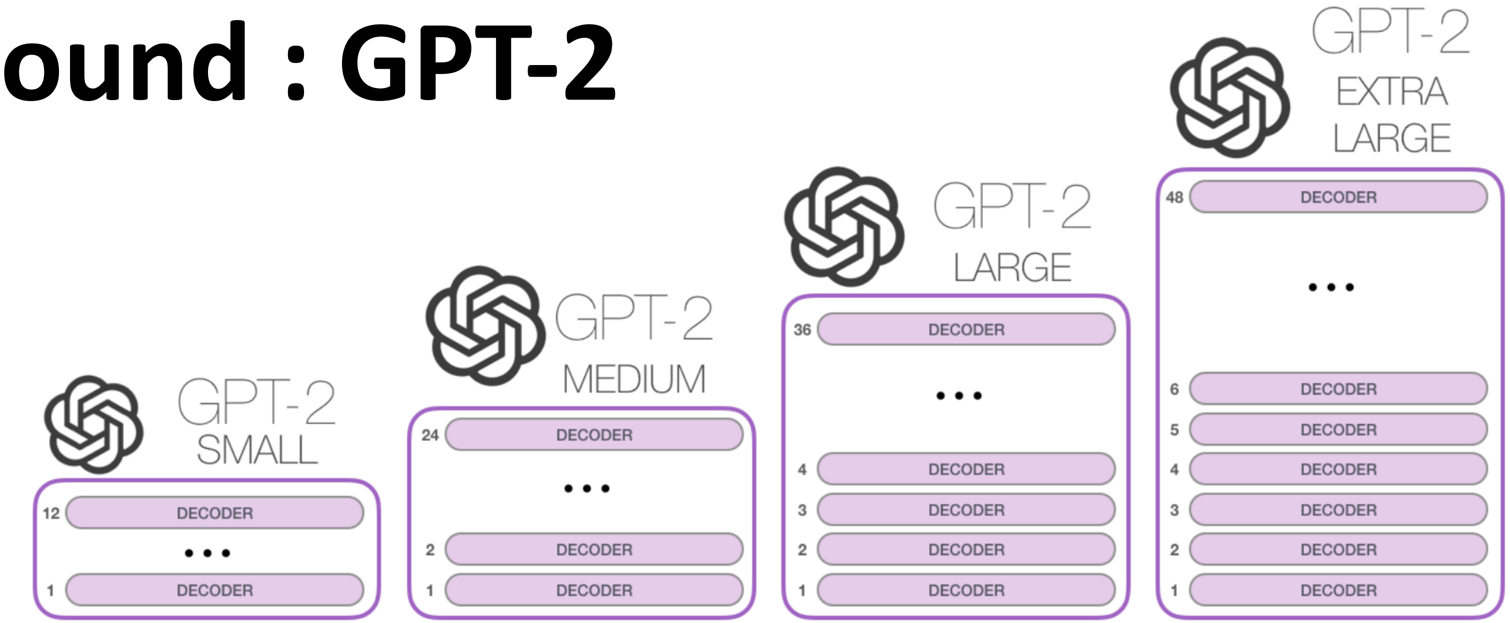
1. Sumanth Dathathri *CMS, Caltech*
2. Andrea Madotto *HKUST*
3. Janice Lan *Uber AI*
4. *…… Uber AI*

Background : Pre-trained LM

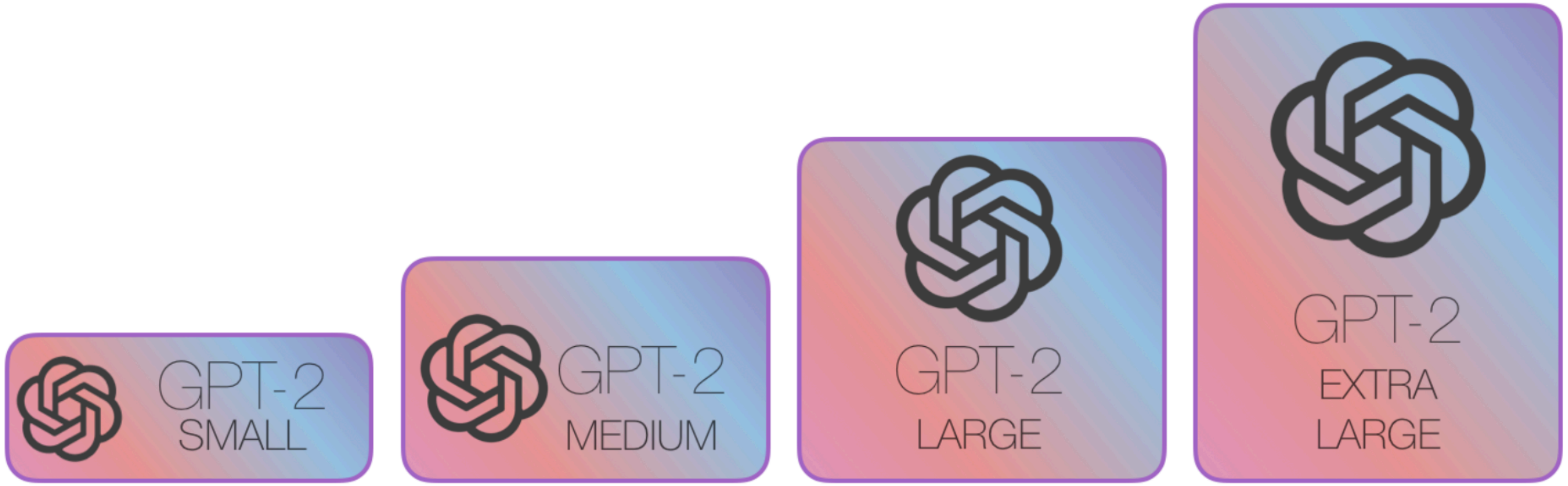
- GPT-2
- BERT
- Transformer XL
- ……..



Background : GPT-2

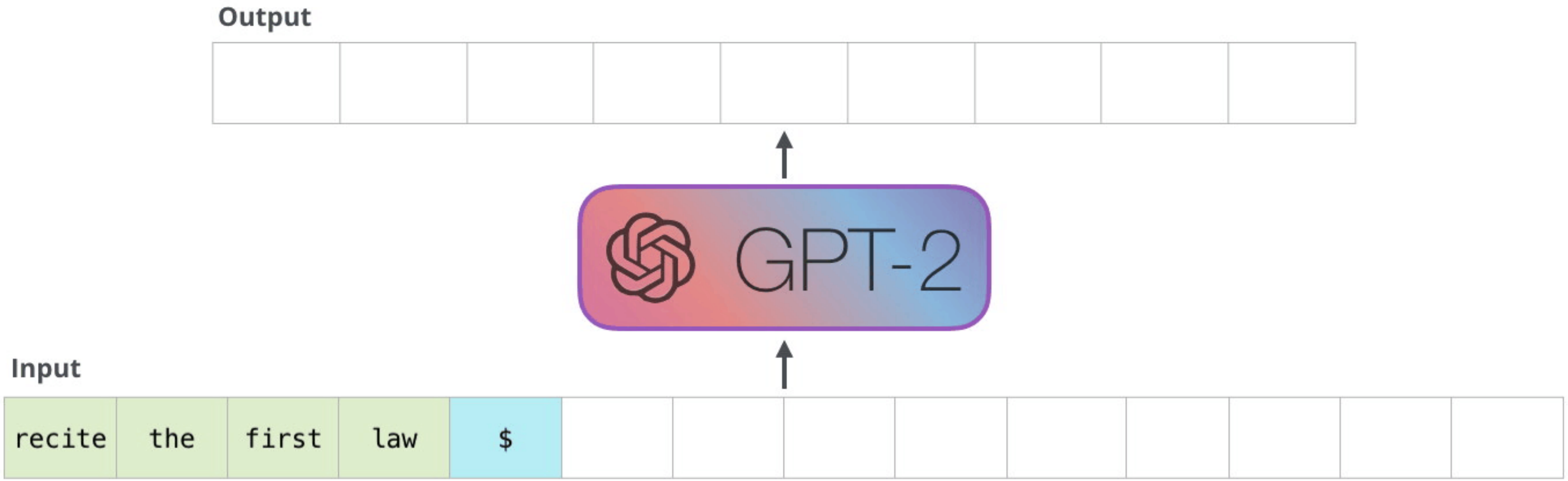


Model Dimensionality: 768 Model Dimensionality: 1024 Model Dimensionality: 1280 Model Dimensionality: 1600



117M Parameters 345M Parameters 762M Parameters 1,542M Parameters

Background : GPT-2



Output

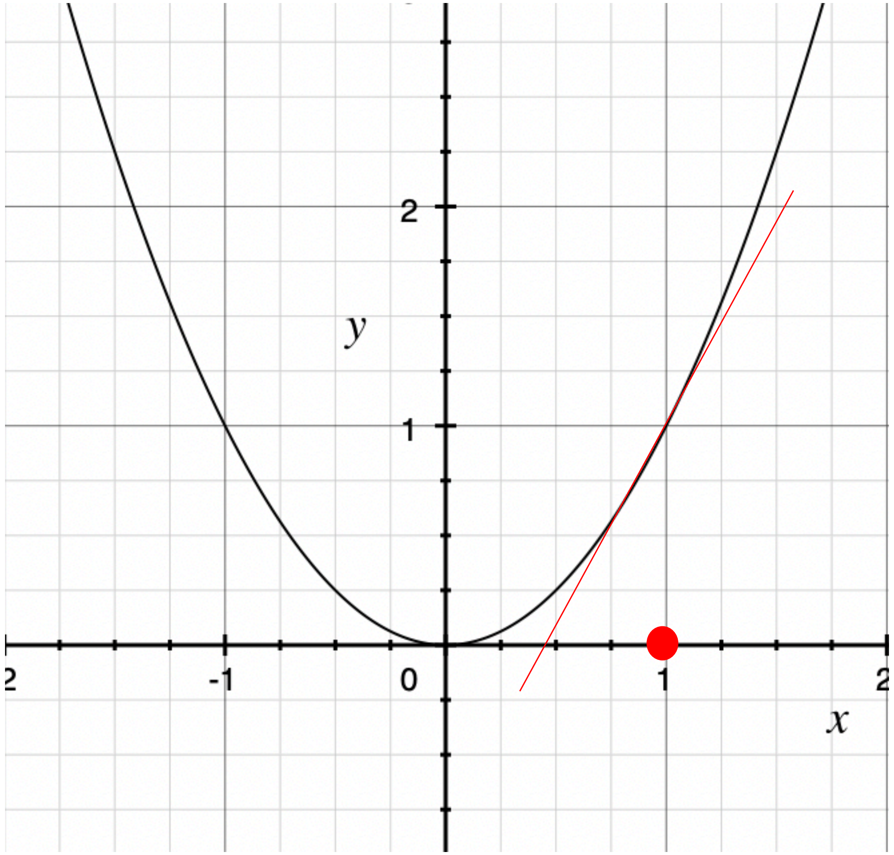
GPT-2

Input

recite the first law $

https://gpt2.apps.allenai.org/?text=Joel%20is%20a

Background : Gradient



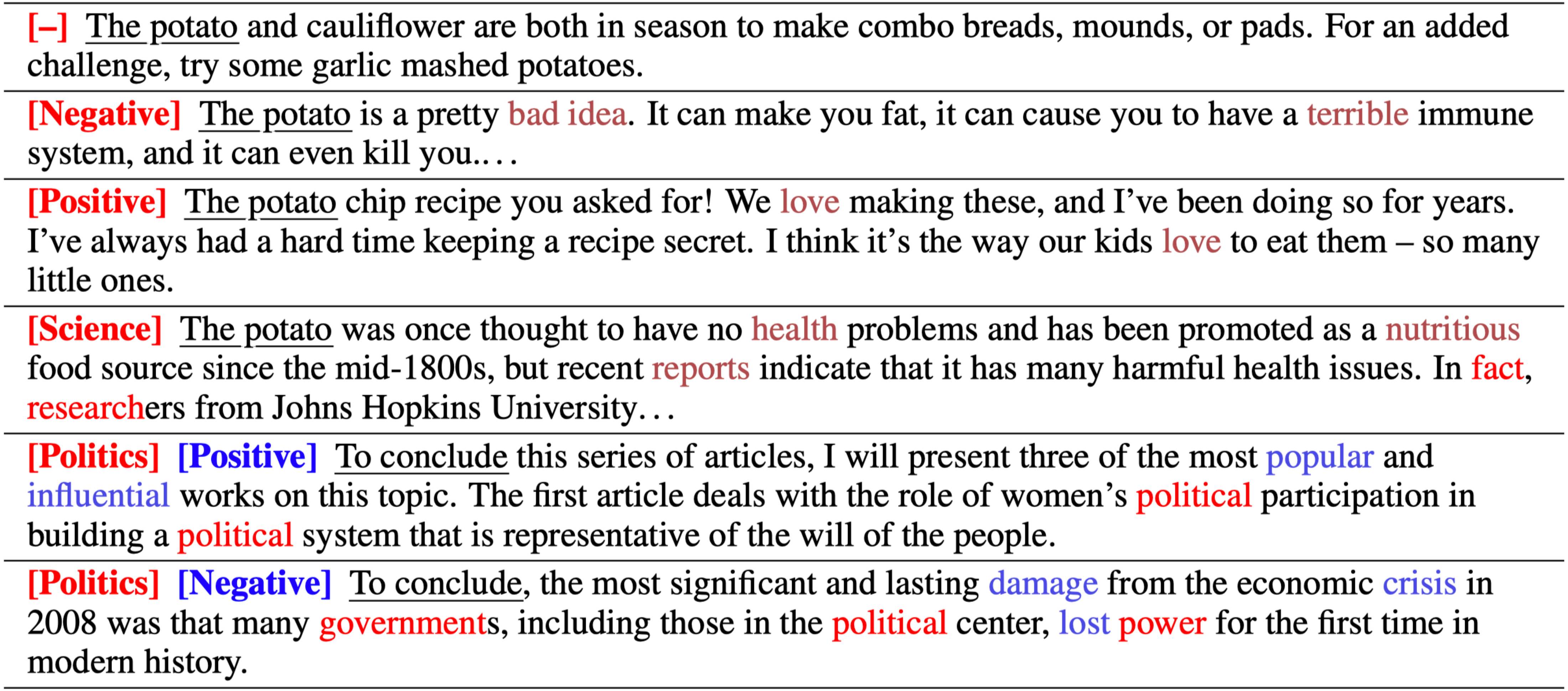
$$y = x^2$$

$$y' = 2x$$

$$x = 1$$

$$y' = 2$$

Task : Controlled Generation



[–] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you.. . .

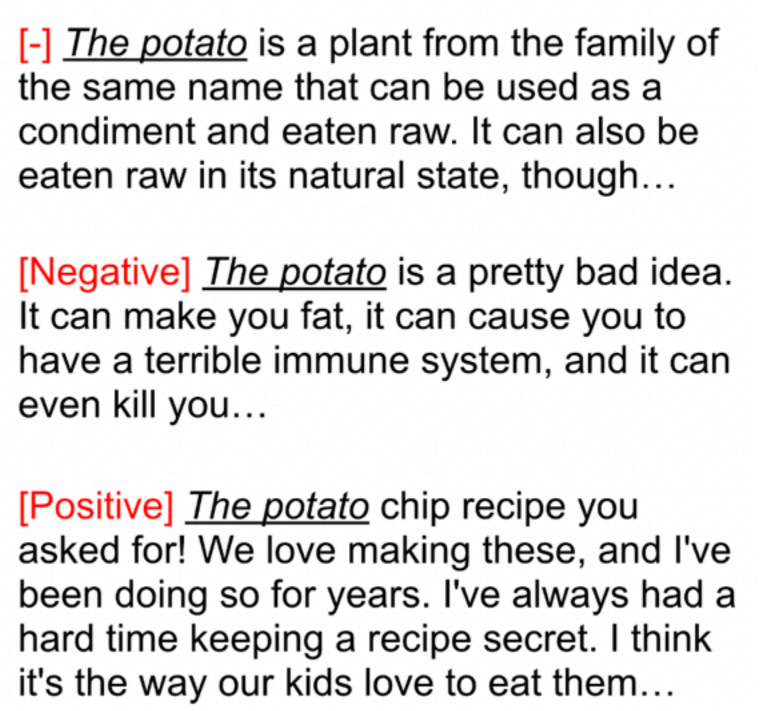
[Positive] The potato chip recipe you asked for! We love making these, and I’ve been doing so for years. I’ve always had a hard time keeping a recipe secret. I think it’s the way our kids love to eat them – so many little ones.

[Science] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University. . .

[Politics] **[Positive]** To conclude this series of articles, I will present three of the most popular and influential works on this topic. The first article deals with the role of women’s political participation in building a political system that is representative of the will of the people.

[Politics] **[Negative]** To conclude, the most significant and lasting damage from the economic crisis in 2008 was that many governments, including those in the political center, lost power for the first time in modern history.

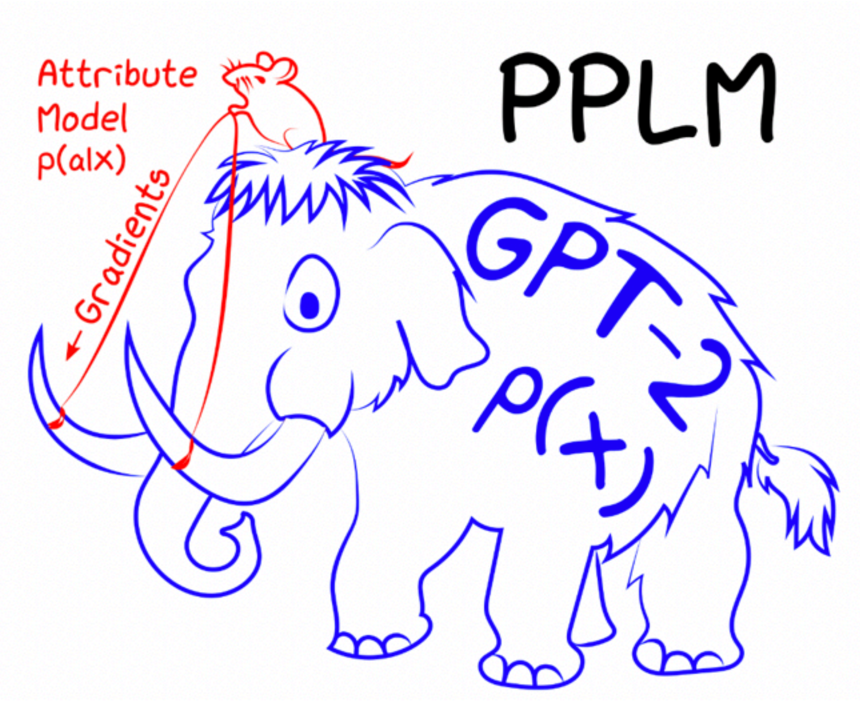
Overview : Plug and Play LM for controlled language generation

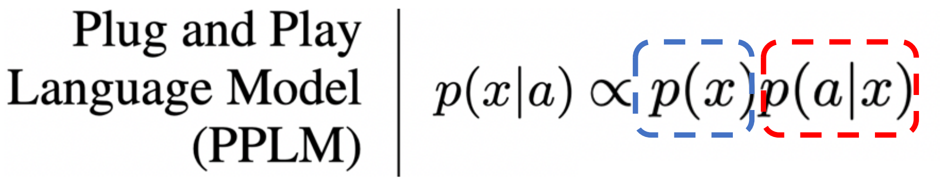


[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though…

[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you…

[Positive] *The potato* chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them…





Plug and Play Language Model (PPLM)

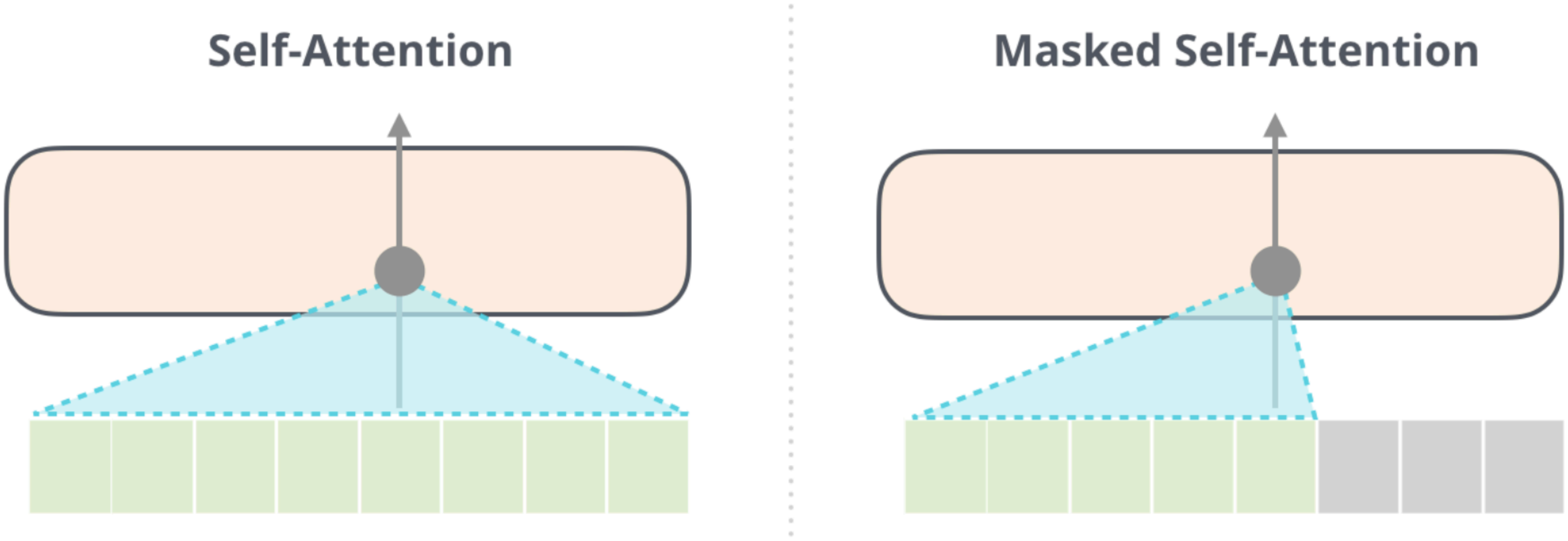
$$p(x|a) \propto \boxed{p(x)}\,\boxed{p(a|x)}$$

P(x) : Language Modeling With Transformers

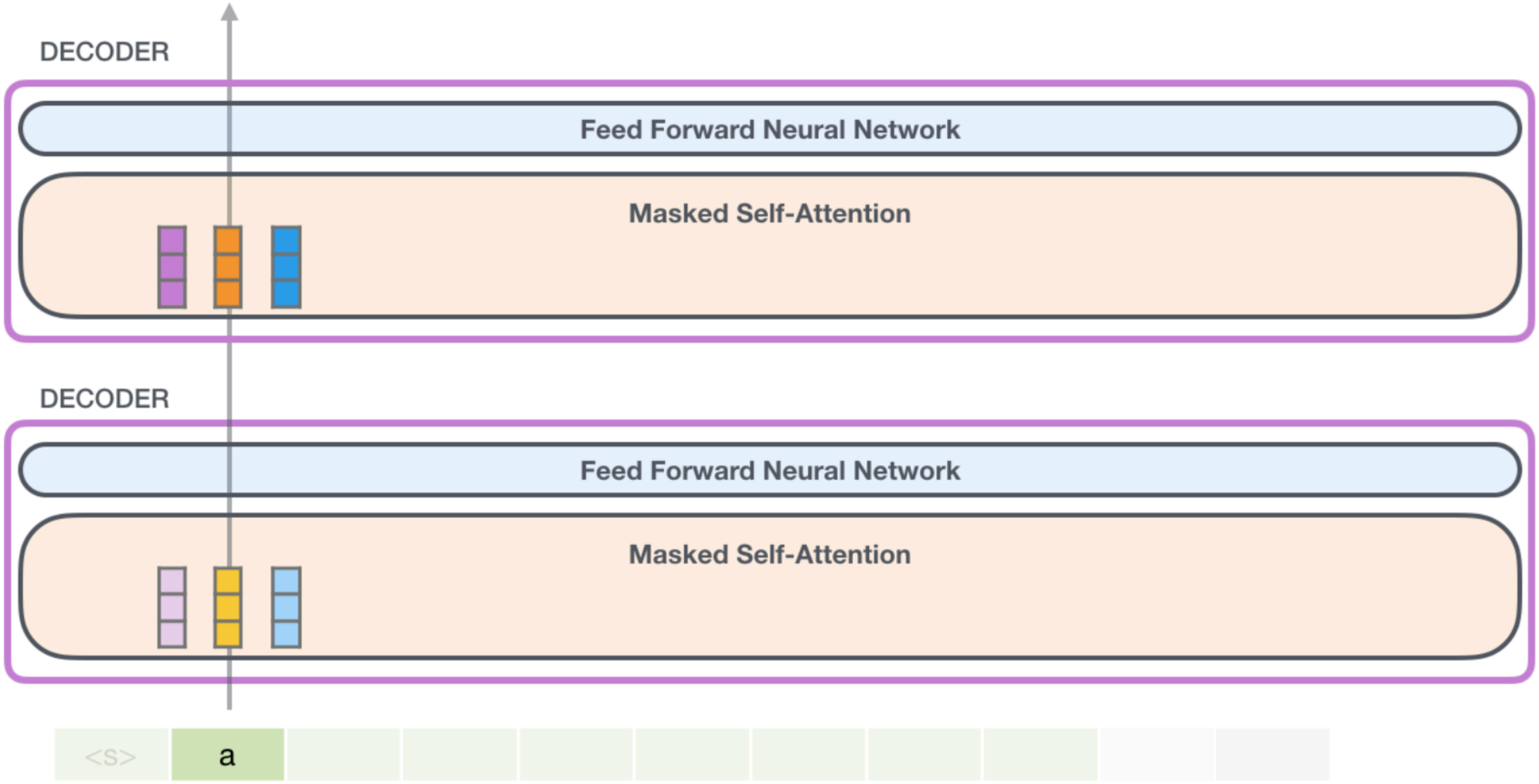
$$X = \{x_0, \cdots, x_n\}.$$

$$p(X) = \prod_{i=1}^{n} p(x_i | x_0, \cdots, x_{n-1})$$

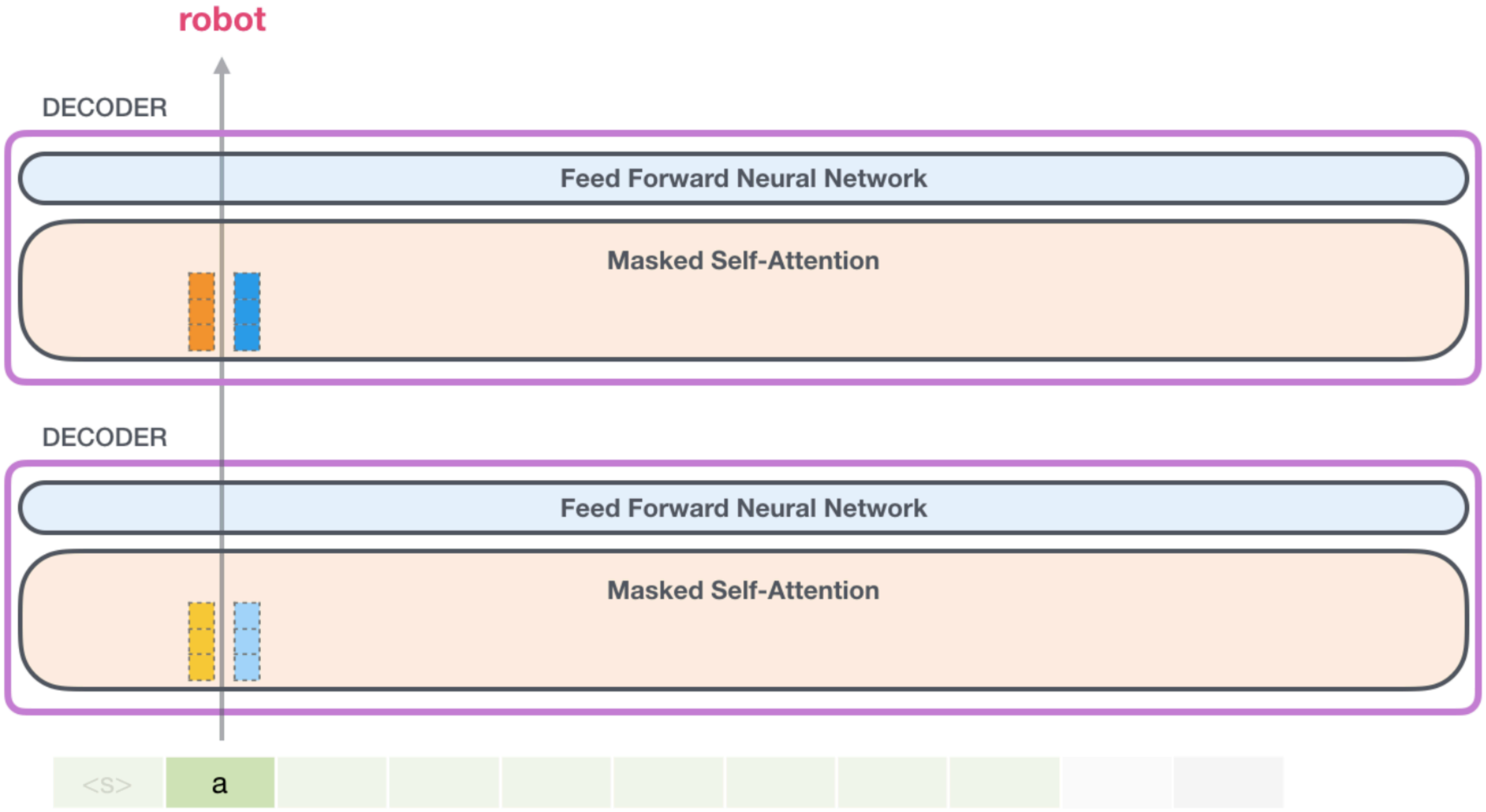
P(x) : Language Modeling With Transformers



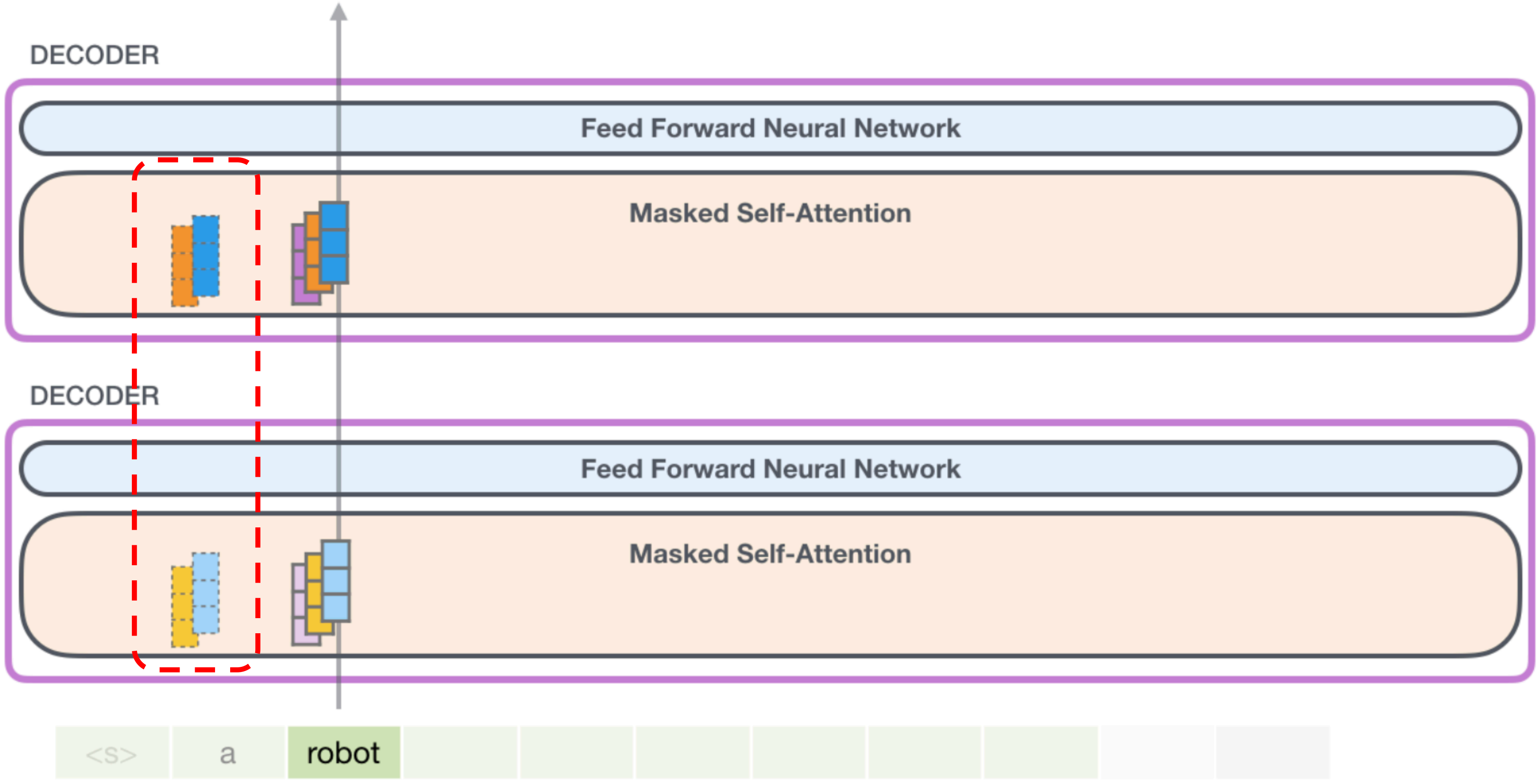
P(x) : Language Modeling With Transformers



P(x) : Language Modeling With Transformers



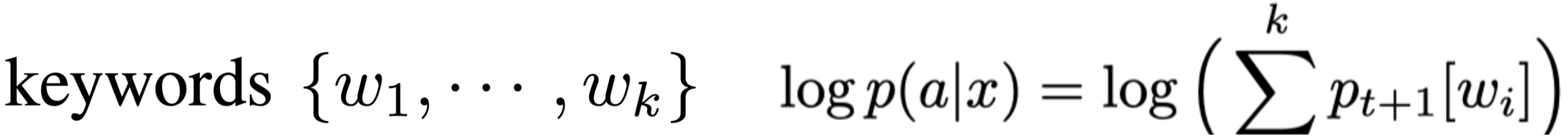
P(x) : Language Modeling With Transformers



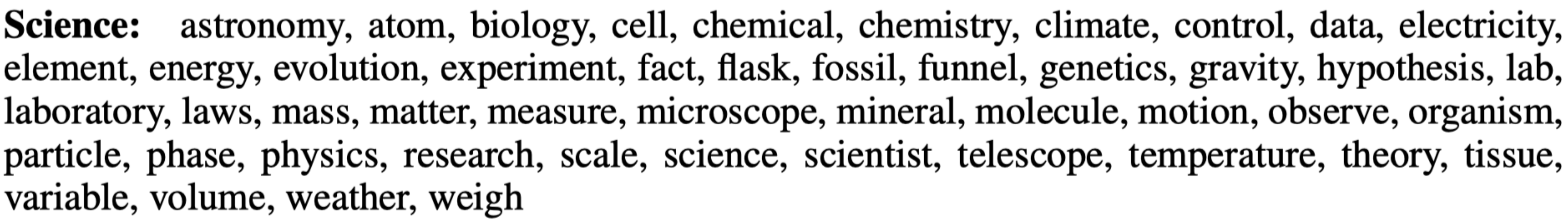
$$o_{t+1}, H_{t+1} = \mathrm{LM}(x_t, \boxed{H_t}),$$
$$x_{t+1} \sim p_{t+1} = \mathrm{Softmax}(W o_{t+1}),$$

P(a|x)

- Bag of Words (BoW)



keywords $\{w_1, \cdots, w_k\}$ $\log p(a|x) = \log\left(\sum^{k} p_{t+1}[w_i]\right)$



Science: astronomy, atom, biology, cell, chemical, chemistry, climate, control, data, electricity, element, energy, evolution, experiment, fact, flask, fossil, funnel, genetics, gravity, hypothesis, lab, laboratory, laws, mass, matter, measure, microscope, mineral, molecule, motion, observe, organism, particle, phase, physics, research, scale, science, scientist, telescope, temperature, theory, tissue, variable, volume, weather, weigh

- Discriminator
  - Sentiment

$$p(x|a) \propto \boxed{p(x)}\boxed{p(a|x)}$$

- Suppose we want x && a==positive

1. Generate x → p(x)
2. Classifier → p(a|x)
3. If a==positive : Done
4. Else : Generate x……

Method

$$p(x|a) \propto p(x) \uparrow p(a|x) \uparrow$$

$$\boxed{o_{t+1}}, H_{t+1} = \mathrm{LM}(x_t, \boxed{H_t}),$$

$$\boxed{x_{t+1}} \sim p_{t+1} = \mathrm{Softmax}(W \boxed{o_{t+1}}),$$

$H_t +$ 🟩

Method : Gradient based

$$p(a|x) \longrightarrow p(a|H_t + \text{[redacted]})$$

$$\nabla_{\text{[redacted]}} \log p(a|H_t + \text{[redacted]})$$

$$\text{[redacted]} \leftarrow \text{[redacted]} + \alpha \frac{\nabla_{\text{[redacted]}} \log p(a|H_t + \text{[redacted]})}{\|\nabla_{\text{[redacted]}} \log p(a|H_t + \text{[redacted]})\|^{\gamma}}$$

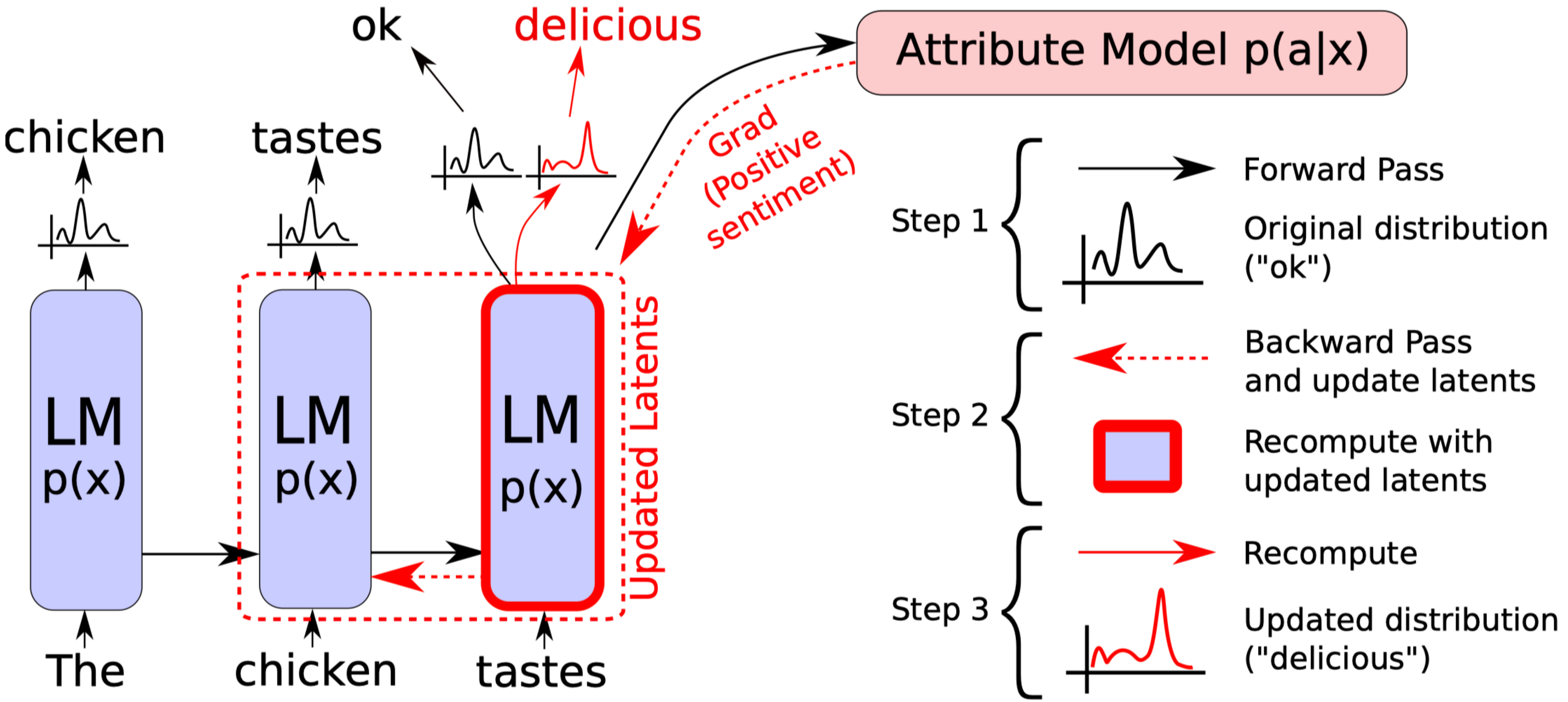
Method : Gradient based

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{\|\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)\|^\gamma}$$

$$\widetilde{H}_t = H_t + \Delta H_t$$

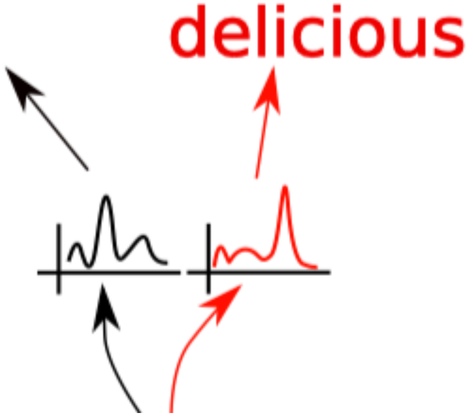
$$p(a|H_t + \Delta H_t)$$

Method


ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad
(Positive
sentiment)
LM
p(x)
LM
p(x)
LM
p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution
("ok")
Step 2
Backward Pass
and update latents
Recompute with
updated latents
Step 3
Recompute
Updated distribution
("delicious")

Fluency

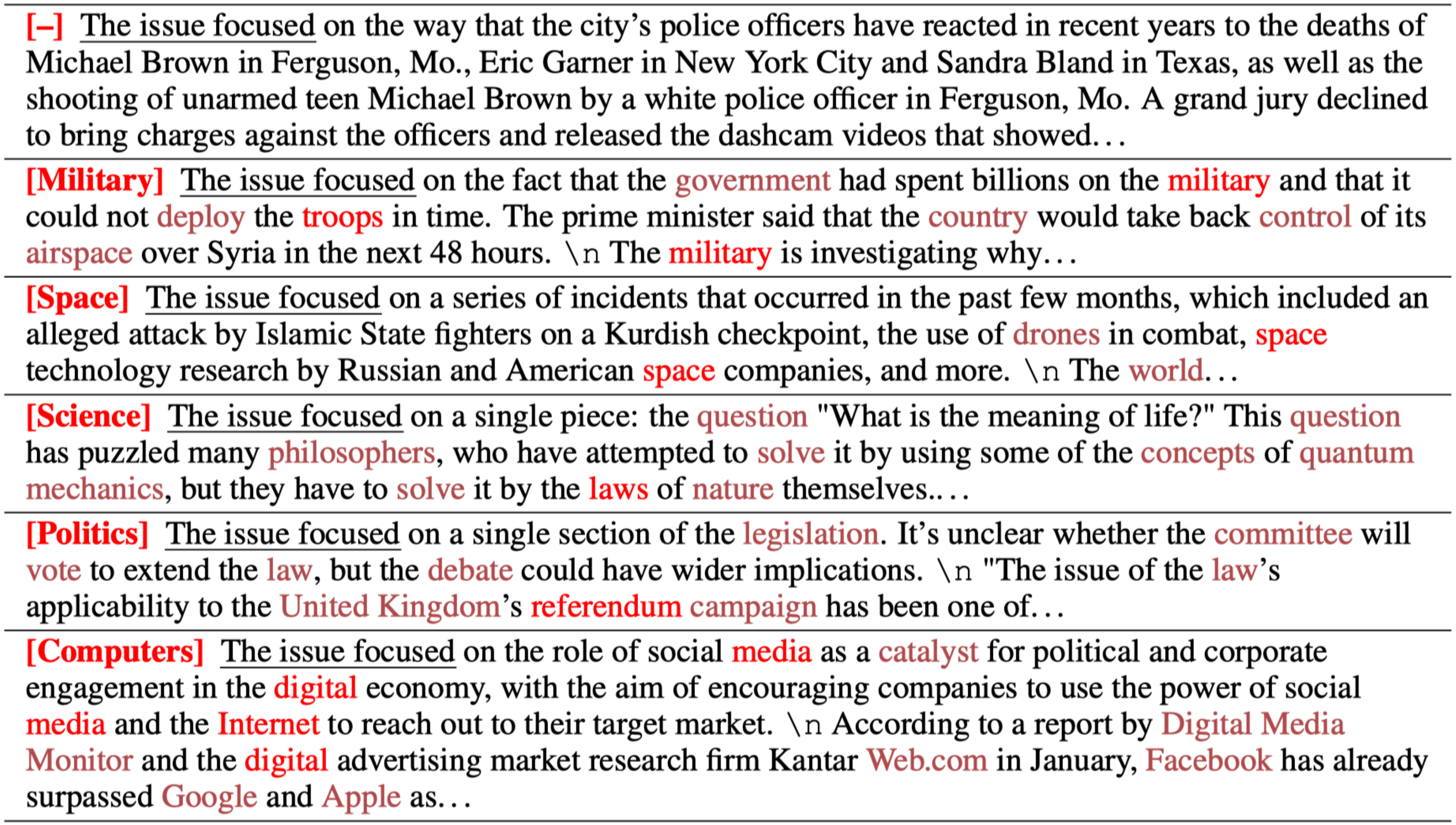
- **Kullback–Leibler (KL) Divergence**



- **Post-norm Geometric Mean Fusion**

$$x_{t+1} \sim 1/\beta \left(\widetilde{p}_{t+1}^{\gamma_{gm}} \, p_{t+1}^{1-\gamma_{gm}} \right)$$

Bag of Words (BoW)



[–] The issue focused on the way that the city's police officers have reacted in recent years to the deaths of Michael Brown in Ferguson, Mo., Eric Garner in New York City and Sandra Bland in Texas, as well as the shooting of unarmed teen Michael Brown by a white police officer in Ferguson, Mo. A grand jury declined to bring charges against the officers and released the dashcam videos that showed…

[Military] The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. \n The military is investigating why…

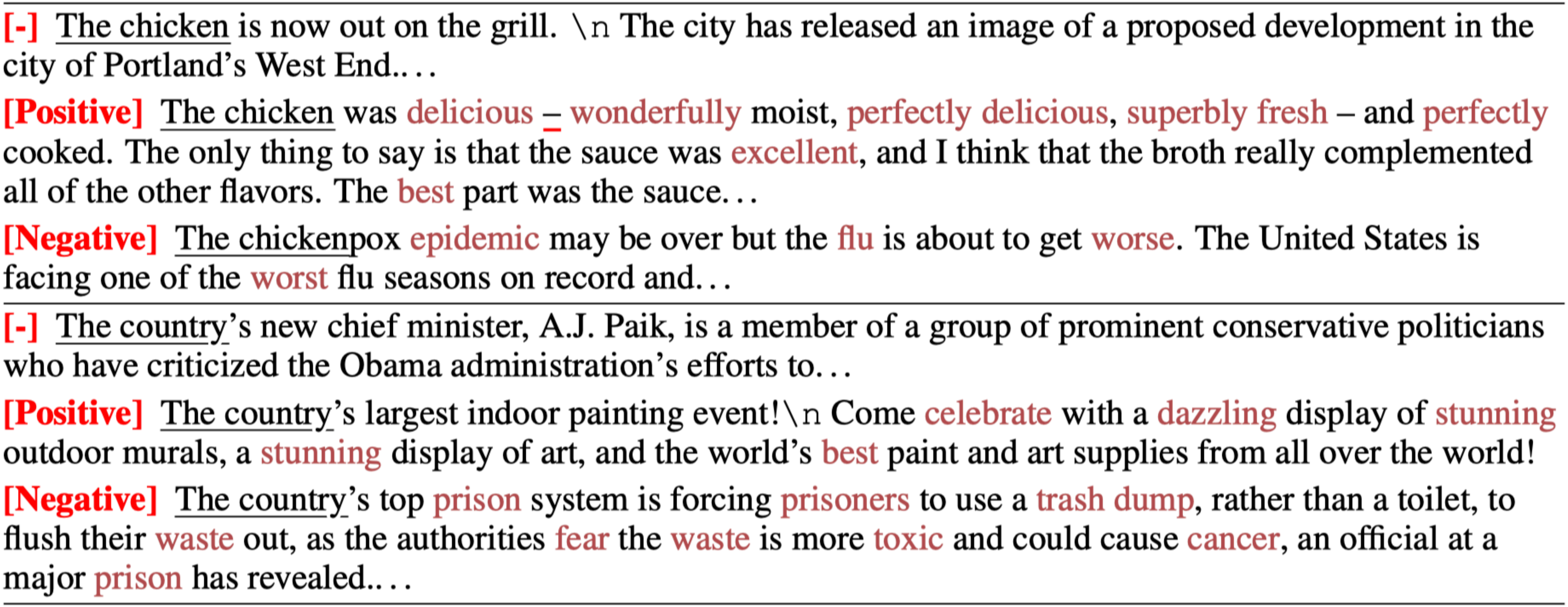
[Space] The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, space technology research by Russian and American space companies, and more. \n The world…

[Science] The issue focused on a single piece: the question "What is the meaning of life?" This question has puzzled many philosophers, who have attempted to solve it by using some of the concepts of quantum mechanics, but they have to solve it by the laws of nature themselves.…

[Politics] The issue focused on a single section of the legislation. It's unclear whether the committee will vote to extend the law, but the debate could have wider implications. \n "The issue of the law's applicability to the United Kingdom's referendum campaign has been one of…

[Computers] The issue focused on the role of social media as a catalyst for political and corporate engagement in the digital economy, with the aim of encouraging companies to use the power of social media and the Internet to reach out to their target market. \n According to a report by Digital Media Monitor and the digital advertising market research firm Kantar Web.com in January, Facebook has already surpassed Google and Apple as…

Discriminator



[-] The chicken is now out on the grill. \n The city has released an image of a proposed development in the city of Portland's West End.. . .

[Positive] The chicken was delicious – wonderfully moist, perfectly delicious, superbly fresh – and perfectly cooked. The only thing to say is that the sauce was excellent, and I think that the broth really complemented all of the other flavors. The best part was the sauce. . .

[Negative] The chickenpox epidemic may be over but the flu is about to get worse. The United States is facing one of the worst flu seasons on record and. . .

[-] The country's new chief minister, A.J. Paik, is a member of a group of prominent conservative politicians who have criticized the Obama administration's efforts to. . .

[Positive] The country's largest indoor painting event!\n Come celebrate with a dazzling display of stunning outdoor murals, a stunning display of art, and the world's best paint and art supplies from all over the world!

[Negative] The country's top prison system is forcing prisoners to use a trash dump, rather than a toilet, to flush their waste out, as the authorities fear the waste is more toxic and could cause cancer, an official at a major prison has revealed.. . .

Thanks!