# Towards Robust Numerical Question Answering: Diagnosing Numerical Capabilities of NLP Systems

Jialiang Xu [1], Mengyu Zhou[2], Xinyi He [3], Shi Han [2], Dongmei Zhang [2]

[1] University of Illinois at Urbana-Champaign [2] Microsoft Research [3] Xi'an Jiaotong University



西安交通大學
XI'AN JIAOTONG UNIVERSITY

EMNLP
2022

# Table of Contents

# Table of Contents

# Motivation: Existing numerical QA systems face challenges

*Numerical Question Answering* requires **Numerical Capabilities**

### *Discrete Reasoning*

**Q**: HV captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack **the next day**. What date did the JNA form a battlegroup to counterattack? counterattack the next day. **What date** did the JNA form a battlegroup to counterattack?

**A**: 3 March 1992

### *Tabular QA*

| Year | Revenue ($) | # Sales |
|---|---|---|
| Feb | 20,000 | 10,000 |
| Mar | 23,000 | 11,000 |
| Apr | 26,000 | 12,500 |

**Q**: What's the **average** revenue **from** February **to** April?

**A**: (20000+23000+26000) / 3 = 23000

### *Math Word Problem*

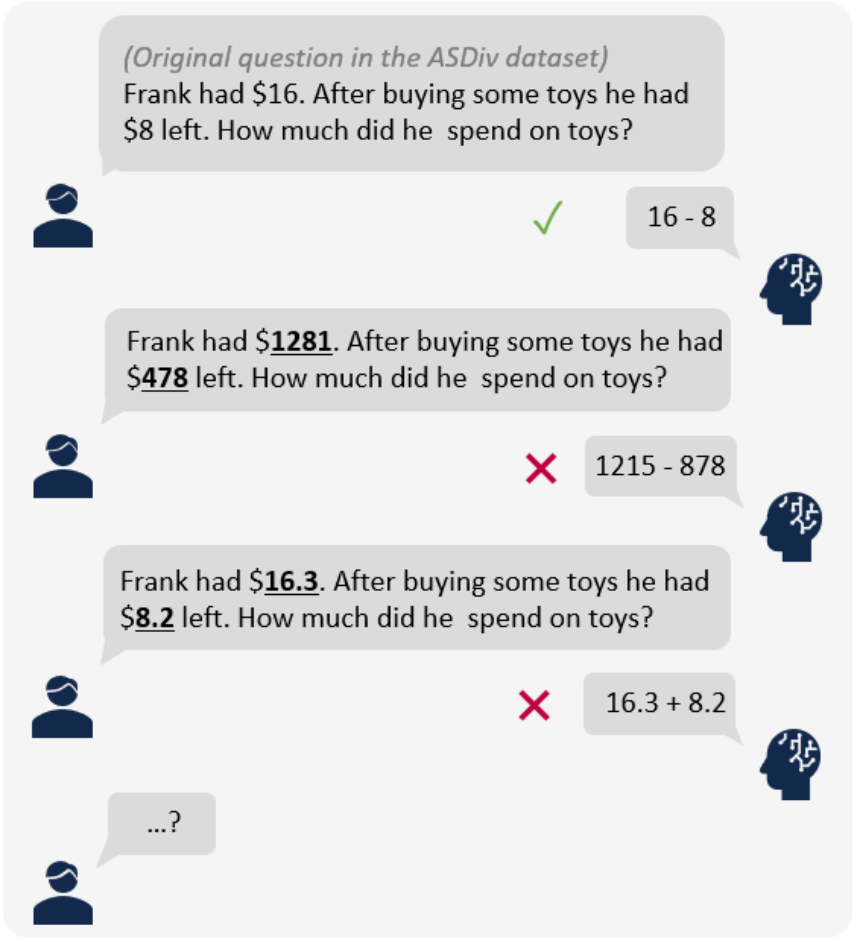
**Q**: Frank had $**16**. After buying some toys he had $**8** left. **How much** did he spend on toys?

**A**: 16 – 8 = 8

Current Numerical QA systems perform well on existing datasets

*SOTA on ASDiv-a: > 80% acc*

**However**…




- Numerical QA systems can be challenged by a variety of simple perturbations.
- Reflects **weakness** of numerical capabilities!

# Motivation: Existing numerical QA systems face challenges

## The Big Questions

- ***Which*** *numerical capabilies are needed?*
- *How to* ***quantify*** *a system's weakness?*
- *Is there a way to* ***alleviate*** *this weakness?*

*…A systematic evaluation framework is needed!*

# Table of Contents

# Methodology: The Hierarchy and Implementation (Goals)

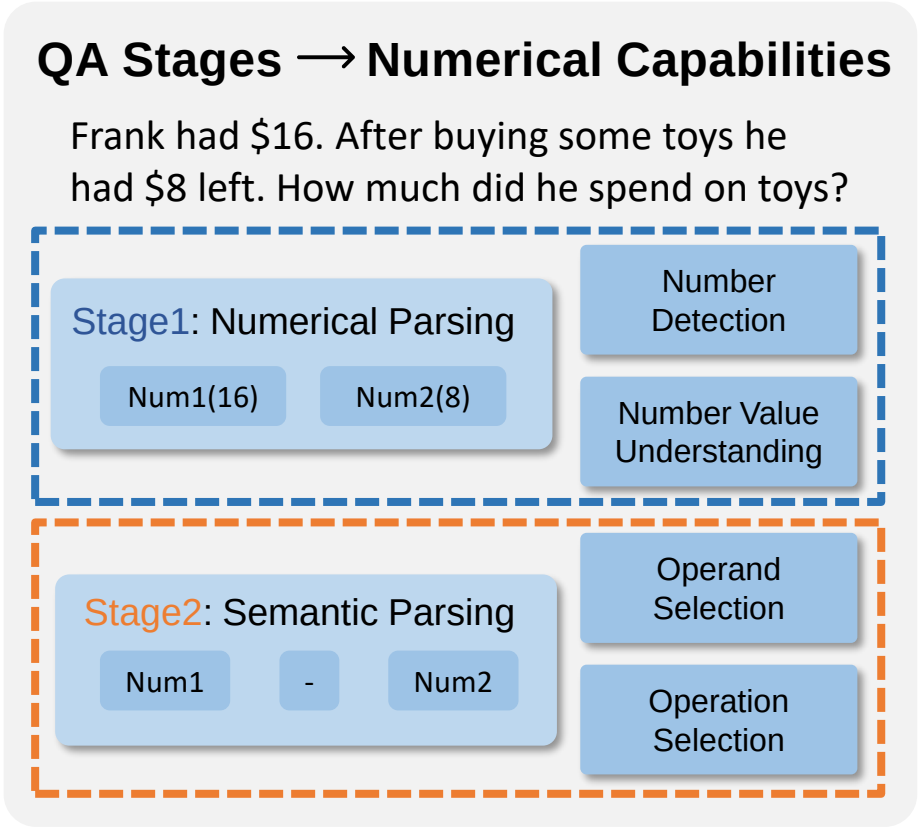
## To Answer the Questions…

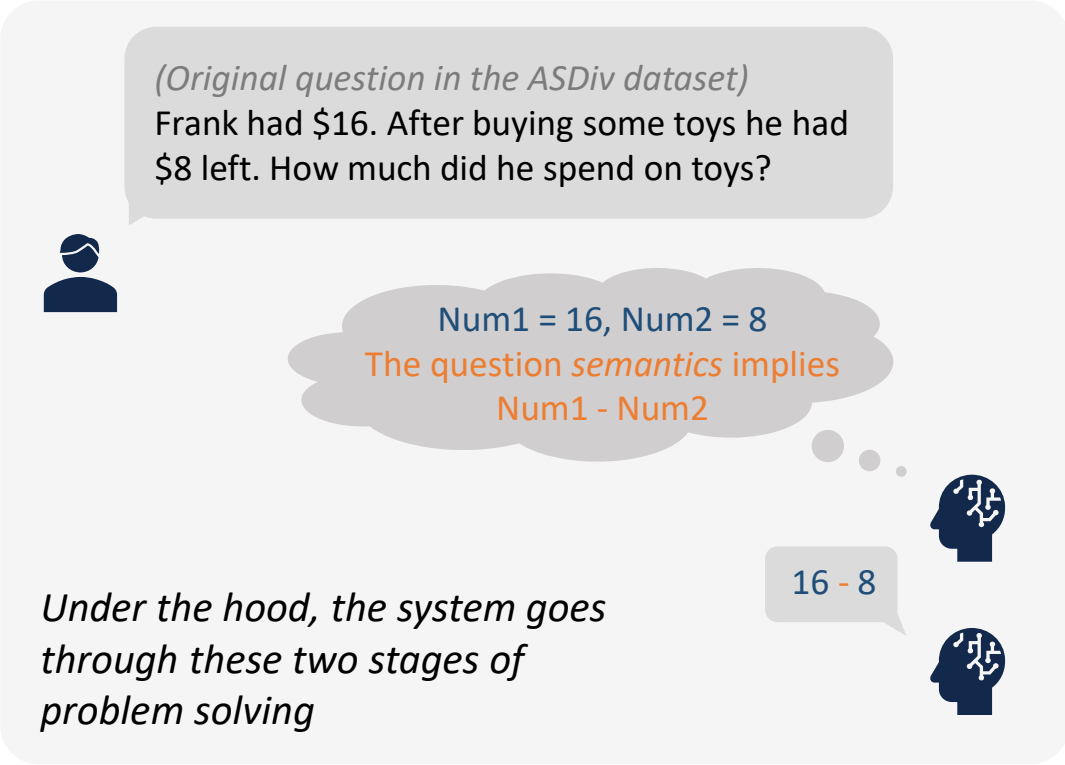
| Question | Goal | Motivation: so system designers can… |
|---|---|---|
| ***Which*** *numerical capabilies are needed?* | #1: To map out the capabilities involved in numerical QA | Have a thorough checklist of needed components. |
| *How to* ***quantify*** *a system's weakness?* | #2: To establish an indicator for systems' lack of numerical capabilities | 1. Know how severely a system lacks a capability.<br>2. Map out the weakness landscape by comparing results for different capabilities. |
| *Is there a way to* ***alleviate*** *this weakness?* | #3: To provide a baseline approach to alleviate the lack | Compare with future improvements on system architecture etc. |

# Methodology: The Hierarchy and Implementation (Goal #1)

## Goal #1

**2** Solving Stages + **4** Numerical Capabilities




These 2 stages delimit the two categories of numerical capabilities.

# Methodology: The Hierarchy and Implementation (Goal #2)

## Goal #2

**8** Perturbations + **Attack** setting

**Numerical Capabilities ⟶ Perturbations**

| Number Detection | Number Value Understanding | Operand Selection | Operation Selection |
|---|---|---|---|
| Language, Type | Noise, Distribution | Verbosity, Extra | Logic, Order |

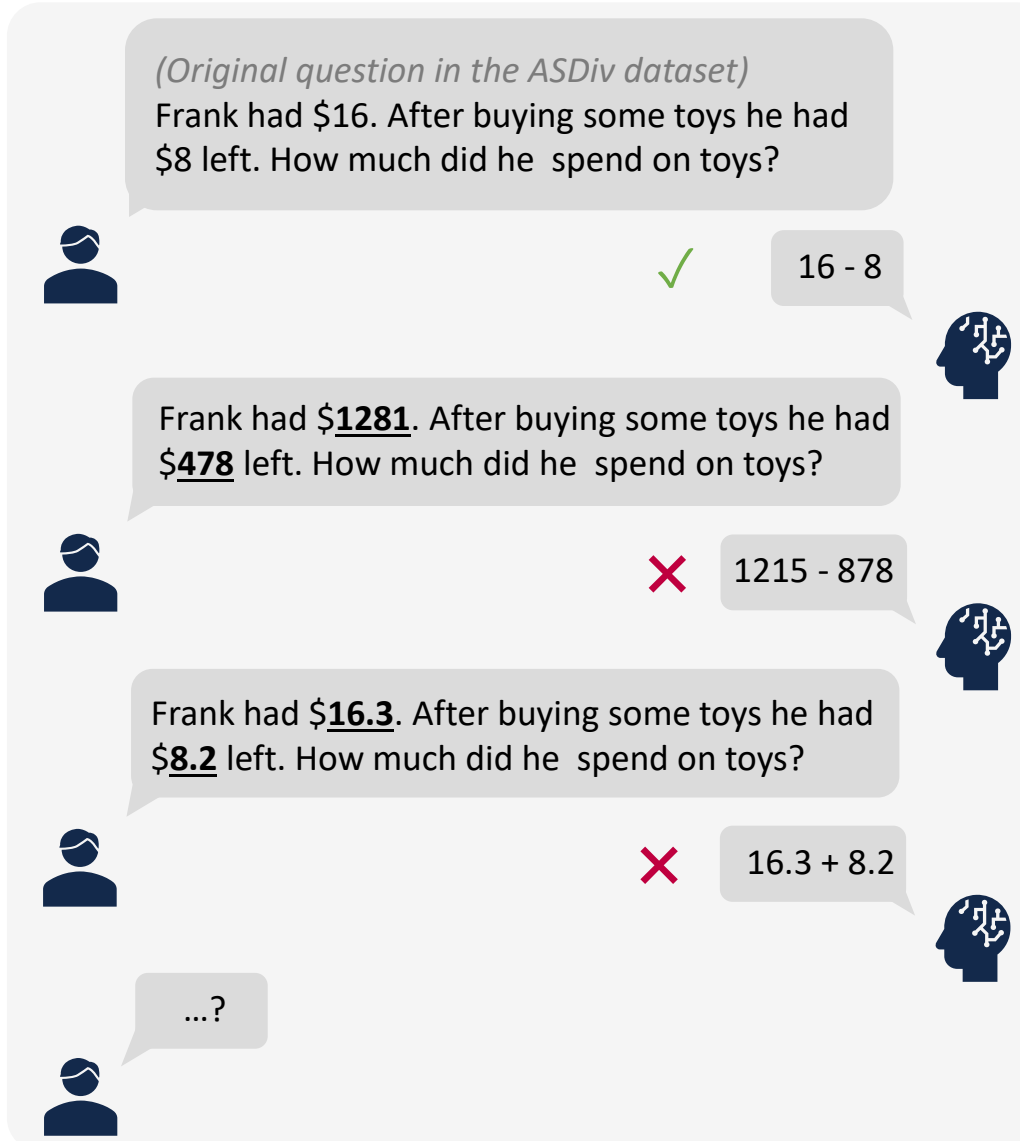
The perturbations are mutually independent

**Attack**

"Frank had $16. After buying some toys he had $8 left. How much did he spend on toys?"

↓ Noise

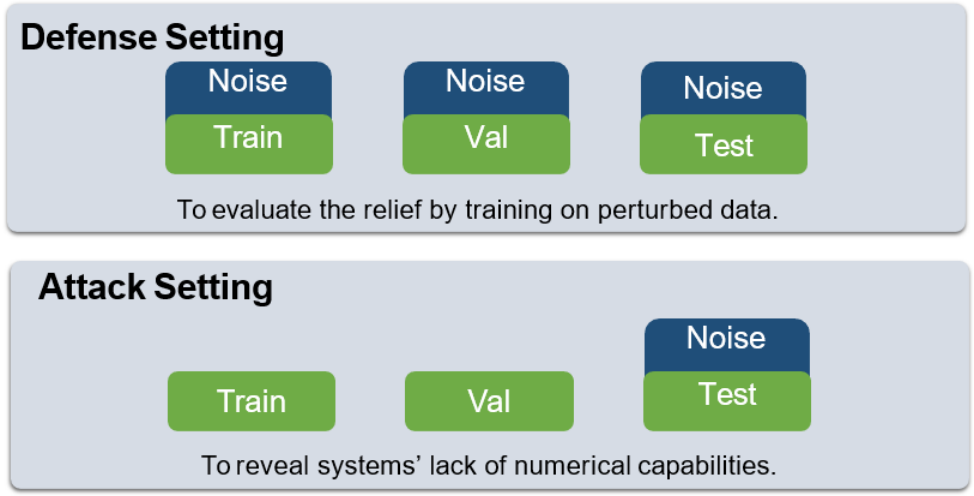
"Frank had $**16.3**. After buying some toys he had $**8.2** left. How much did he spend on toys?"



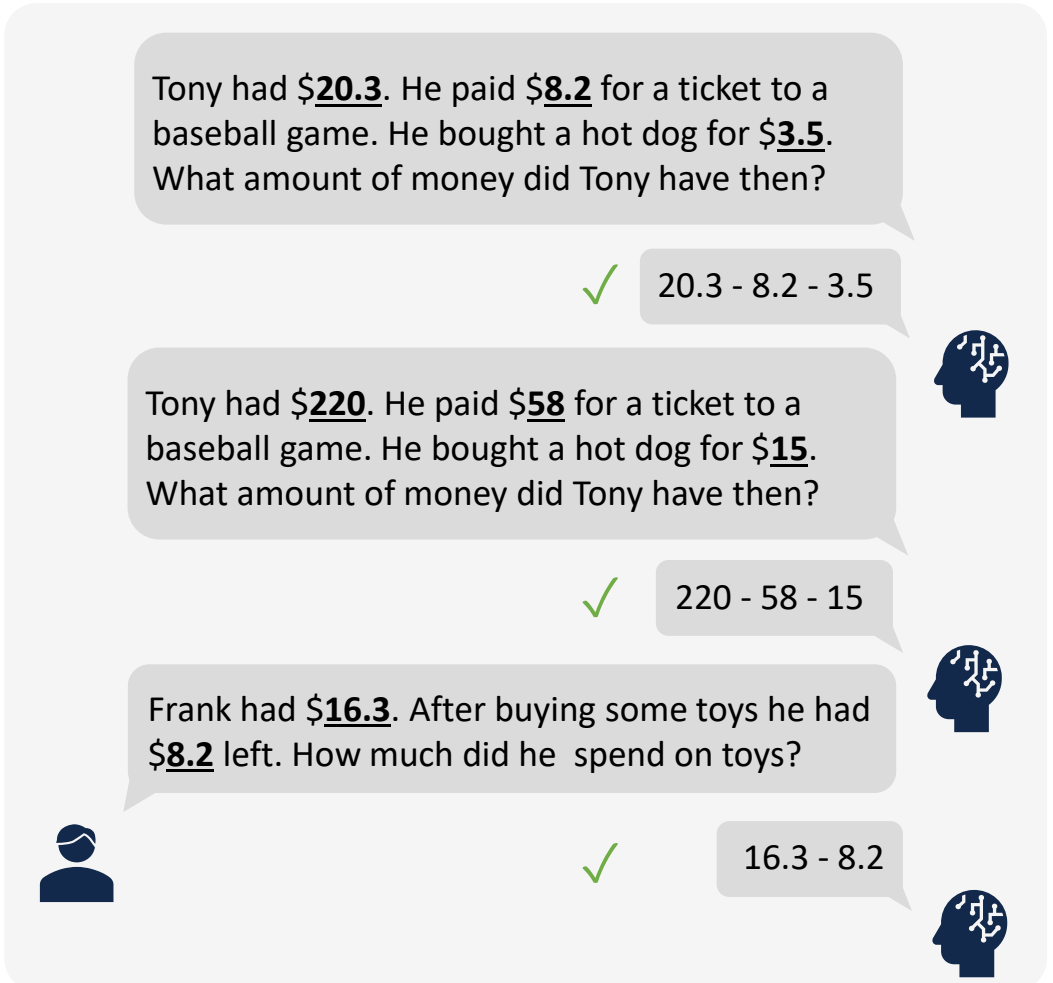
# Methodology: The Hierarchy and Implementation (Goal #3)

## Goal #3

## **Defense** setting




- The difference between the two settings is shown during training




- The systems are trained on additional samples with the perturbation first

# Methodology: The Hierarchy and Implementation (Summary)

## To Aggregate

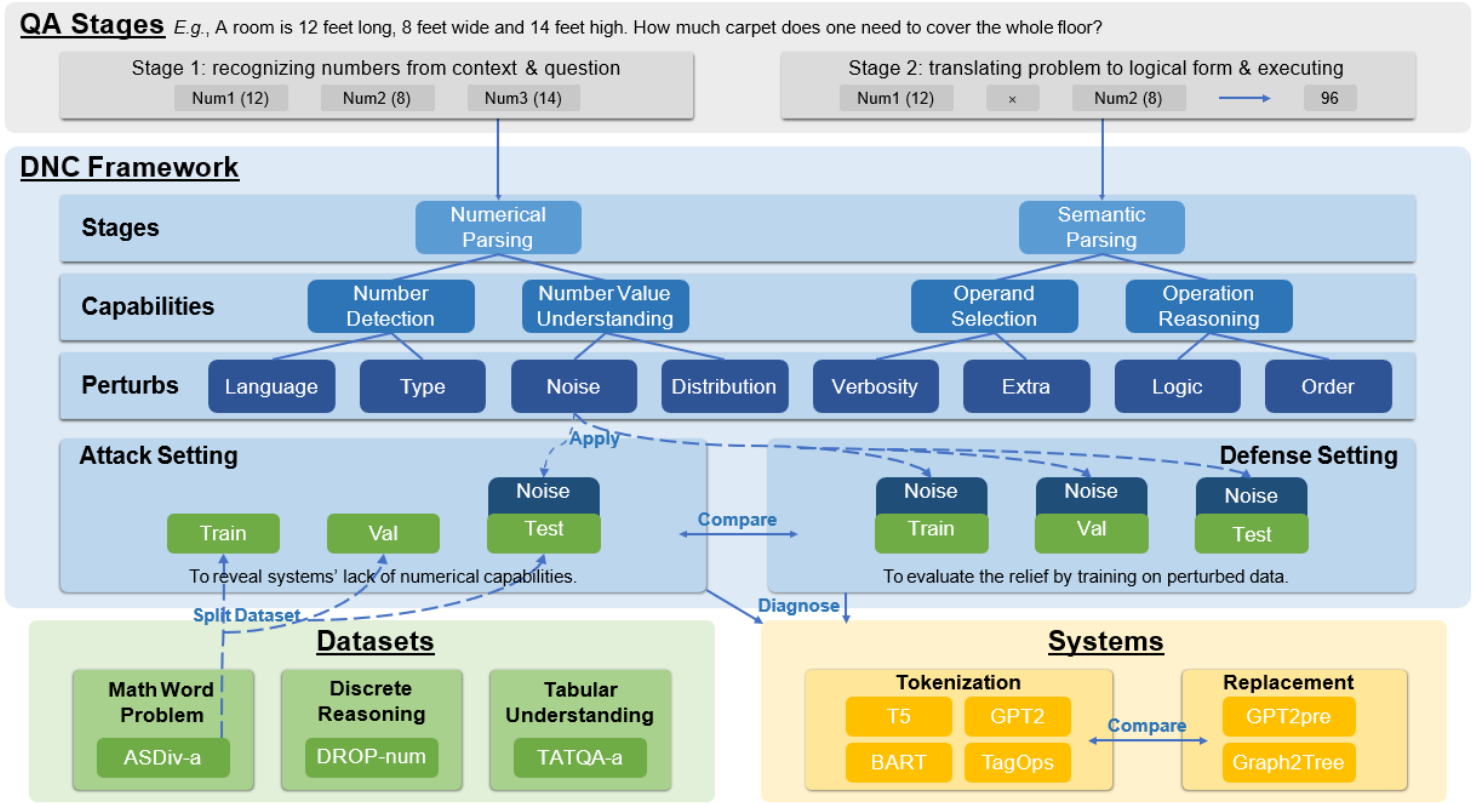


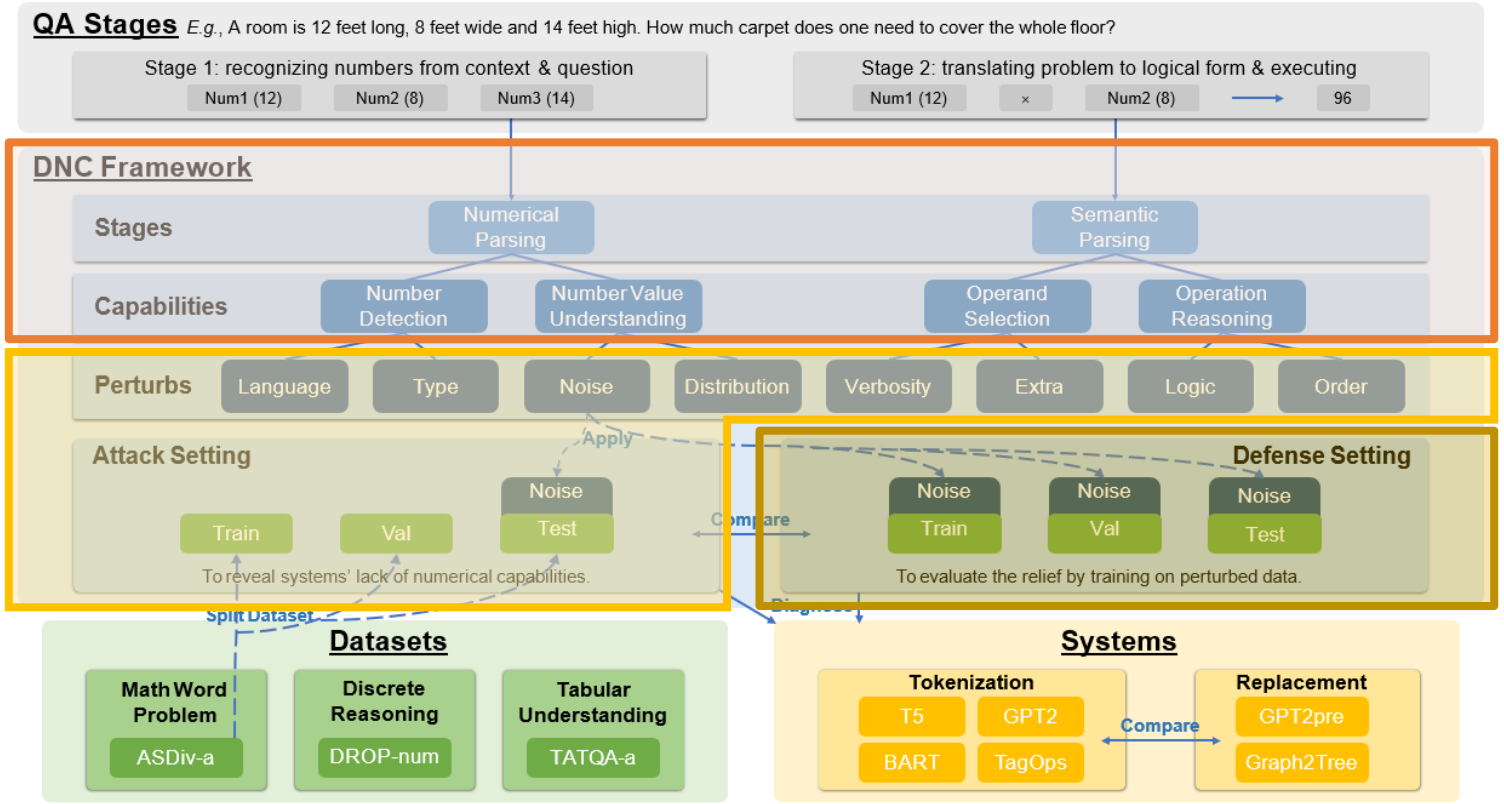

Figure 1: Overview of DNC Framework. (In our paper)



https://arxiv.org/pdf/2211.07455.pdf

# Methodology: The Hierarchy and Implementation (Summary)

## To Aggregate




Goal #1

Goal #2

Goal #3

Figure 1: Overview of DNC Framework. (In our paper)



https://arxiv.org/pdf/2211.07455.pdf

# Table of Contents

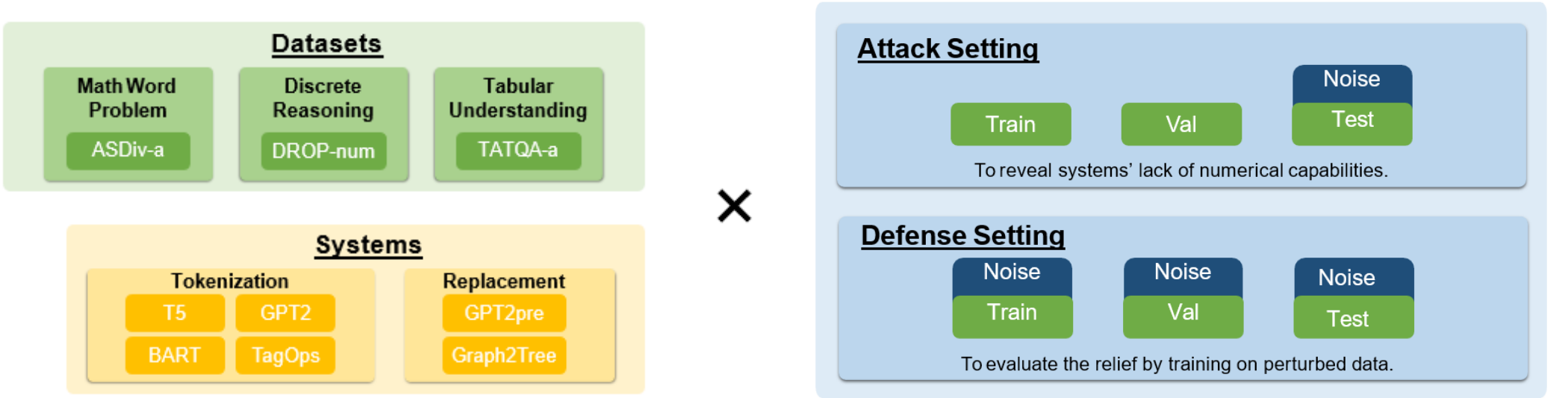
# Experiment Settings

**3** Datasets + **5** Systems

**7** Combinations of Tasks & Systems

**8** Perturbations × **2** Evaluation Settings

**1** Nvidia V100 GPU + **1** Hour on Avg



# Experiments: Results & Insights

## Experiment Results

| Configuration | | ASDiv-a | | | | | | | | DROP-num | | TATQA-a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T5 | | BART | | GPT2 | | Graph2Tree | | T5 | BART | TagOps |
| Setting | Perturbation | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | Acc | Acc | Acc |
| Attack (Δ) | Language | -18.85% | -18.85% | -23.77% | -27.05% | -12.30% | -12.30% | -7.65% | -7.38% | -10.62% | -14.73% | -18.62% |
| | Type | -37.70% | -11.48% | -32.79% | -15.57% | -17.21% | -10.66% | 0.27% | 1.09% | -7.70% | -11.06% | -5.34% |
| | Noise | -36.89% | -36.89% | -18.85% | -21.31% | -9.84% | -9.02% | 0.27% | 0.55% | - | - | - |
| | Distribution | -16.39% | -14.75% | -29.51% | -18.03% | -13.11% | -13.11% | -6.56% | -6.56% | - | - | - |
| | Verbosity | -41.80% | -44.26% | -25.41% | -29.51% | -10.66% | -11.48% | -33.33% | -33.88% | -9.58% | -13.31% | -1.90% |
| | Extra | -25.41% | -27.87% | -41.80% | -45.90% | -28.69% | -28.69% | -53.83% | -54.64% | -11.79% | -11.67% | -1.21% |
| | Logic | -29.51% | -27.87% | -36.89% | -35.25% | -25.41% | -23.77% | -28.42% | -21.86% | - | - | -14.29% |
| | Order | -34.43% | -5.74% | -33.61% | -4.10% | -27.87% | -7.38% | -33.33% | -7.10% | - | - | 1.12% |
| Defense (Δ) | Language | -12.30% | -13.93% | -19.67% | -24.59% | 2.46% | 2.46% | -7.65% | -7.38% | 0.07% | -1.84% | -7.59% |
| | Type | -11.48% | -12.30% | -4.92% | -6.56% | 3.28% | 4.10% | 1.64% | 1.91% | 0.46% | -0.95% | 2.93% |
| | Noise | -14.75% | -14.75% | -3.28% | -4.92% | 3.28% | 4.10% | 0.55% | 0.27% | - | - | - |
| | Distribution | -20.49% | -20.49% | -8.20% | -9.84% | -8.20% | -9.02% | -6.83% | -6.01% | - | - | - |
| | Verbosity | -15.57% | -16.39% | -5.74% | -7.38% | -0.82% | 0.00% | -0.27% | 1.09% | -5.13% | -1.84% | 2.25% |
| | Extra | 0.00% | 1.64% | -2.46% | -4.10% | -17.21% | -18.03% | -20.22% | -17.76% | -11.32% | -10.44% | -9.14% |
| | Logic | - | - | - | - | - | - | - | - | - | - | 13.64% |
| | Order | -25.41% | -4.10% | -27.87% | -7.38% | -1.64% | 23.77% | -29.23% | -7.92% | - | - | 19.47% |
| Original | None | 68.03% | 72.95% | 67.21% | 72.95% | 44.26% | 45.08% | 66.94% | 68.58% | 49.42% | 50.36% | 42.41% |

# Experiment Results and Insights

*Performance Change*

| Configuration | | ASDiv-a | | | | | | | | DROP-num | | TATQA-a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T5 | | BART | | GPT2 | | Graph2Tree | | T5 | BART | TagOps |
| Setting | Perturbation | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | Acc | Acc | Acc |
| Attack (Δ) | Language | -18.85% | -18.85% | -23.77% | -27.05% | -12.30% | -12.30% | -7.65% | -7.38% | -10.62% | -14.73% | -18.62% |
| | Type | -37.70% | -11.48% | -32.79% | -15.57% | -17.21% | -10.66% | 0.27% | 1.09% | -7.70% | -11.06% | -5.34% |
| | Noise | -36.89% | -36.89% | -18.85% | -21.31% | -9.84% | -9.02% | 0.27% | 0.55% | - | - | - |
| | Distribution | -16.39% | -14.75% | -29.51% | -18.03% | -13.11% | -13.11% | -6.56% | -6.56% | - | - | - |
| | Verbosity | -41.80% | -44.26% | -25.41% | -29.51% | -10.66% | -11.48% | -33.33% | -33.88% | -9.58% | -13.31% | -1.90% |
| | Extra | -25.41% | -27.87% | -41.80% | -45.90% | -28.69% | -28.69% | -53.83% | -54.64% | -11.79% | -11.67% | -1.21% |
| | Logic | -29.51% | -27.87% | -36.89% | -35.25% | -25.41% | -23.77% | -28.42% | -21.86% | - | - | -14.29% |
| | Order | -34.43% | -5.74% | -33.61% | -4.10% | -27.87% | -7.38% | -33.33% | -7.10% | - | - | 1.12% |
| Defense (Δ) | Language | -12.30% | -13.93% | -19.67% | -24.59% | 2.46% | 2.46% | -7.65% | -7.38% | 0.07% | -1.84% | -7.59% |
| | Type | -11.48% | -12.30% | -4.92% | -6.56% | 3.28% | 4.10% | 1.64% | 1.91% | 0.46% | -0.95% | 2.93% |
| | Noise | -14.75% | -14.75% | -3.28% | -4.92% | 3.28% | 4.10% | 0.55% | 0.27% | - | - | - |
| | Distribution | -20.49% | -20.49% | -8.20% | -9.84% | -8.20% | -9.02% | -6.83% | -6.01% | - | - | - |
| | Verbosity | -15.57% | -16.39% | -5.74% | -7.38% | -0.82% | 0.00% | -0.27% | 1.09% | -5.13% | -1.84% | 2.25% |
| | Extra | 0.00% | 1.64% | -2.46% | -4.10% | -17.21% | -18.03% | -20.22% | -17.76% | -11.32% | -10.44% | -9.14% |
| | Logic | - | - | - | - | - | - | - | - | - | - | 13.64% |
| | Order | -25.41% | -4.10% | -27.87% | -7.38% | -1.64% | 23.77% | -29.23% | -7.92% | - | - | 19.47% |
| Original | None | 68.03% | 72.95% | 67.21% | 72.95% | 44.26% | 45.08% | 66.94% | 68.58% | 49.42% | 50.36% | 42.41% |

## Attack

Systems experience significant performance drop from the perturbations.

## Defense

Defense mechanism helps to alleviate systems' lack of corresponding numerical capabilities.

# Experiment Results and Insights

*Most Senstive Stage*

| Configuration | | ASDiv-a | | | | | | | | DROP-num | | TATQA-a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T5 | | BART | | GPT2 | | Graph2Tree | | T5 | BART | TagOps |
| Setting | Perturbation | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | Acc | Acc | Acc |
| Attack (Δ) | Language | -18.85% | -18.85% | -23.77% | -27.05% | -12.30% | -12.30% | -7.65% | -7.38% | -10.62% | -14.73% | -18.62% |
| | Type | -37.70% | -11.48% | -32.79% | -15.57% | -17.21% | -10.66% | 0.27% | 1.09% | -7.70% | -11.06% | -5.34% |
| | Noise | -36.89% | -36.89% | -18.85% | -21.31% | -9.84% | -9.02% | 0.27% | 0.55% | - | - | - |
| | Distribution | -16.39% | -14.75% | -29.51% | -18.03% | -13.11% | -13.11% | -6.56% | -6.56% | - | - | - |
| | Verbosity | -41.80% | -44.26% | -25.41% | -29.51% | -10.66% | -11.48% | -33.33% | -33.88% | -9.58% | -13.31% | -1.90% |
| | Extra | -25.41% | -27.87% | -41.80% | -45.90% | -28.69% | -28.69% | -53.83% | -54.64% | -11.79% | -11.67% | -1.21% |
| | Logic | -29.51% | -27.87% | -36.89% | -35.25% | -25.41% | -23.77% | -28.42% | -21.86% | - | - | -14.29% |
| | Order | -34.43% | -5.74% | -33.61% | -4.10% | -27.87% | -7.38% | -33.33% | -7.10% | - | - | 1.12% |
| Defense (Δ) | Language | -12.30% | -13.93% | -19.67% | -24.59% | 2.46% | 2.46% | -7.65% | -7.38% | 0.07% | -1.84% | -7.59% |
| | Type | -11.48% | -12.30% | -4.92% | -6.56% | 3.28% | 4.10% | 1.64% | 1.91% | 0.46% | -0.95% | 2.93% |
| | Noise | -14.75% | -14.75% | -3.28% | -4.92% | 3.28% | 4.10% | 0.55% | 0.27% | - | - | - |
| | Distribution | -20.49% | -20.49% | -8.20% | -9.84% | -8.20% | -9.02% | -6.83% | -6.01% | - | - | - |
| | Verbosity | -15.57% | -16.39% | -5.74% | -7.38% | -0.82% | 0.00% | -0.27% | 1.09% | -5.13% | -1.84% | 2.25% |
| | Extra | 0.00% | 1.64% | -2.46% | -4.10% | -17.21% | -18.03% | -20.22% | -17.76% | -11.32% | -10.44% | -9.14% |
| | Logic | - | - | - | - | - | - | - | - | - | - | 13.64% |
| | Order | -25.41% | -4.10% | -27.87% | -7.38% | -1.64% | 23.77% | -29.23% | -7.92% | - | - | 19.47% |
| Original | None | 68.03% | 72.95% | 67.21% | 72.95% | 44.26% | 45.08% | 66.94% | 68.58% | 49.42% | 50.36% | 42.41% |

## Attack

Semantic Parsing causes a more severe challenge.

## Defense

The lack according to Semantic Parsing gets more recovery

# Experiment Results and Insights

*Most Senstive System*

| Configuration | | ASDiv-a | | | | | | | | DROP-num | | TATQA-a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T5 | | BART | | GPT2 | | Graph2Tree | | T5 | BART | TagOps |
| Setting | Perturbation | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | $Acc_{eq}$ | $Acc_{ans}$ | Acc | Acc | Acc |
| Attack (Δ) | Language | -18.85% | -18.85% | -23.77% | -27.05% | -12.30% | -12.30% | -7.65% | -7.38% | -10.62% | -14.73% | -18.62% |
| | Type | -37.70% | -11.48% | -32.79% | -15.57% | -17.21% | -10.66% | 0.27% | 1.09% | -7.70% | -11.06% | -5.34% |
| | Noise | -36.89% | -36.89% | -18.85% | -21.31% | -9.84% | -9.02% | 0.27% | 0.55% | - | - | - |
| | Distribution | -16.39% | -14.75% | -29.51% | -18.03% | -13.11% | -13.11% | -6.56% | -6.56% | - | - | - |
| | Verbosity | -41.80% | -44.26% | -25.41% | -29.51% | -10.66% | -11.48% | -33.33% | -33.88% | -9.58% | -13.31% | -1.90% |
| | Extra | -25.41% | -27.87% | -41.80% | -45.90% | -28.69% | -28.69% | -53.83% | -54.64% | -11.79% | -11.67% | -1.21% |
| | Logic | -29.51% | -27.87% | -36.89% | -35.25% | -25.41% | -23.77% | -28.42% | -21.86% | - | - | -14.29% |
| | Order | -34.43% | -5.74% | -33.61% | -4.10% | -27.87% | -7.38% | -33.33% | -7.10% | - | - | 1.12% |
| Defense (Δ) | Language | -12.30% | -13.93% | -19.67% | -24.59% | 2.46% | 2.46% | -7.65% | -7.38% | 0.07% | -1.84% | -7.59% |
| | Type | -11.48% | -12.30% | -4.92% | -6.56% | 3.28% | 4.10% | 1.64% | 1.91% | 0.46% | -0.95% | 2.93% |
| | Noise | -14.75% | -14.75% | -3.28% | -4.92% | 3.28% | 4.10% | 0.55% | 0.27% | - | - | - |
| | Distribution | -20.49% | -20.49% | -8.20% | -9.84% | -8.20% | -9.02% | -6.83% | -6.01% | - | - | - |
| | Verbosity | -15.57% | -16.39% | -5.74% | -7.38% | -0.82% | 0.00% | -0.27% | 1.09% | -5.13% | -1.84% | 2.25% |
| | Extra | 0.00% | 1.64% | -2.46% | -4.10% | -17.21% | -18.03% | -20.22% | -17.76% | -11.32% | -10.44% | -9.14% |
| | Logic | - | - | - | - | - | - | - | - | - | - | 13.64% |
| | Order | -25.41% | -4.10% | -27.87% | -7.38% | -1.64% | 23.77% | -29.23% | -7.92% | - | - | 19.47% |
| Original | None | 68.03% | 72.95% | 67.21% | 72.95% | 44.26% | 45.08% | 66.94% | 68.58% | 49.42% | 50.36% | 42.41% |

## Attack

Transformer-based Seq2Seq systems have larger performance drops.

## Defense

Transformer-based Seq2Seq systems benefit more from Defense.

# Thank you!


Jialiang Xu
*jx17@illinois.edu*


ArXiv Page
https://arxiv.org/abs/2211.07455


Mengyu Personal Page
*zmy.io*
*Corresponding Author*
*mezho@microsoft.com*


Project Page
Code, Data, Video

Jialiang Personal Page
*I'm applying for grad school starting Fall 2023!*