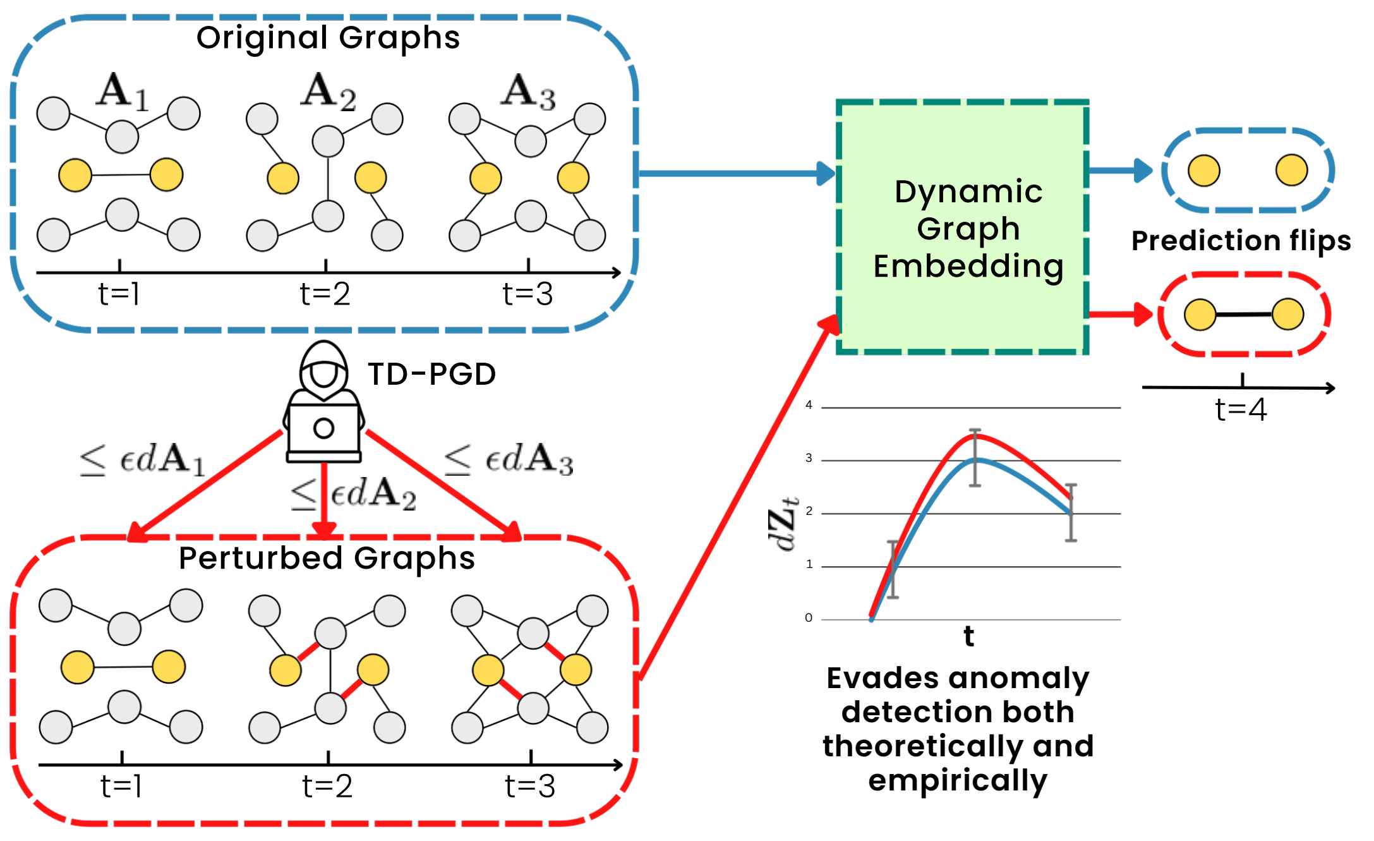
Intention in tension: Towards Robust and Controllable Machine Learning

KARTIK SHARMA
Georgia Tech
ksartik@gatech.edu

Robustness (tension)

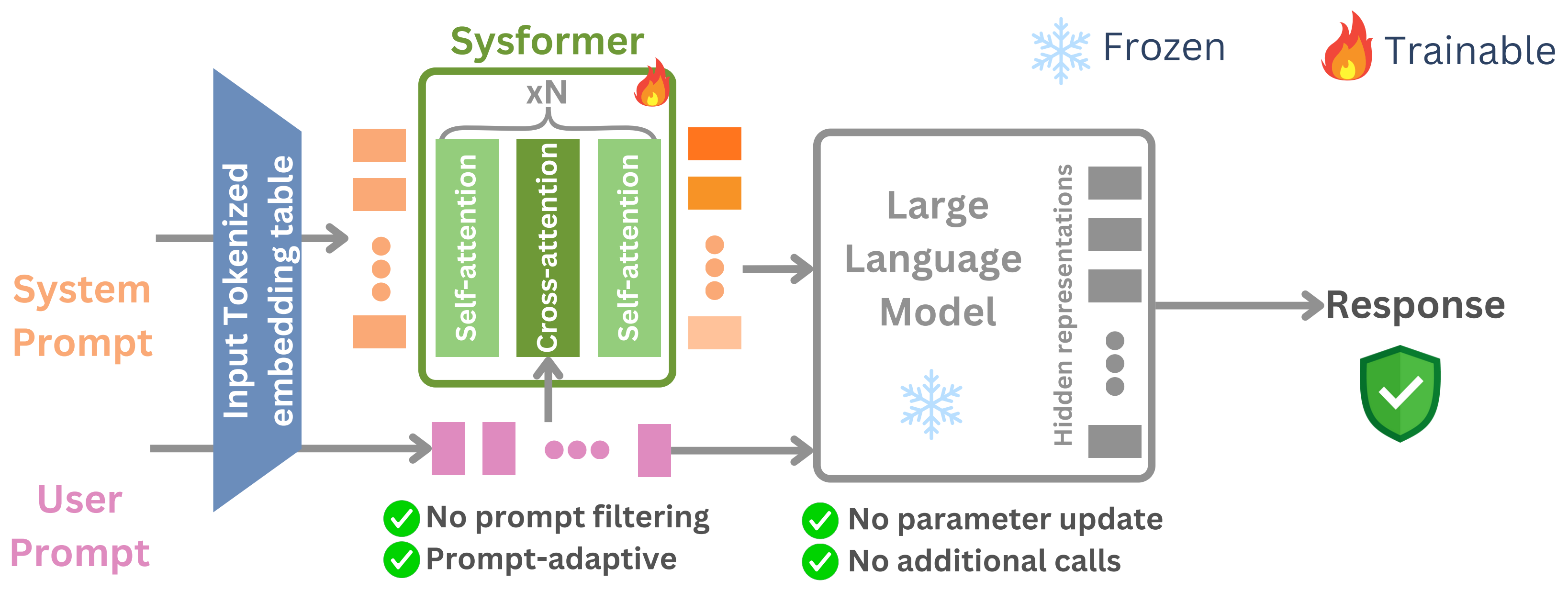
stretching the model, finding its break point

Adversarial vulnerabilities



Temporal Dynamics-Aware Adversarial Attacks on Discrete-Time Dynamic Graph Models.
Kartik Sharma, Rakshit Trivedi, Rohit Sridhar, Srijan Kumar. KDD 2023

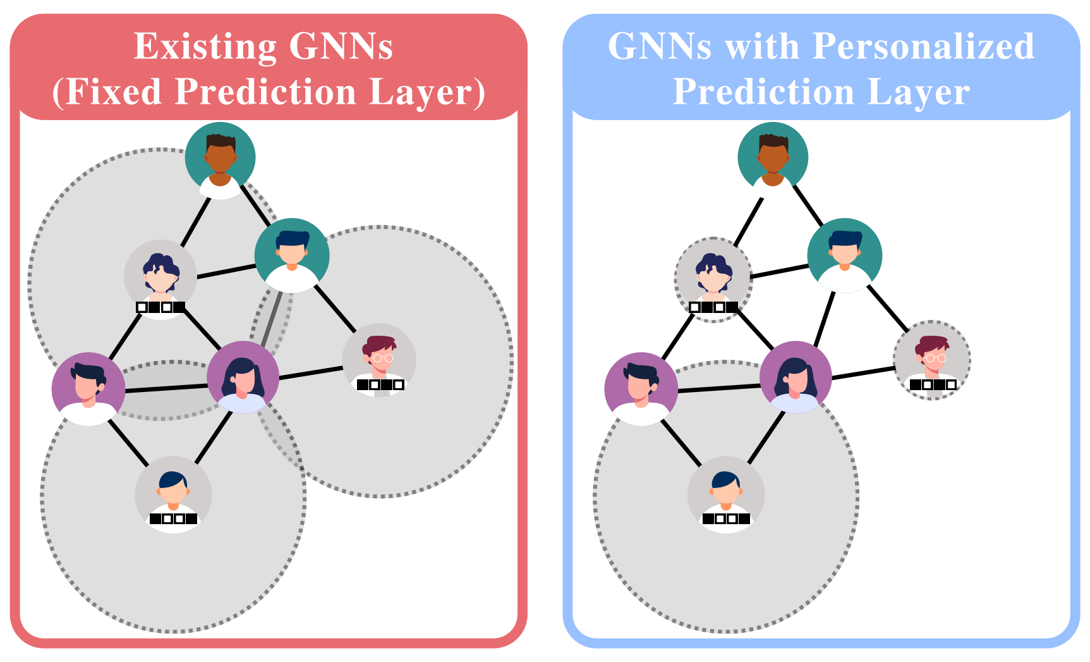
Adaptive defenses



✓ No prompt filtering
✓ Prompt-adaptive
✓ No parameter update
✓ No additional calls

Sysformer: Safeguarding Frozen Large Language Models with Adaptive System Prompts.
Kartik Sharma, Yiqiao Jin, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, Srijan Kumar. arXiv:2506.15751

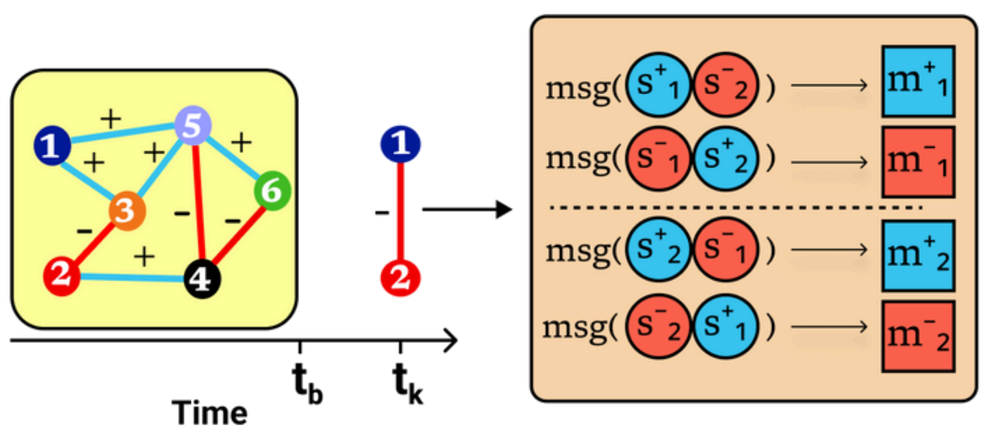
- Node-adaptive layer selection in GNNs for more robust feature aggregation in local neighborhood


Existing GNNs (Fixed Prediction Layer)
GNNs with Personalized Prediction Layer

Personalized Layer Selection for Graph Neural Networks
Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Srijan Kumar, Mahashweta Das. TMLR 2025

Invariance and Equivariance

Learning sign-equivariant representations in dynamic networks through social balance theory

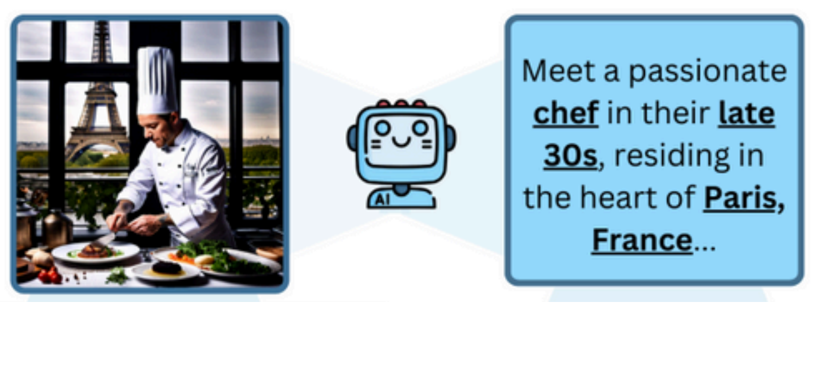

1
2
3
4
5
6
+
-
Time
t_b
t_k
$\text{msg}(s^+_1, s^-_2) \rightarrow m^+_1$
$\text{msg}(s^-_1, s^+_2) \rightarrow m^-_1$
$\text{msg}(s^+_2, s^-_1) \rightarrow m^+_2$
$\text{msg}(s^-_2, s^+_1) \rightarrow m^-_2$

Representation Learning in Continuous-Time Dynamic Signed Networks. Kartik Sharma*, Mohit Raghavendra*, Yeon Chang Lee, Anand Kumar M, Srijan Kumar. CIKM 2023

Evaluating modality invariance in LLM personas

Persona “A **35**-year-old **male chef** from **Paris**”

Scenario "You're coordinating a playlist for your high school reunion after-party. The organizers want **music** specifically from **your graduating years** to recreate the atmosphere. You ..."



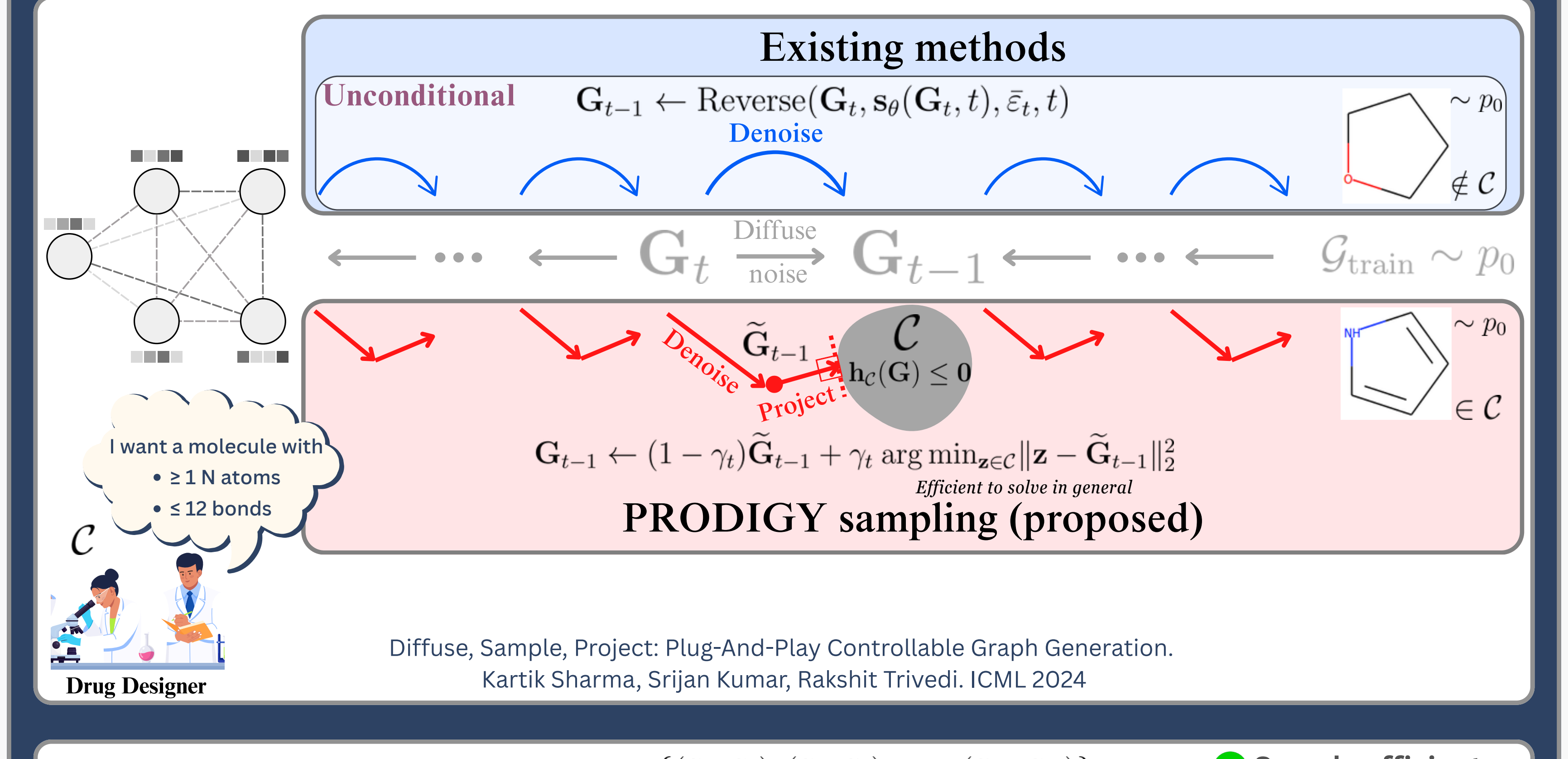
Meet a passionate **chef** in their **late 30s**, residing in the heart of **Paris, France**...

A Thousand Words or An Image: Studying the Influence of Persona Modality in Multimodal LLMs.
Julius Broomfield ⋆, Kartik Sharma ⋆, Srijan Kumar.
arXiv:2502.20504

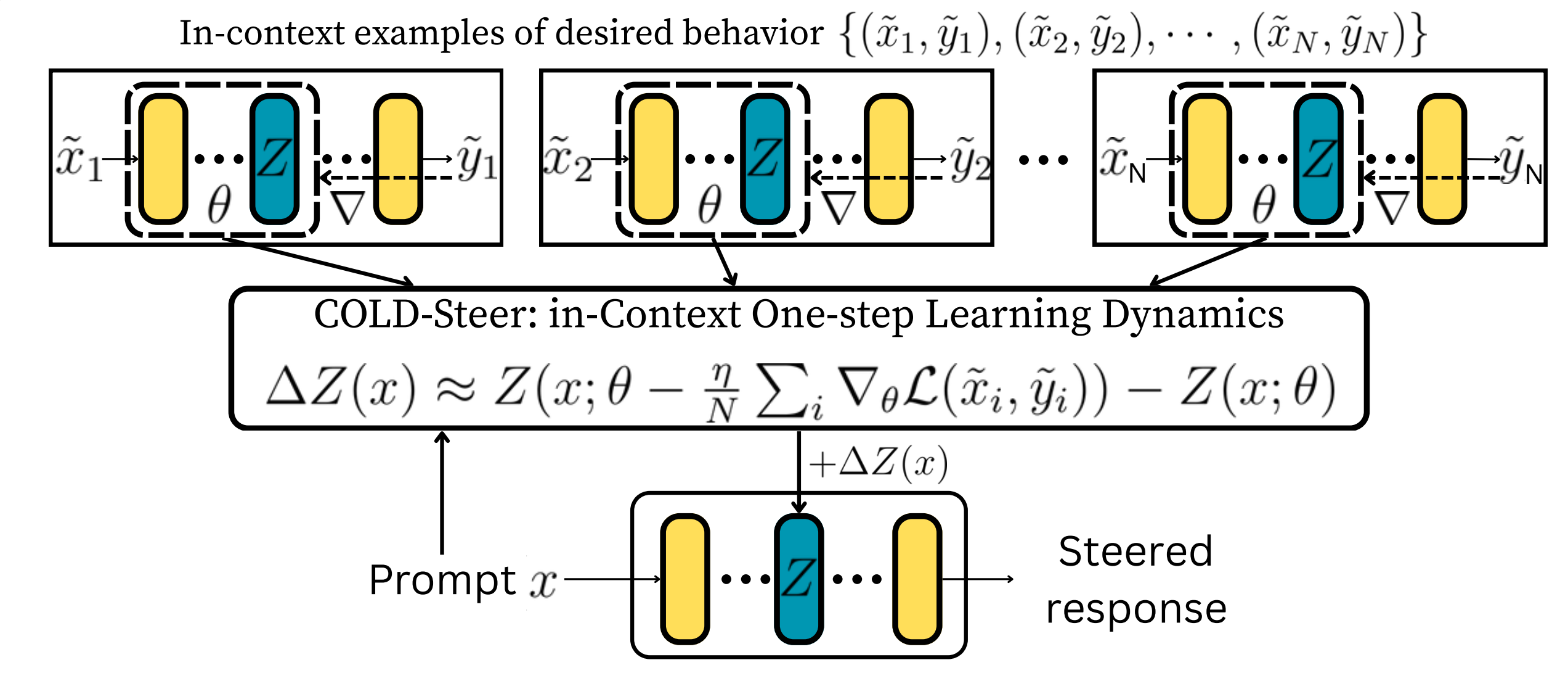
Controllable (intention)

controllable by users, adapting to user intentions

Test-time control


Existing methods
Unconditional
$\mathbf{G}_{t-1} \leftarrow \text{Reverse}(\mathbf{G}_t, \mathbf{s}_\theta(\mathbf{G}_t, t), \bar{\varepsilon}_t, t)$
Denoise
$\sim p_0$
$\notin \mathcal{C}$
$\mathbf{G}_t \xrightarrow[\text{noise}]{\text{Diffuse}} \mathbf{G}_{t-1}$
$\mathcal{G}_{\text{train}} \sim p_0$
$\widetilde{\mathbf{G}}_{t-1}$
$\mathcal{C}$
$\mathbf{h}_{\mathcal{C}}(\mathbf{G}) \leq 0$
Denoise
Project
$\sim p_0$
$\in \mathcal{C}$
$\mathbf{G}_{t-1} \leftarrow (1-\gamma_t)\widetilde{\mathbf{G}}_{t-1} + \gamma_t \arg\min_{\mathbf{z}\in\mathcal{C}} \|\mathbf{z} - \widetilde{\mathbf{G}}_{t-1}\|_2^2$
Efficient to solve in general
PRODIGY sampling (proposed)
I want a molecule with
• ≥ 1 N atoms
• ≤ 12 bonds
$\mathcal{C}$
Drug Designer

Diffuse, Sample, Project: Plug-And-Play Controllable Graph Generation.
Kartik Sharma, Srijan Kumar, Rakshit Trivedi. ICML 2024


In-context examples of desired behavior $\{(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), \cdots, (\tilde{x}_N, \tilde{y}_N)\}$
$\tilde{x}_1$
$\tilde{y}_1$
$\tilde{x}_2$
$\tilde{y}_2$
$\tilde{x}_N$
$\tilde{y}_N$
θ
Z
∇
COLD-Steer: in-Context One-step Learning Dynamics
$\Delta Z(x) \approx Z(x; \theta - \frac{\eta}{N}\sum_i \nabla_\theta \mathcal{L}(\tilde{x}_i, \tilde{y}_i)) - Z(x; \theta)$
$+\Delta Z(x)$
Prompt x
Steered response

✓ Sample efficient

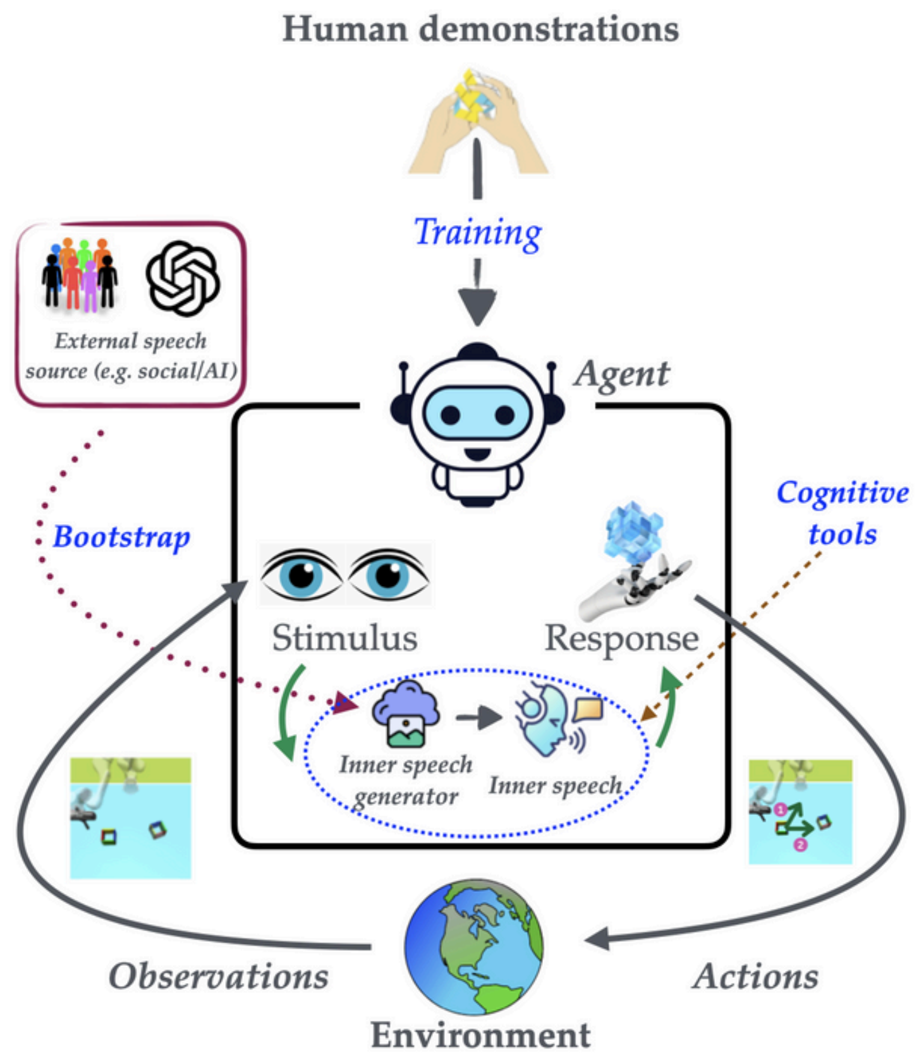
✓ Loss-driven target
✓ Gradient signal

✓ Optimization free

COLD-Steer: Steering Large Language Models via in-Context One-Step Learning Dynamics. Kartik Sharma, Rakshit Trivedi. Under Review

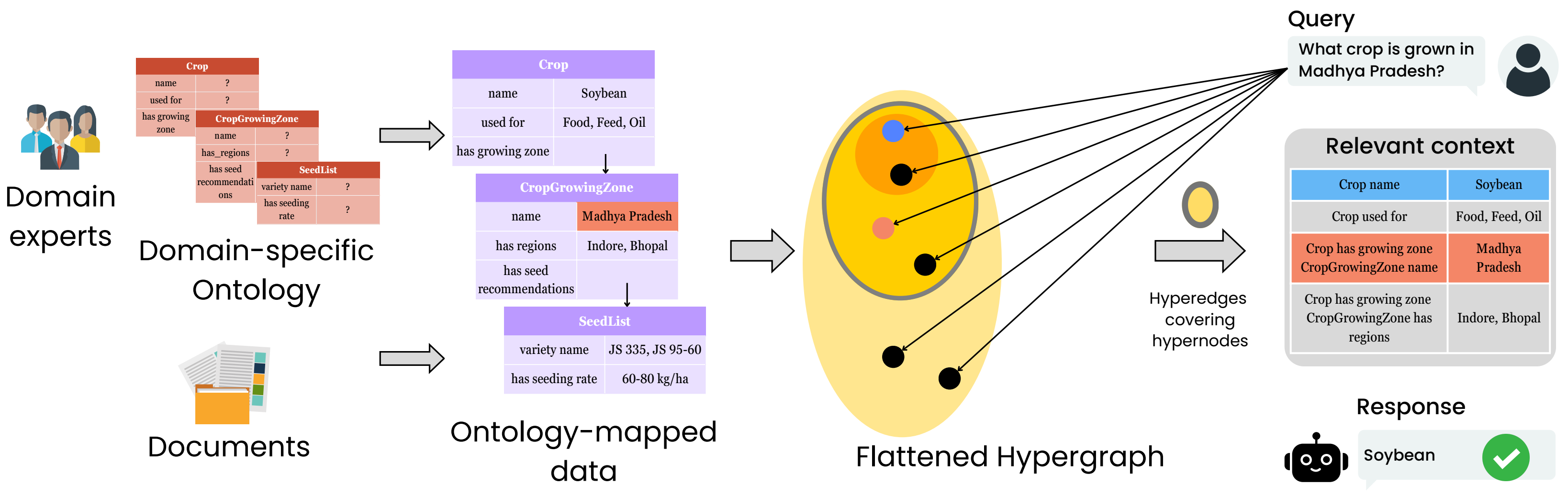
Capture intrinsic diversity

- Using VLMs to discriminate behavior in language as “inner speech”
- Conditioning behavior generation on inner speech to capture diverse demonstrations and enable steerability



Inner Speech as Behavior Guides: Steerable Imitation of Diverse Behaviors for Human-AI coordination.
Rakshit Trivedi ⋆, Kartik Sharma ⋆, David C. Parkes. NeurIPS Spotlight, 2025

Flexible domain adaptation



OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models.
Kartik Sharma, Peeyush Kumar, Yunqing Li. EMNLP Main, 2025

Future: Putting AI under *tensional intention* and *intentional tension*

- **Tensional intention:** *ambiguity within intention*
  - Understanding vague and counteracting user intentions
  - Pluralistic output from contradictory inputs
- **Intentional tension:** *deliberate tension for training*
  - Learning generalizable policies from weak feedback
  - Multi-agent systems for competition and cooperation