Towards Consistent Language Models Using Controlled Prompting and Decoding

Jasmin Mousavi, Arash Termehchy

Oregon State University

IDEA
IDEA
Foundations of
Data-**Ce**ntric
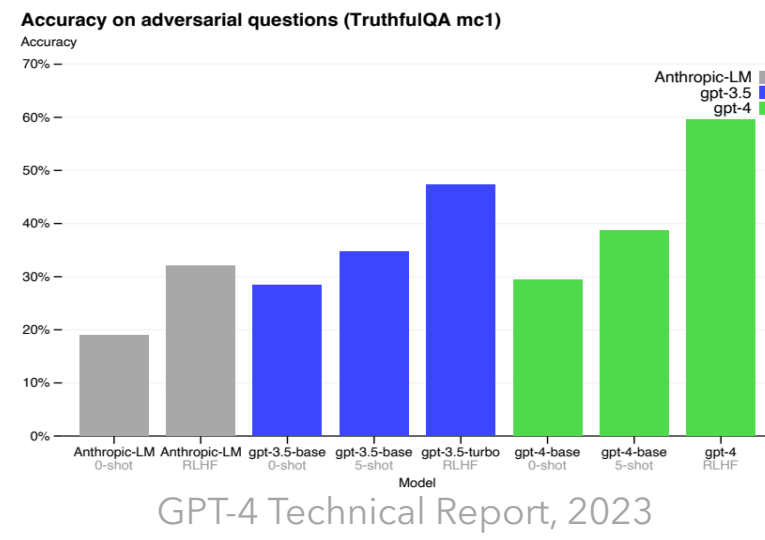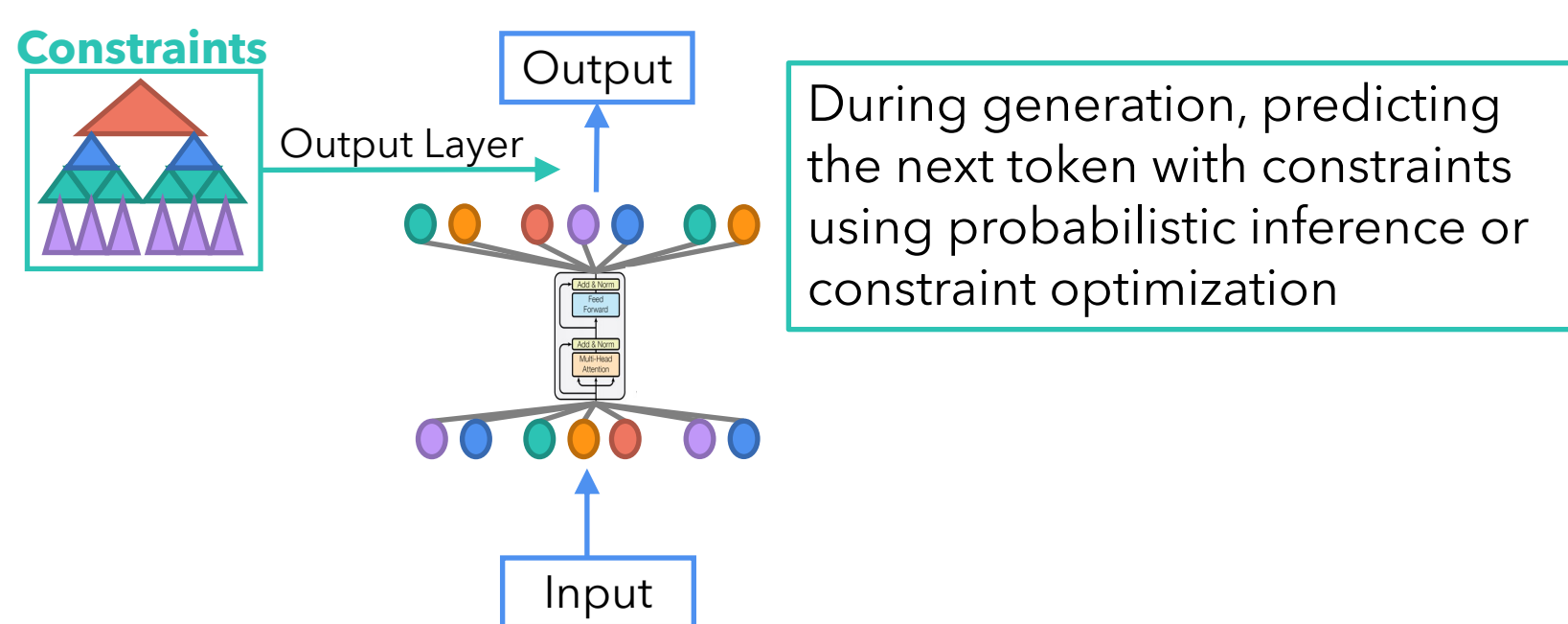Systems **La**b

1. LLMs Hallucinate

- Produces inconsistent information and inaccurate results
- LLMs *over-generalize patterns and relationships* from pretraining data
  - Causes inconsistent and inaccurate results
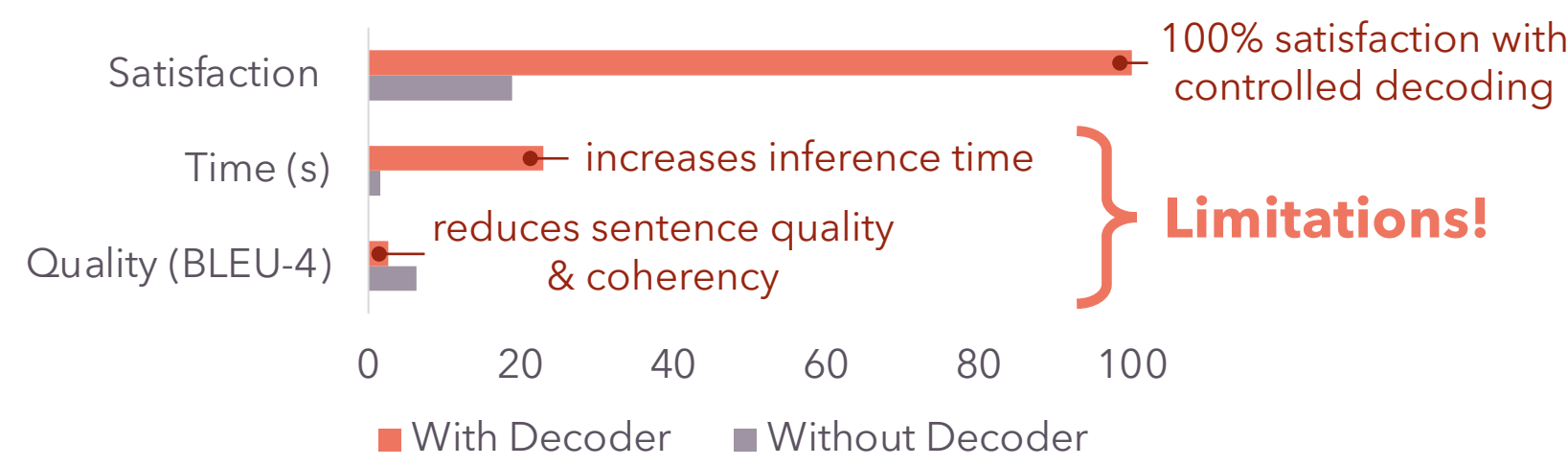
GPT-4: 40% hallucination rate

Accuracy on adversarial questions (TruthfulQA mc1)

GPT-4 Technical Report, 2023

2. Current Efforts for Reducing Inconsistencies

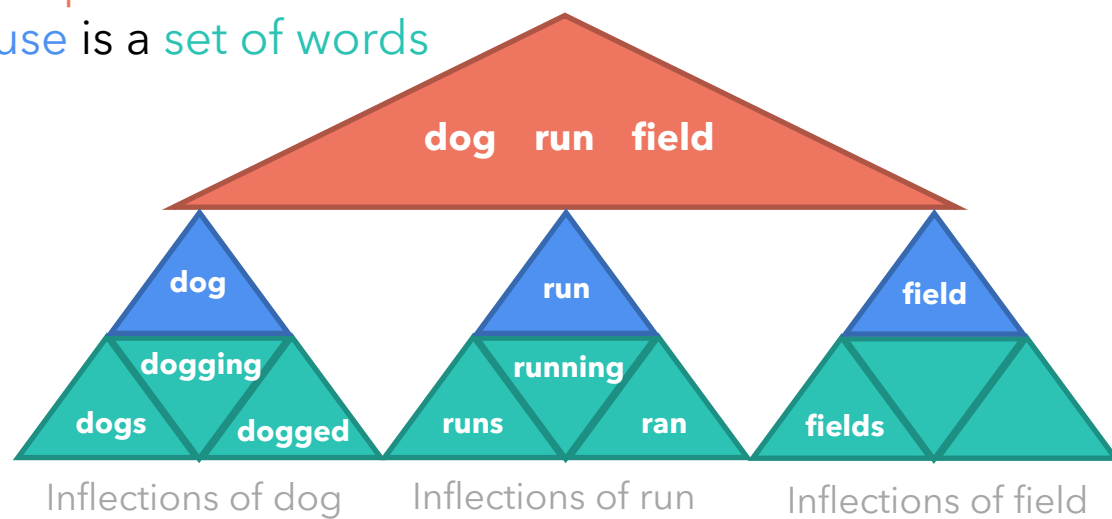- Ensuring the output of LLMs is consistent with domain constraints

During generation, predicting the next token with constraints using probabilistic inference or constraint optimization

- CommonGen constraints = contains keywords or their inflections

- 100% satisfaction with controlled decoding
- increases inference time
- reduces sentence quality & coherency

Limitations!

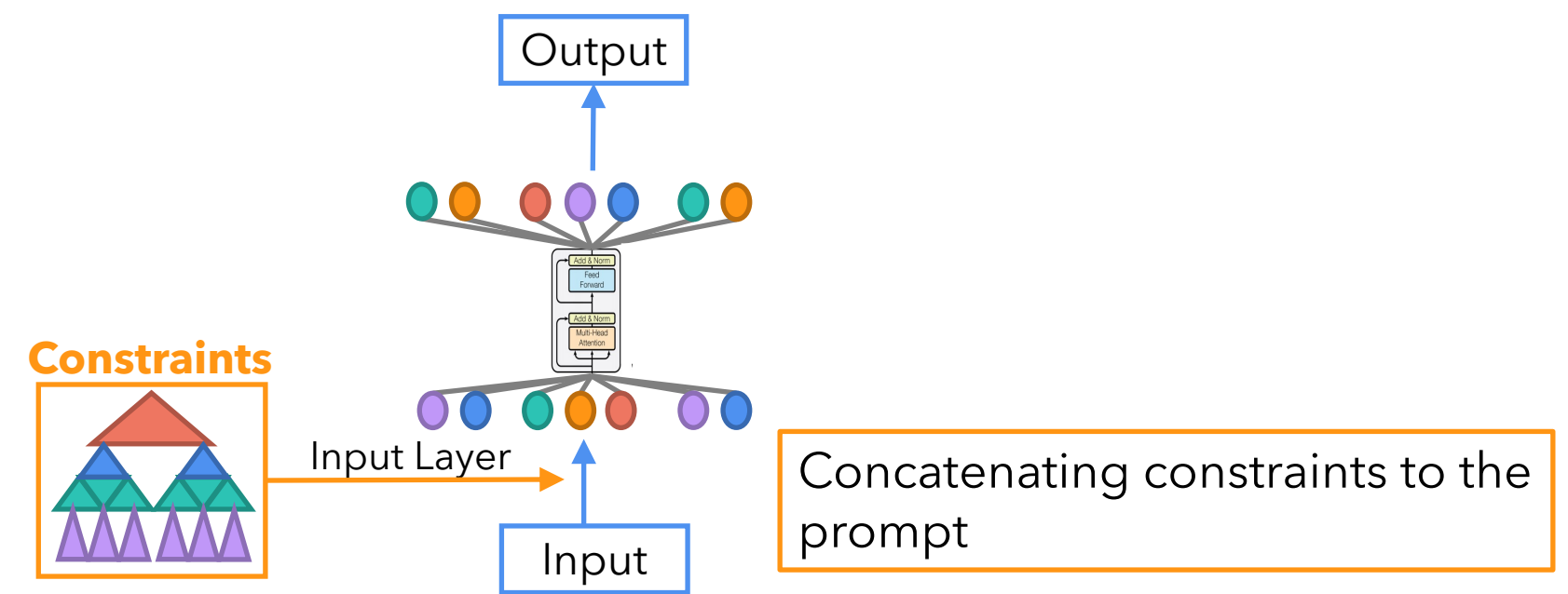Legend: With Decoder / Without Decoder

3. Constraints in CommonGen

- **Given**: Concept word set = [dog, field, run]
- **Goal**: Write a sentence using all concepts from the set

Representing Constraints in Conjunctive Normal Form (CNF)
- Each concept is a clause
- Each clause is a set of words

dog run field

dog / dogging / dogs / dogged
Inflections of dog

run / running / runs / ran
Inflections of run

field / fields
Inflections of field

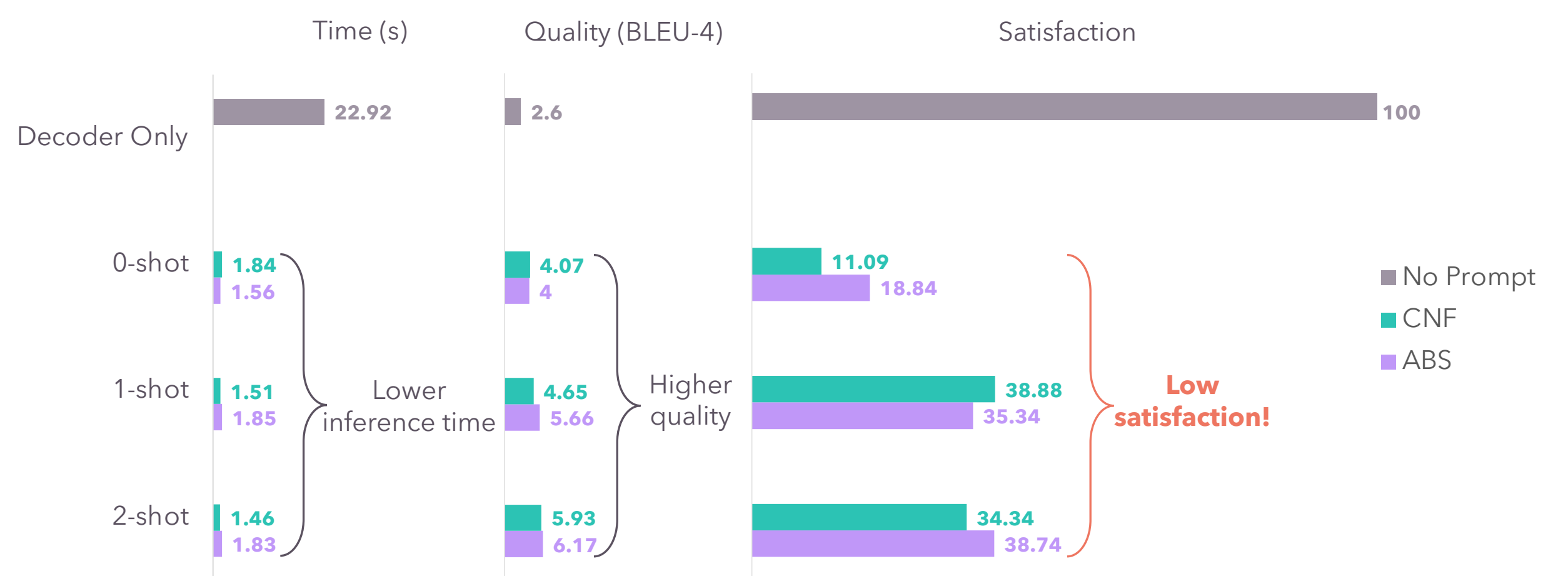(dog ∨ dogs ∨ dogging ∨ dogged) ∧ (run ∨ runs ∨ running ∨ ran) ∧ (field ∨ fields)

4. Can Prompting with Constraints Help Mitigate Decoder Limitations?

- Adding constraints to the input of LLMs
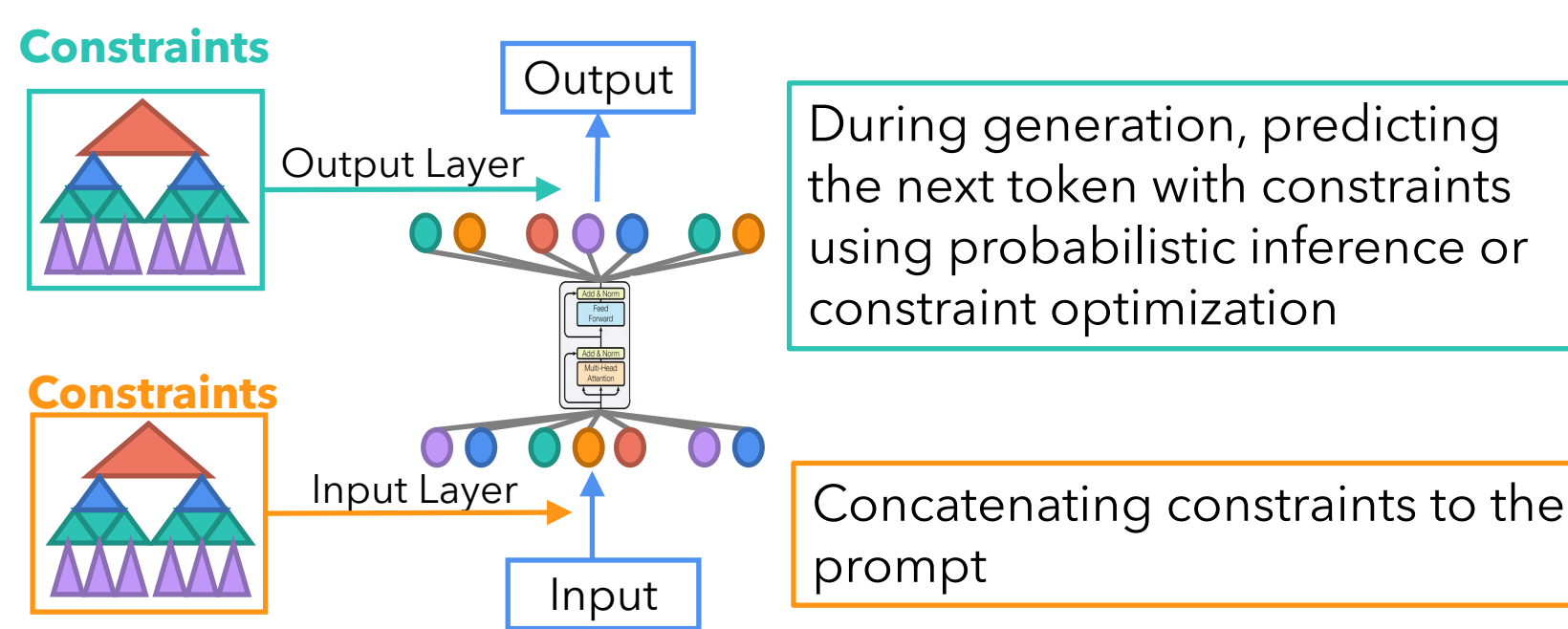

Concatenating constraints to the prompt

CNF — Conjunctive Normal Form
CNF constraint as text

Write a sentence using the words (dog or dogs or …) and (run or running or …) and (field or fields)

ABS — Abstract
Abstract instance of constraint

Given a set of words x, write a sentence using all words in x or inflections of x

	Time (s)	Quality (BLEU-4)	Satisfaction
Decoder Only	22.92	2.6	100
0-shot	1.84 / 1.56	4.07 / 4	11.09 / 18.84
1-shot	1.51 / 1.85	4.65 / 5.66	38.88 / 35.34
2-shot	1.46 / 1.83	5.93 / 6.17	34.34 / 38.74

Lower inference time / Higher quality / Low satisfaction!

Legend: No Prompt / CNF / ABS

5. Empirical Study

During generation, predicting the next token with constraints using probabilistic inference or constraint optimization

Concatenating constraints to the prompt

Metrics

1. Generation Quality (perfect score = 100.0)
   - Automated metrics: **ROUGE-L**, **BLEU-4**, **CIDEr**, **SPICE**
2. Satisfaction (perfect score = 100.0)
   - If *all* CNF clauses in constraint C are met: $\bigwedge_{i=1}^{|C|} C_i$
3. Time
   - Inference time (seconds)

Decoder Strategies

NL — NeuroLogic
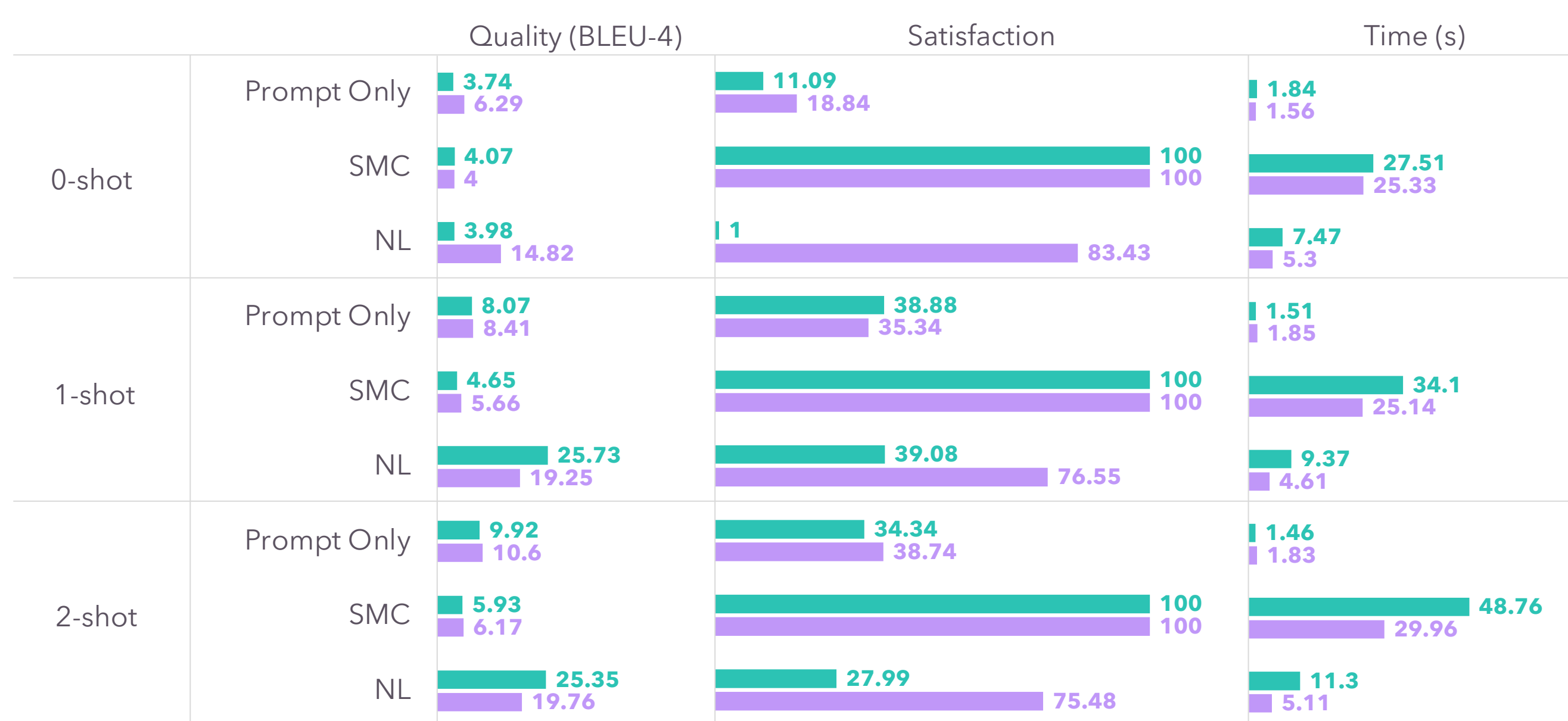- Soft constraint decoder
- Beam-based look ahead strategy

NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints , 2021
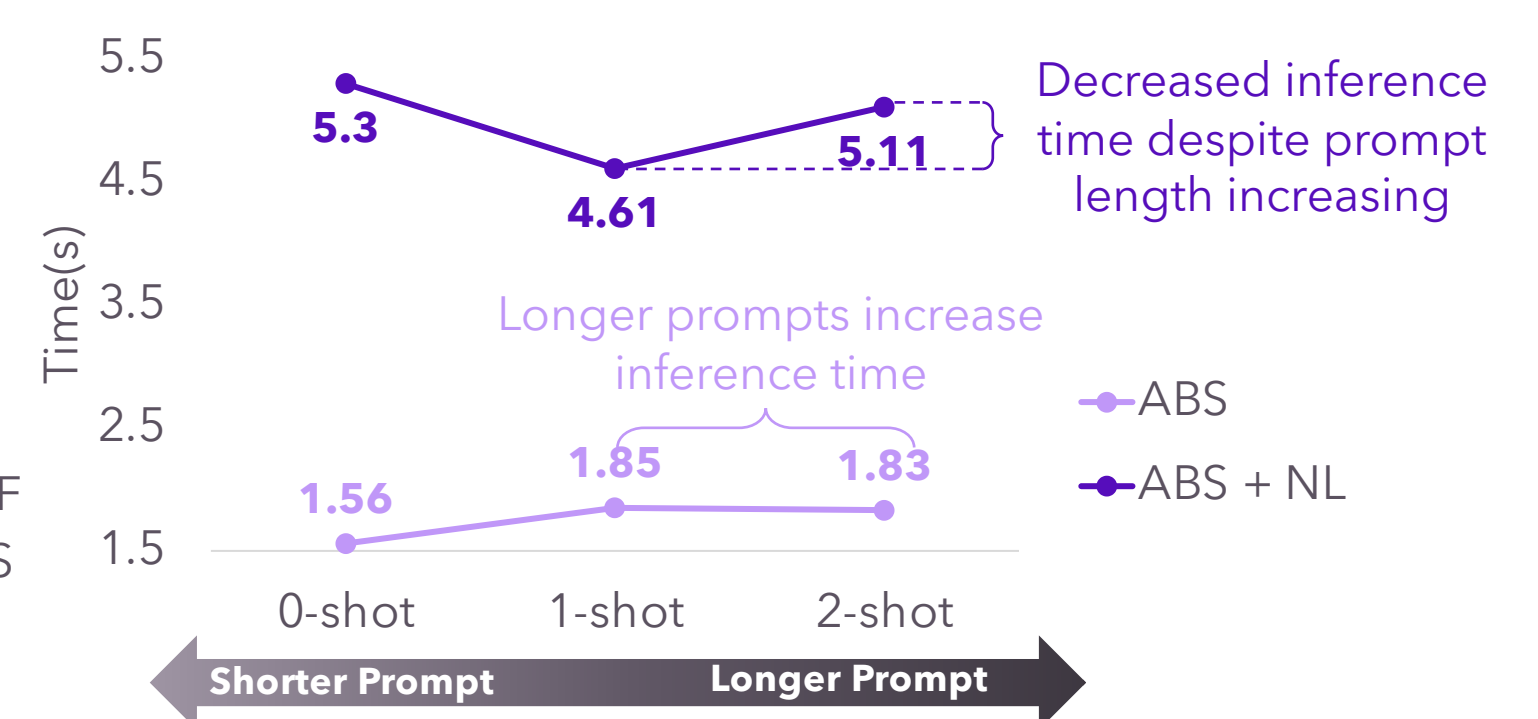
SMC — Sequential Monte Carlo
- Hard constraint decoder
- Masked vocabulary strategy

Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs , 2023
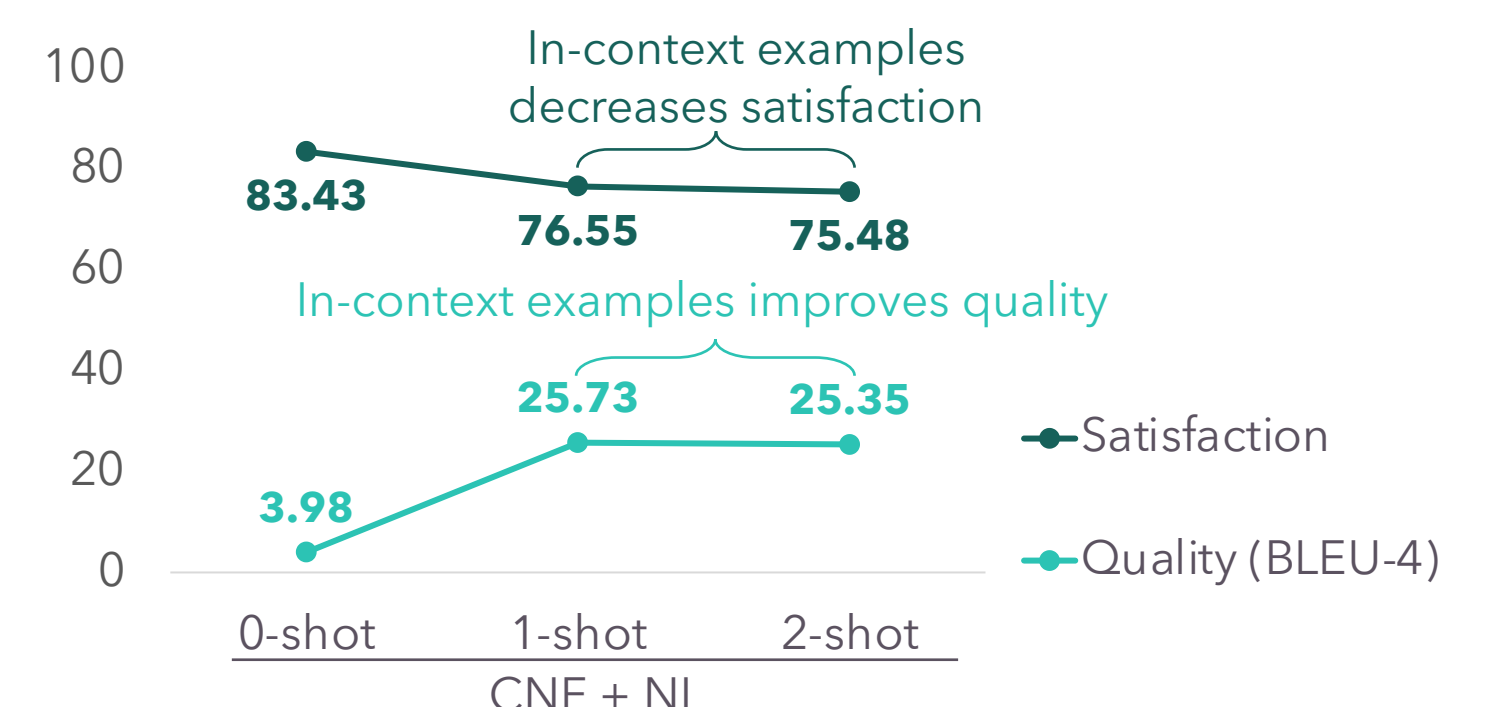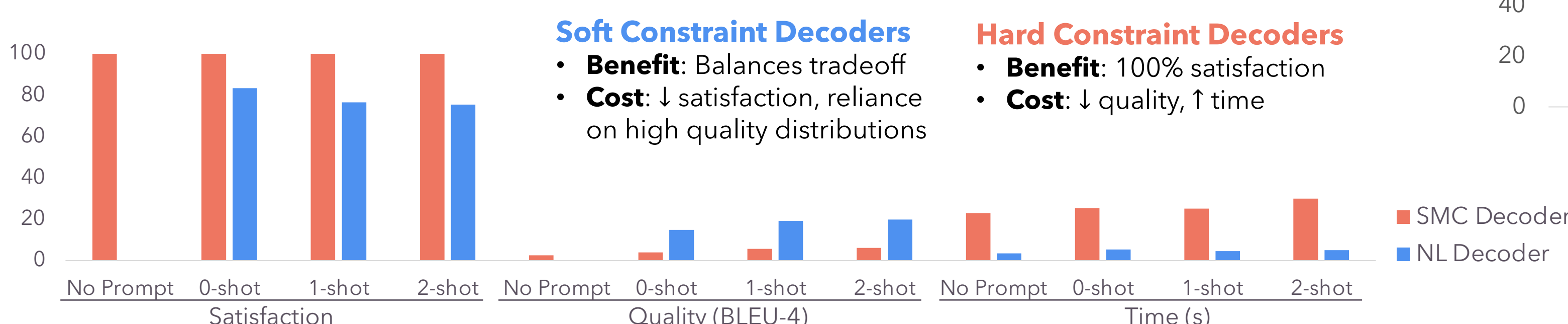
6. Results & Key Findings

		Quality (BLEU-4)	Satisfaction	Time (s)
0-shot	Prompt Only	3.74 / 6.29	11.09 / 18.84	1.84 / 1.56
	SMC	4.07 / 4	100 / 100	27.51 / 25.33
	NL	3.98 / 14.82	1 / 83.43	7.47 / 5.3
1-shot	Prompt Only	8.07 / 8.41	38.88 / 35.34	1.51 / 1.85
	SMC	4.65 / 5.66	100 / 100	34.1 / 25.14
	NL	25.73 / 19.25	39.08 / 76.55	9.37 / 4.61
2-shot	Prompt Only	9.92 / 10.6	34.34 / 38.74	1.46 / 1.83
	SMC	5.93 / 6.17	100 / 100	48.76 / 29.96
	NL	25.35 / 19.76	27.99 / 75.48	11.3 / 5.11

Legend: CNF / ABS

Prompting reduces the search space for decoders using search strategies

- 5.3
- 4.61
- 5.11

Decreased inference time despite prompt length increasing

Longer prompts increase inference time

- 1.56
- 1.85
- 1.83

ABS / ABS + NL

Shorter Prompt ← → Longer Prompt

Irrelevant context hurts satisfaction

In-context examples decreases satisfaction

- 83.43
- 76.55
- 75.48

In-context examples improves quality

- 3.98
- 25.73
- 25.35

Satisfaction / Quality (BLEU-4)

CNF + NL

Tradeoff between constraint satisfaction and inference time + quality

Soft Constraint Decoders
- **Benefit**: Balances tradeoff
- **Cost**: ↓ satisfaction, reliance on high quality distributions

Hard Constraint Decoders
- **Benefit**: 100% satisfaction
- **Cost**: ↓ quality, ↑ time

Legend: SMC Decoder / NL Decoder

Project Page   Read our paper