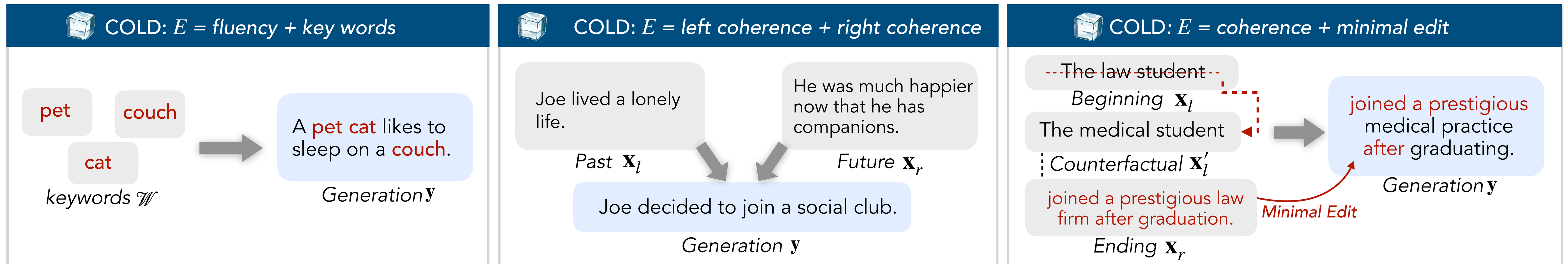
COLD Decoding:
Energy-based Constrained Text Generation with Langevin Dynamics

Lianhui Qin · Sean Welleck · Daniel Khashabi · Yejin Choi

UNIVERSITY of WASHINGTON · PAUL G. ALLEN SCHOOL · AI2

Text generation requires producing text that is not only **fluent**, but also satisfies different **constraints** that control the semantics or style of the generated text.



The dominant approach: fine-tune a pretrained LM with task-specific data

- prohibitively expensive
- can hardly scale to the infinite possible combinations of constraints

This work: constrained generation as sampling from an **energy-based model (EBM):**

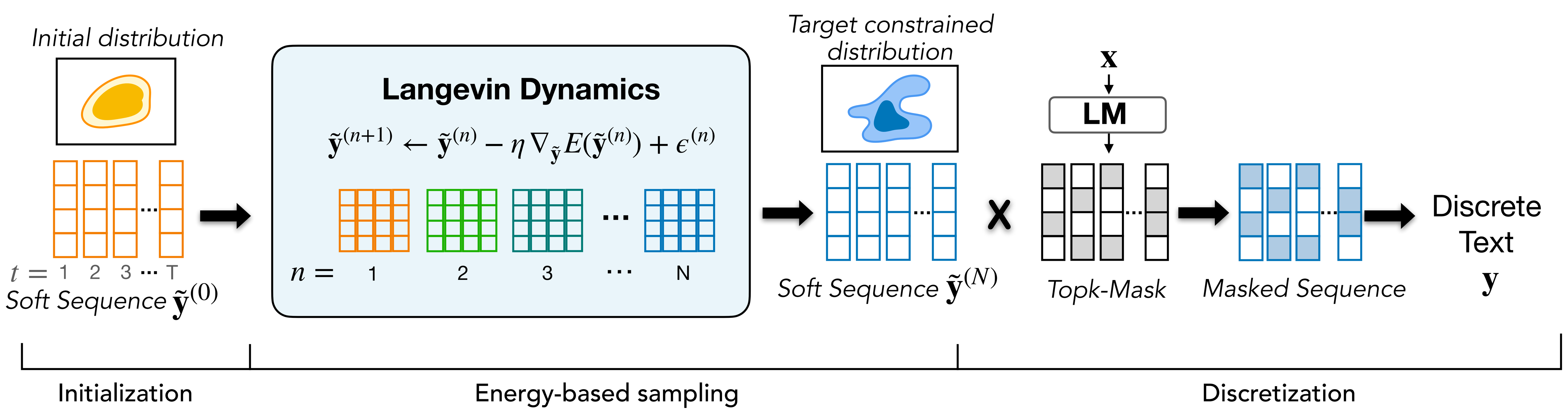
- specify an energy function by plugging in any desired constraint functions
- then sample from the induced energy-based distribution
- No training/finetuning — control on the fly!

$$p(\mathbf{y}) = \exp\left\{\sum_i \lambda_i f_i(\mathbf{y})\right\}/Z,$$

Key **challenge** of **sampling from the text EBM:**

- the normalizing factor Z is intractable
- the common discrete MCMC methods (e.g., Gibbs sampling) is too inefficient!

Solution: Use **gradient-based** MCMC, **Langevin dynamics**, for efficient sampling!



- continuous relaxation of discrete text: each token y_t is modeled with its logit vector $\tilde{\mathbf{y}}_t$
- Langevin dynamics: $\tilde{\mathbf{y}}^{(n+1)} \leftarrow \tilde{\mathbf{y}}^{(n)} - \eta \nabla_{\tilde{\mathbf{y}}} E(\tilde{\mathbf{y}}^{(n)}) + \epsilon^{(n)}$
- Discretize the sampled continuous text vector with *top-k filtering* (see paper for more details)

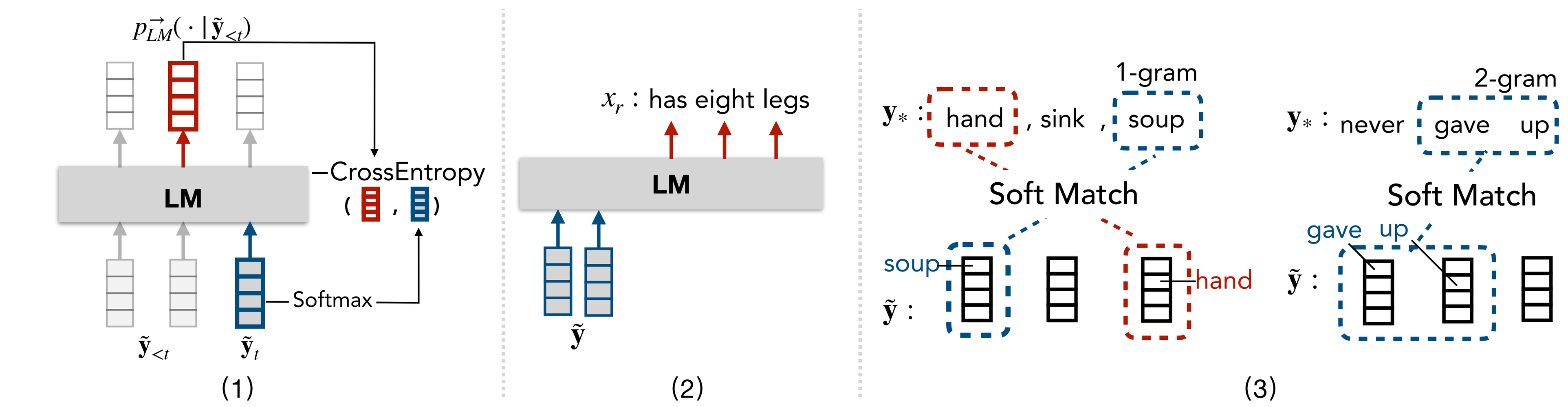
Algorithm of COLD

Algorithm 1 Constrained Decoding w/ Langevin Dynamics.

input Constraints $\{f_i\}$, length T, iterations N.
output Sample sequence $\mathbf{y}$.

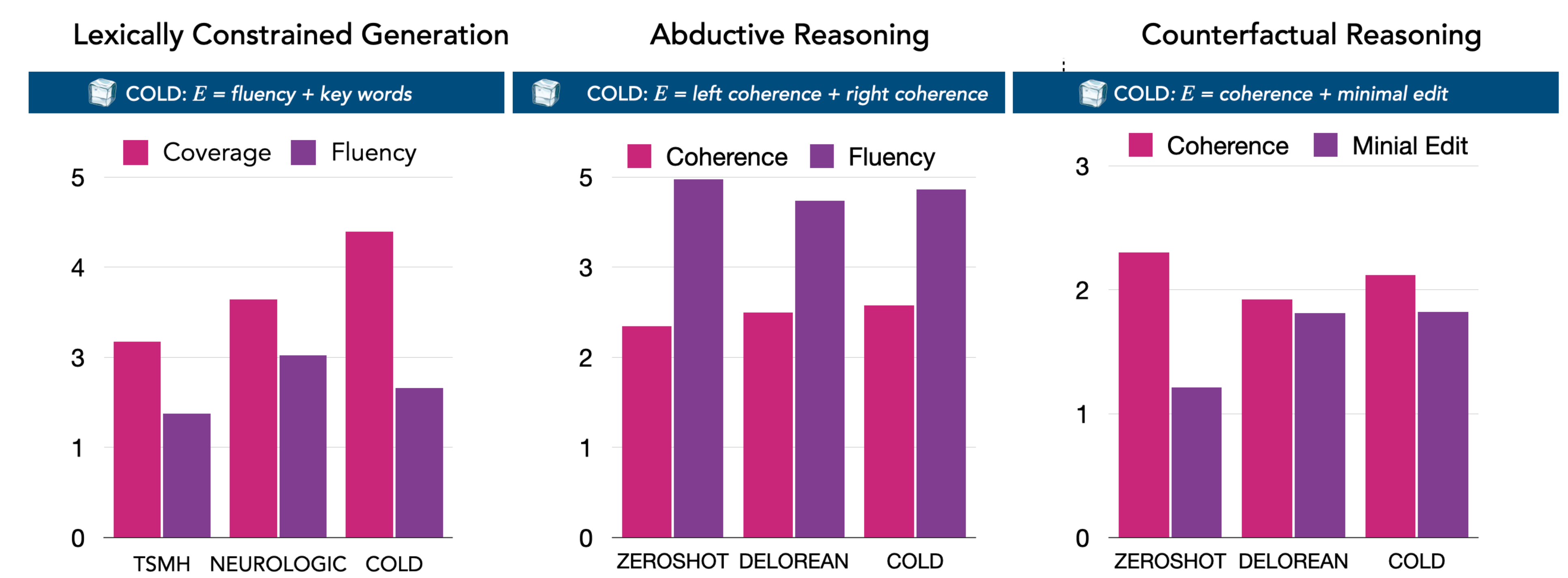
```
ỹ_t^(0) ← init() for all position t  // init soft-tokens
for n ∈ {1, ..., N} do
    E^(n) ← E(ỹ^(n); {f_i})  // compute energy (§3.2)
    ỹ_t^(n+1) ← ỹ_t^(n) − η ∇_{ỹ_t} E^(n) + ε_t^(n) for all t  // update soft tokens (Eq.2)
end for
y_t = arg max_v topk-filter(ỹ_t^(N)(v)) for all t // discretize (Eq.6)
return: y = (y_1, ..., y_T)
```

Illustrations of example differentiable constraints



(1) fluency constraint (2) future contextualization constraint (3) n-gram similarity constraint

Experiment Results



Check out COLD decoding paper!!