Institute of Automation
Chinese Academy of Sciences

Bag of Tricks for Training Data Extraction from Language Models

ICML
40 Years
International Conference On Machine Learning

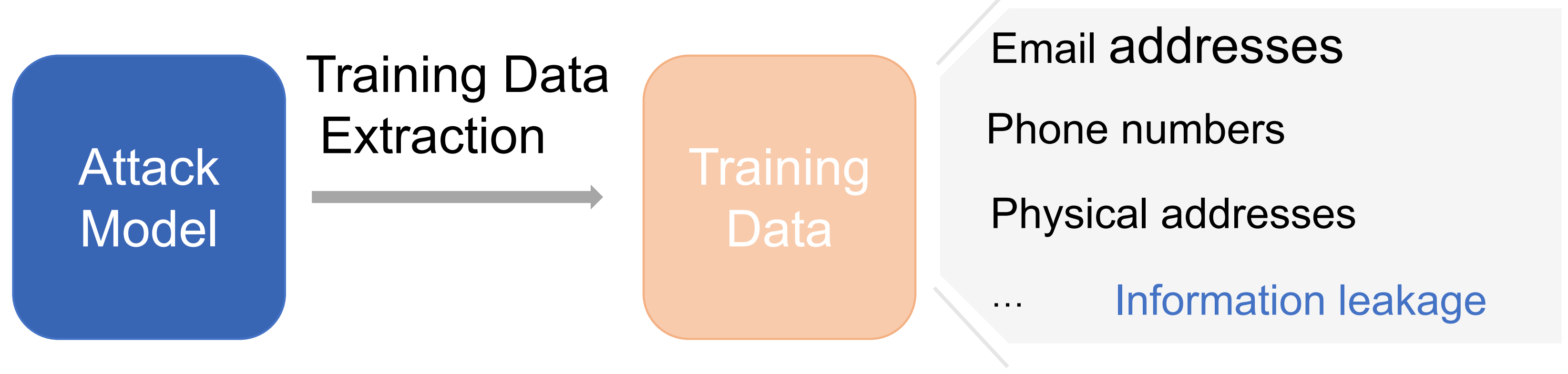
sea AI Lab

Weichen Yu[1,2], Tianyu Pang[2], Qian Liu[2], Chao Du[2], Bingyi Kang[2], Yan Huang[1], Min Lin[2], Shuicheng Yan[2]

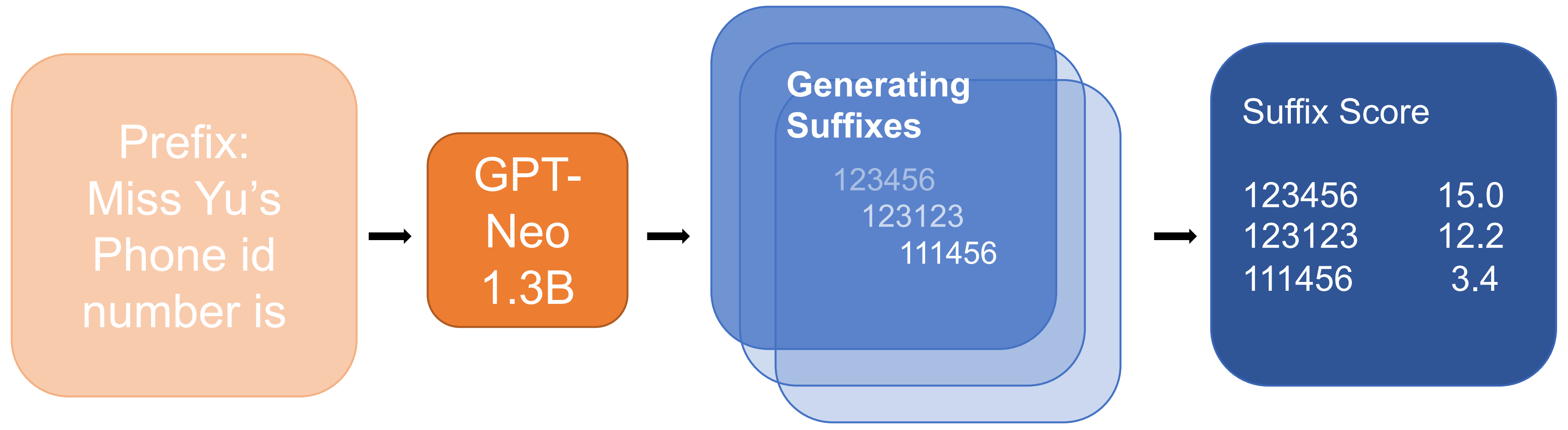
[1] Institute of Automation, Chinese Academy of Sciences   [2] Sea AI Lab

Significance of Training Data Extraction:

An effective tool to evaluate the privacy preserving ability of language models.



Pipeline: Generating-then-ranking



123456	15.0
123123	12.2
111456	3.4

Ranking by perplexity $\mathcal{P} = \exp\left(-\frac{1}{N}\sum_{n=0}^{N}\log f_\theta(x_n|x_{[0:n-1]})\right)$

1-eidetic (Carlini et al. (2021)): the sentence [*p*, *s*] appears in at most *1* example in the training data.

Evaluation Metrics:

Precision, Recall, and Hamming distance.

Bag of tricks

Probability adjustment

- temperature
- repetition penalty

Table 3. Results of $\mathcal{M}_\mathrm{P}$, $\mathcal{M}_\mathrm{R}$, and $\mathcal{M}_\mathrm{H}$ under different repetition penalty. Repetition penalty $r = 1$ is the baseline. All results are reported on 5 trials.

Repetition penalty	$\mathcal{M}_\mathrm{P}$ (%)(↑)	$\mathcal{M}_\mathrm{R}$ (%)(↑)	$\mathcal{M}_\mathrm{H}$ (↓)
0.9	19.8	66.4	27.927
1	37.0	76.5	19.614
1.1	37.3	76.5	20.181
1.2	37.1	76.5	20.323
1.3	36.7	76.4	20.332
1.5	34.7	75.7	21.154

Dynamic context window

$$f_\theta\left(x_n;\mathcal{W}\right) = h_\mathcal{W}\left(f_\theta\left(x_n|x_{[n-w_1,n-1]}\right), ..., f_\theta\left(x_n|x_{[n-w_m,n-1]}\right)\right),$$

$$f_\theta\left(x_n;\mathcal{W}_w\right) = \frac{\sum_{i=1}^{m}\epsilon_i f_\theta\left(x_n|x_{[n-w_i,n-1]}\right)}{\sum_{i=1}^{m}\epsilon_i},$$

$$f_\theta(x_n;\mathcal{W}_v) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{V}(f_\theta(x_n|x_{[n-w_i,n-1]});$$

Dynamic position shifting

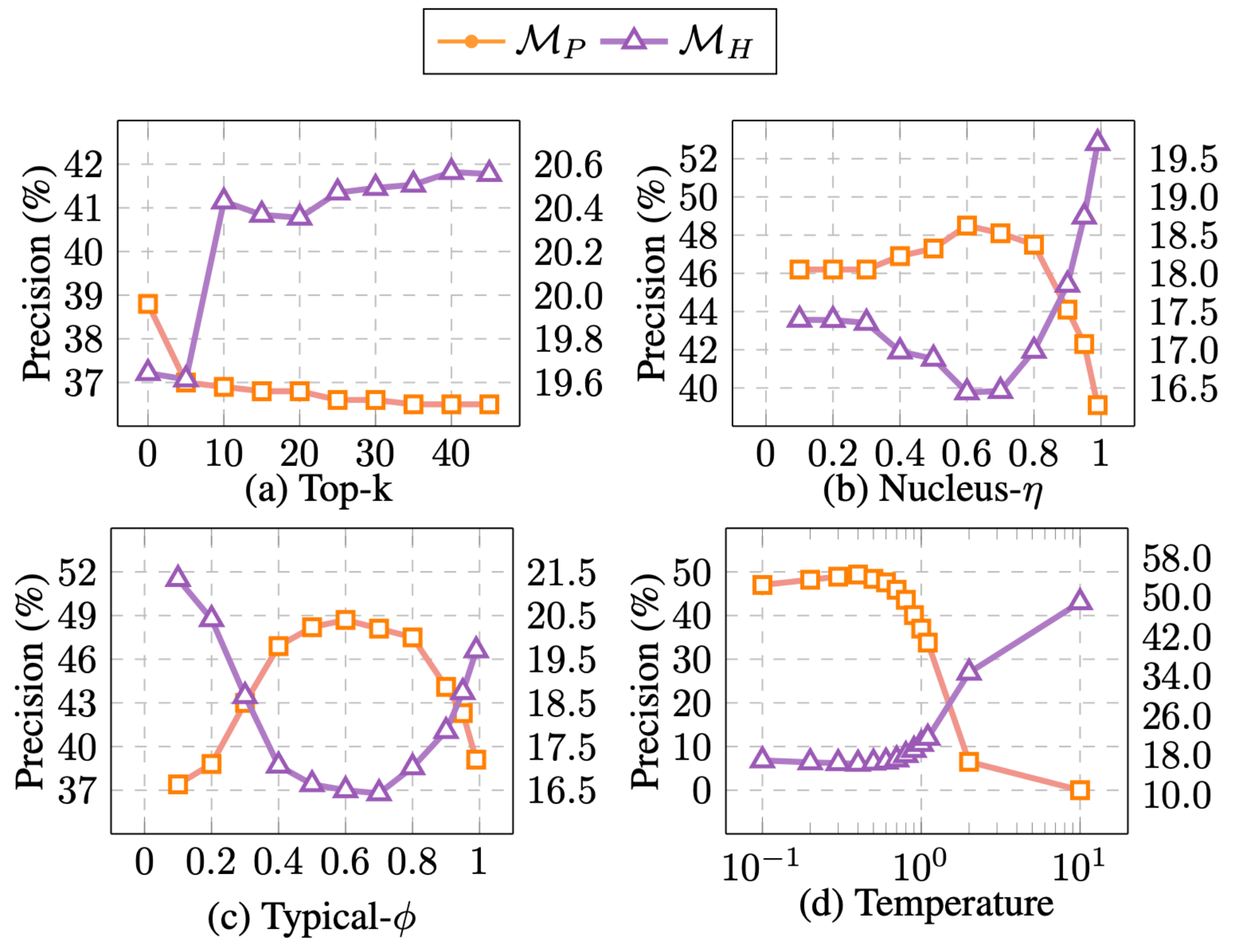
$$c = \operatorname{argmin}\mathcal{P}(p, c^i); \quad \hat{\phi}(x_i) = \psi(c_n) + \phi(x_n),$$

Table 4. Results of $\mathcal{M}_\mathrm{P}$, $\mathcal{M}_\mathrm{R}$, and $\mathcal{M}_\mathrm{H}$ under context window length adjustments. All results are reported on a single trial.

	$\mathcal{M}_\mathrm{P}$ (%)(↑)	$\mathcal{M}_\mathrm{R}$ (%)(↑)	$\mathcal{M}_\mathrm{H}$ (↓)
Baseline	19.5	65.6	26.948
Context Win $\mathcal{W}_w$	47.4	77.6	16.993
Context Win $\mathcal{W}_v$	46.7	77.5	17.164
Position Shifting	16.4	39.0	21.154

Sampling strategy

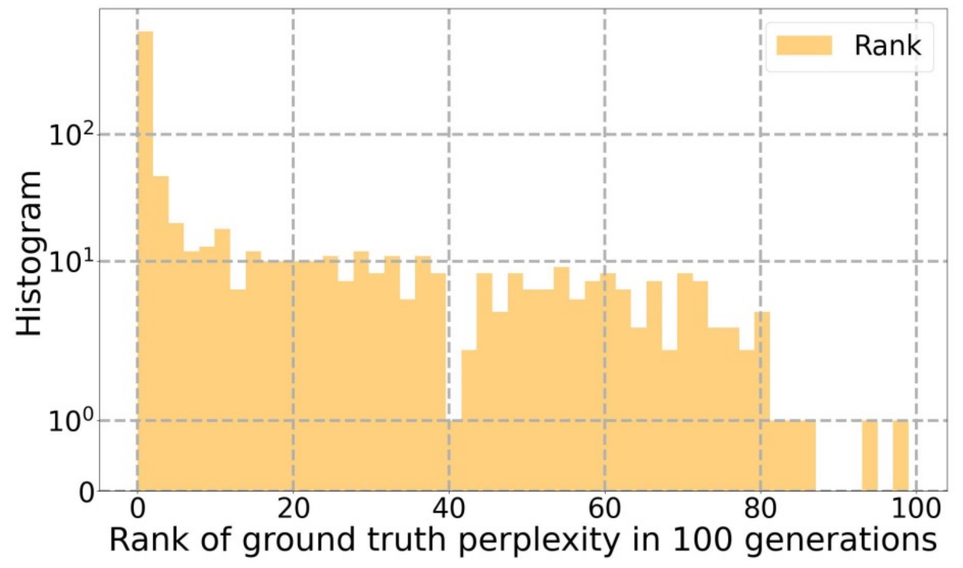
- top-k sampling
- nucleus sampling
- typical sampling



(a) Top-k (b) Nucleus-η (c) Typical-ϕ (d) Temperature

Cumprod

$$\mathcal{L}_c = (\prod_{n=0}^{N}\log p(x_n|x_0, .., x_{n-1}))^{-N}$$

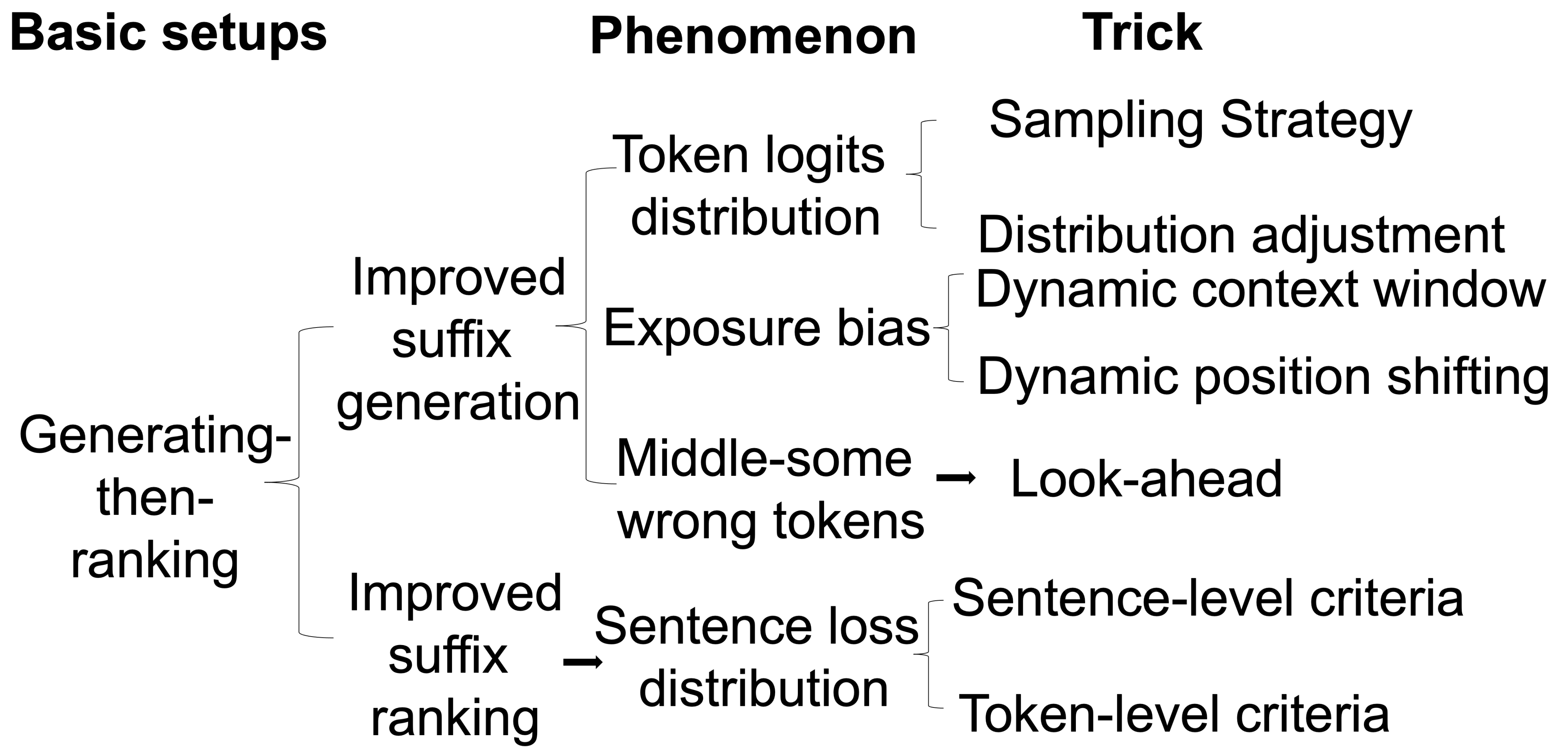


Look-ahead

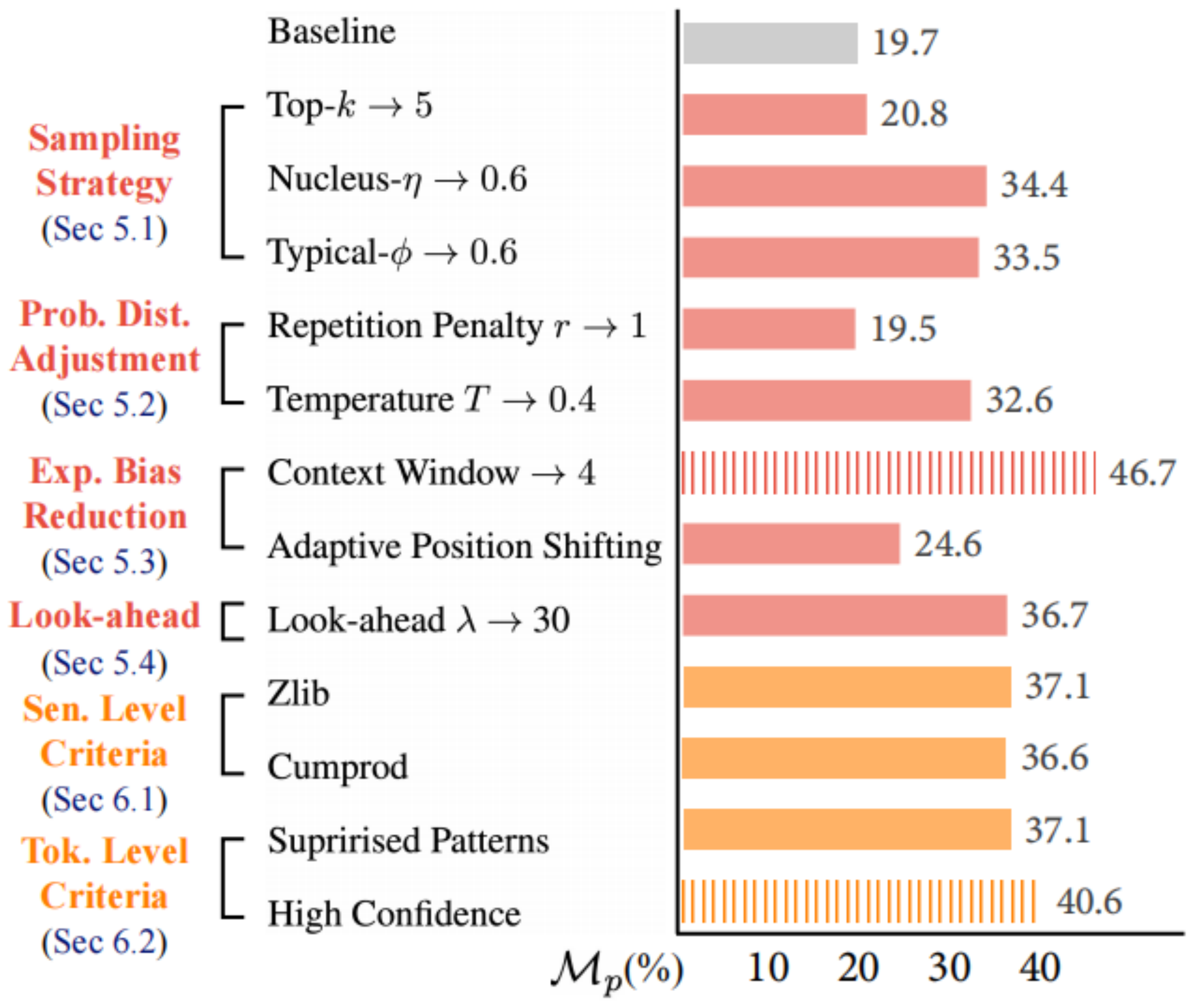
$$f_\theta\left(x_n|x_{n+1}, x_{<n}\right) = \frac{f_\theta\left(x_{n+1}|x_n, x_{<n}\right) f_\theta\left(x_n|x_{<n}\right)}{\sum_{x'_n} f_\theta\left(x_{n+1}|x'_n, x_{<n}\right) f_\theta\left(x'_n|x_{<n}\right)},$$

$$\mathcal{X} = \{x'_n|\mathcal{R}(f_\theta\left(x'_n|x_{<n}\right)) \geq \lambda\}.$$

Taxonomy of the evaluated tricks


Basic setups
Phenomenon
Trick
Generating-then-ranking
Improved suffix generation
Improved suffix ranking
Token logits distribution
Exposure bias
Middle-some wrong tokens
Sentence loss distribution
Sampling Strategy
Distribution adjustment
Dynamic context window
Dynamic position shifting
Look-ahead
Sentence-level criteria
Token-level criteria

Overall evaluation



The code is available at https://github.com/weichen-yu/LM-Extraction, primary contact Weichen at weichen.yu@cripac.ia.ac.cn.