FACTUM

Mechanistic Detection of Citation Hallucination in Long-Form RAG

Maxime Dassen[1] · Rebecca Kotula[2] · Kenton Murray[3] · Andrew Yates[3]
Dawn Lawrie[3] · Efsun Kayi[4] · James Mayfield[3] · Kevin Duh[3]

[1] *University of Amsterdam* [2] *Department of Defense* [3] *HLTCOE, Johns Hopkins University* [4] *APL, Johns Hopkins*

Content

Problem statement

What we aim to solve and why it matters for the field.

Content

Problem statement

What we aim to solve and why it matters for the field.

Our method

How the residual stream decomposition is utilized

Content

Problem statement

What we aim to solve and why it matters for the field.

Our method

How the residual stream decomposition is utilized

Results & Findings

Our findings and performance evaluation

Content

Problem statement

What we aim to solve and why it matters for the field.

Our method

How the residual stream decomposition is utilized

Results & Findings

Our findings and performance evaluation

Conclusions

What are the key takeaways of this study

Long-Form RAG: Reports with Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"

↓

Generated report example

Pocket parks are small-scale urban open spaces that provide a refuge against city crowds **[7]**. They are a solution to offset the lack of green spaces in urban centers **[7]**. The benefits include reducing air pollution **[1]**. They also improve rainwater management **[1]**. Pocket parks increase urban green space per capita **[1]**.

Looks good right?

Long-Form RAG: Reports with Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"

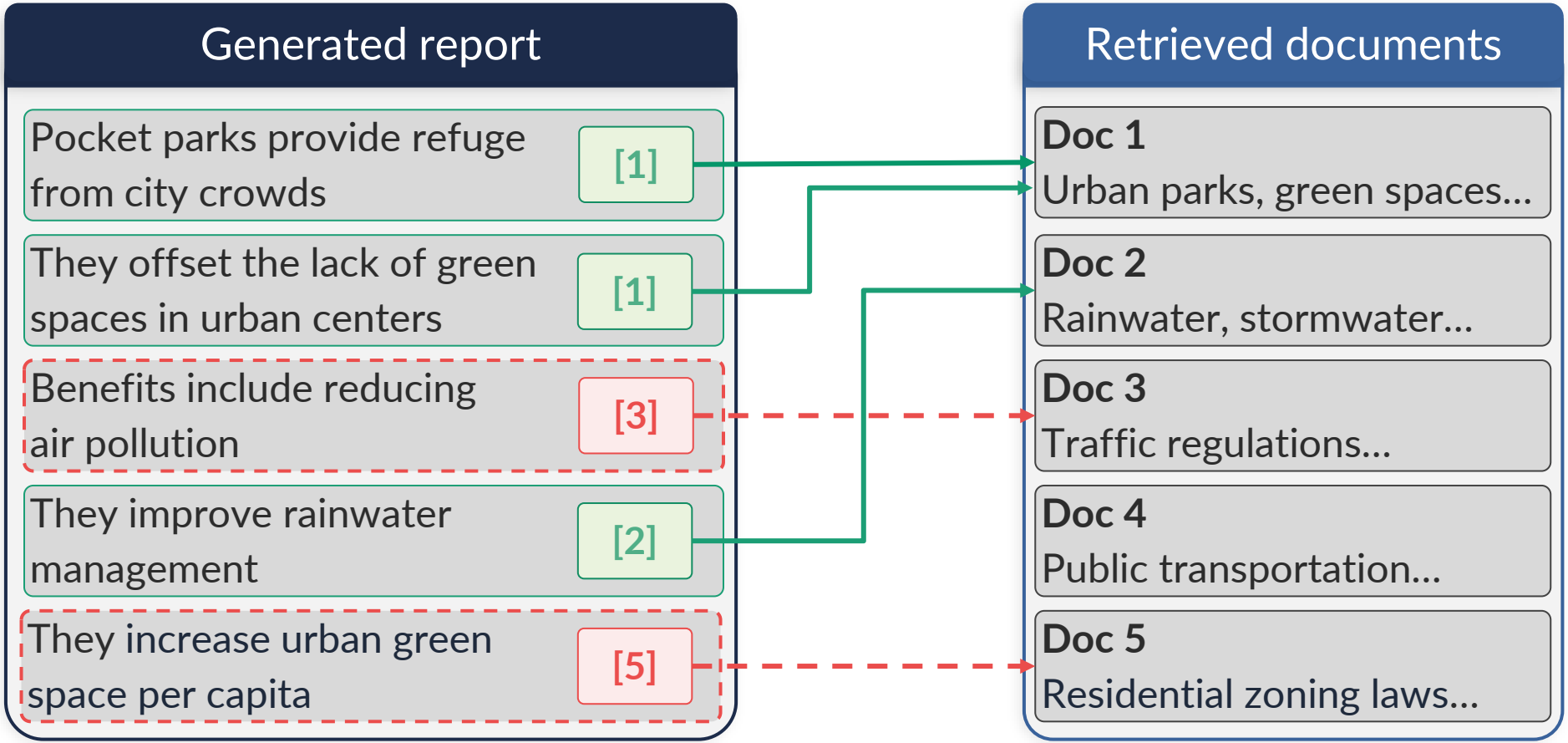
↓

Generated report example

Pocket parks are small-scale urban open spaces that provide a refuge against city crowds [7]. They are a solution to offset the lack of green spaces in urban centers [7]. The benefits include reducing air pollution [1]. They also improve rainwater management [1]. Pocket parks increase urban green space per capita [1].

Correct citation:
cited source attests the claim

Hallucinated citation:
cited source does not attest


Generated report
Pocket parks provide refuge from city crowds
[1]
They offset the lack of green spaces in urban centers
[1]
Benefits include reducing air pollution
[3]
They improve rainwater management
[2]
They increase urban green space per capita
[5]
Retrieved documents
Doc 1
Urban parks, green spaces...
Doc 2
Rainwater, stormwater...
Doc 3
Traffic regulations...
Doc 4
Public transportation...
Doc 5
Residential zoning laws...

Goal: Detect Hallucinated Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"



Generated report example

Pocket parks are small-scale urban ope spaces that pro a refuge against city crowds [7]. They are a solution to o set the lack of green sp es in urban centers [7]. The benefits include reducing a pollution [1]. They also i prove rainwater management [1]. Pocket parks incre e urban green space pe capita [1].

Goal: Detect Hallucinated Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"

Generated report example

Pocket parks are small-scale urban open spaces that provide a refuge against city crowds [7]. They can help to offset the lack of green spaces in urban centers [7]. The benefits include reducing air pollution [1]. They also improve rainwater management [1]. Pocket parks increase urban green space per capita [1].



Goal: Detect Hallucinated Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"



Generated report example

Pocket parks are small-scale urban ope spaces that pro a refuge against city crowds [7]. They are a solution to o set the lack of green spa es in urban centers [7]. The benefits include reducing a pollution [1]*. They also prove rainwater management [1]. Pocket parks incre e urban green space pe capita [1].

Goal: Detect Hallucinated Citations

Query: "Report on the benefits of pocket parks in high-density urban settings"

↓

Generated report example

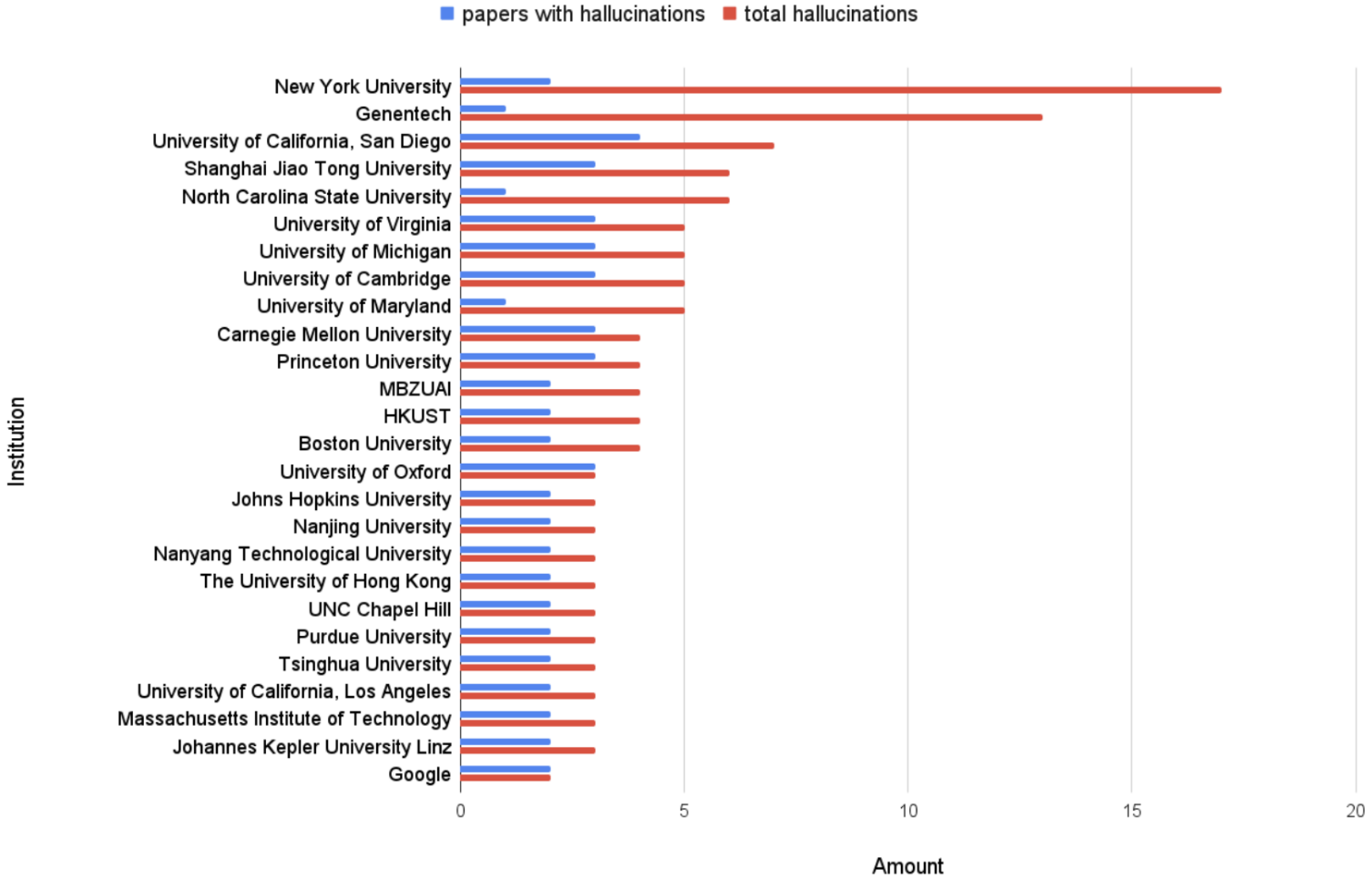
Pocket parks are small-scale urban open spaces that provide a refuge against city crowds [7]. They are a solution to offset the lack of green spaces in urban centers [7]. The benefits include reducing air pollution [1]*. They also improve rainwater management [1]. Pocket parks increase urban green space per capita [1]*.

Visual warning indicating the citation is highly suspect.


[1]
[1]
[1]
[1]
[1]
[1]
[1]
[1]
[1]
Pocket parks are small-scale urban open spaces
that provide a refegue against city crowds [7].
They are a solution to offset the lack of green
spaces in urban centers [7]. The benefits
prcvide reducing air pollution [11].
The also nimprove rainwater management [11].
Pocket parks increase urban green space can cita.


[7]
[7]
[7]
[7]
[7]
[1]
[1]
[1]
[1]
Pocket parks are small-scale urban open spaces that provide a refegue against city crowds [7]. They are a solution to offset the lack of green spaces in urban centers [7]. The benefits provide reducing air pollution [1], The also improve rainwater management [11]. Pocket parks increase urban green space capita cita
Detected hallunicated citation

Number of hallucinations by author affiliation



Published Paper	GPTZero Scan	Example of Verified Hallucination	Comment
SimWorld: An Open-ended Simulator for Agents in Physical and Social Worlds	Sources AI	John Doe and Jane Smith. Webvoyager: Building an end-to-end web agent with large multimodal models. arXiv preprint arXiv:2401.00001, 2024.	Article with a matching title exists here. Authors are obviously fabricated. arXiv ID links to a different article.
Unmasking Puppeteers: Leveraging Biometric Leakage to Expose Impersonation in AI-Based Videoconferencing	Sources AI*	John Smith and Jane Doe. Deep learning techniques for avatar-based interaction in virtual environments. IEEE Transactions on Neural Networks and Learning Systems, 32(12):5600-5612, 2021. doi: 10.1109/TNNLS.2021.3071234. URL https://ieeexplore.ieee.org/document/307123	No author or title match. Doesn't exist in publication. URL and DOI are fake.

Research questions

RQ 1. How can we detect hallucinated citations?

Research questions

RQ 1. How can we detect hallucinated citations?

RQ 2. How can we utilize model internals to detect hallucinated citations?

RQ 1. How can we detect hallucinated citations?

Prior work

Black-Box Approaches

- LLM judge
- Follow-up QA
- Output consistency sampling

Can flag hallucinations, but cannot reveal the underlying mechanistic failure.

White-Box Approaches

- **Logits**, **Attention-based** or **FFN-based** signals

To measure model confidence

To measure retrieved evidence

To measure the model's parametric knowledge

Prior work

Black-Box Approaches

- LLM judge
- Follow-up QA
- Output consistency sampling

Can flag hallucinations, but cannot reveal the underlying mechanistic failure.

White-Box Approaches

- **Logits**, **Attention-based** or **FFN-based** signals


Hallucinations can be highly confident


Dilutes context signal with prompt, query, and attention sink noise

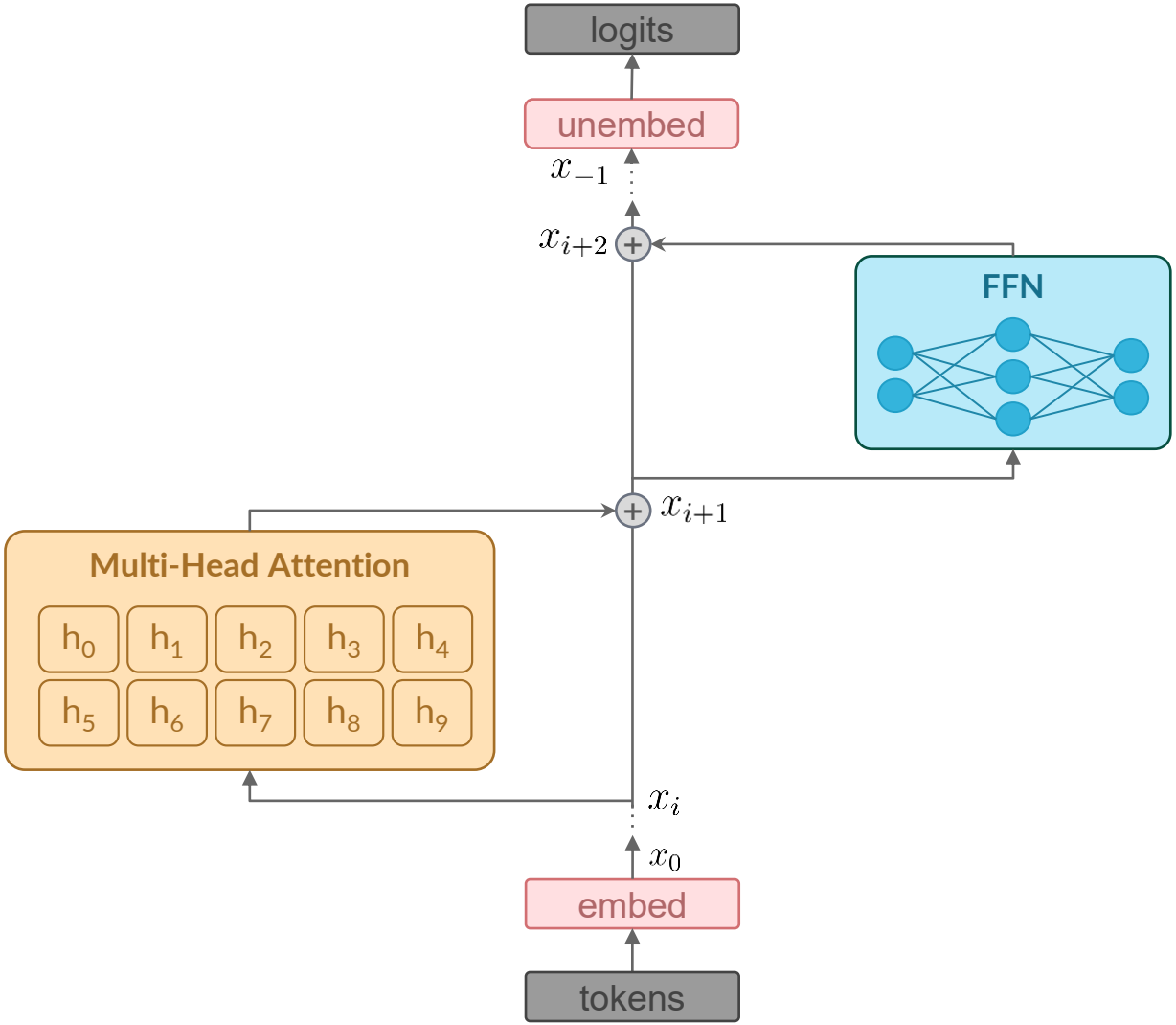

Relies on logits: unreliable proxy for FFN activity

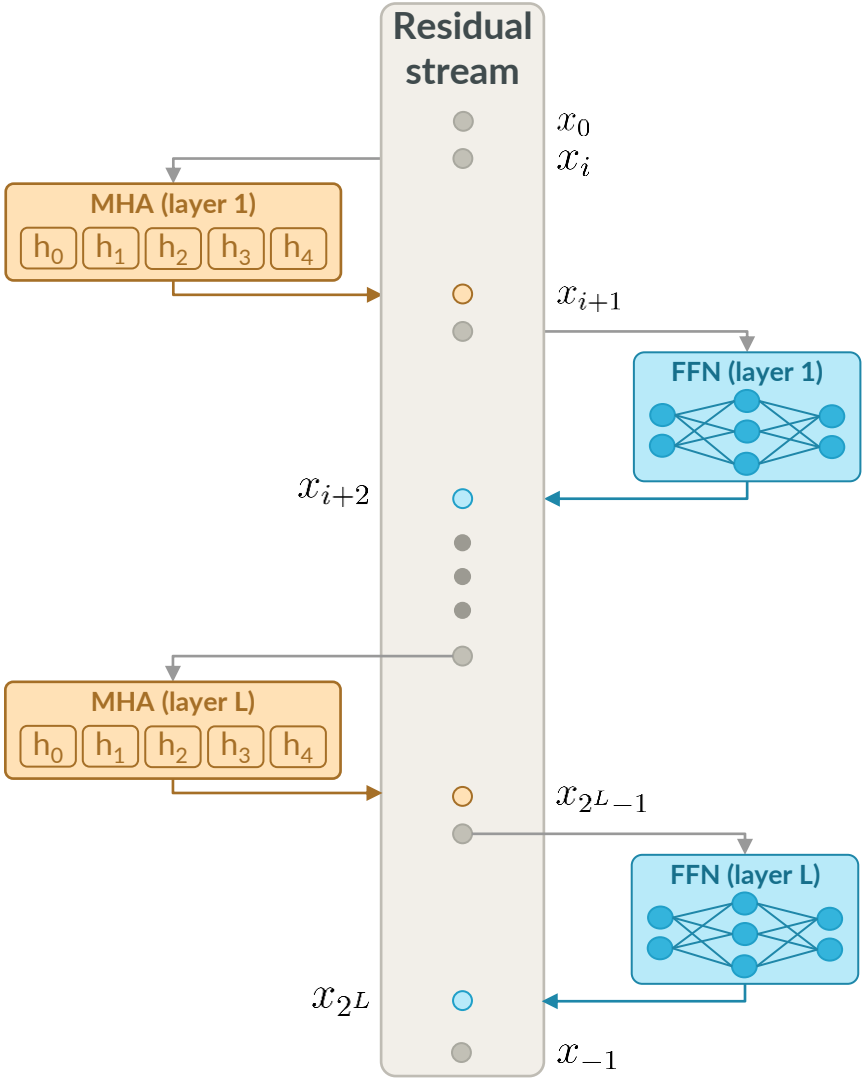

Treat Attn & FFN as independent

FACTUM improves on this by considering **cleaner**, more **targeted** signals, and by measuring how ***Attention*** and ***FFN*** **coordinate**, not just what they each do independently.

RQ 2. How can we utilize model internals to detect hallucinated citations?

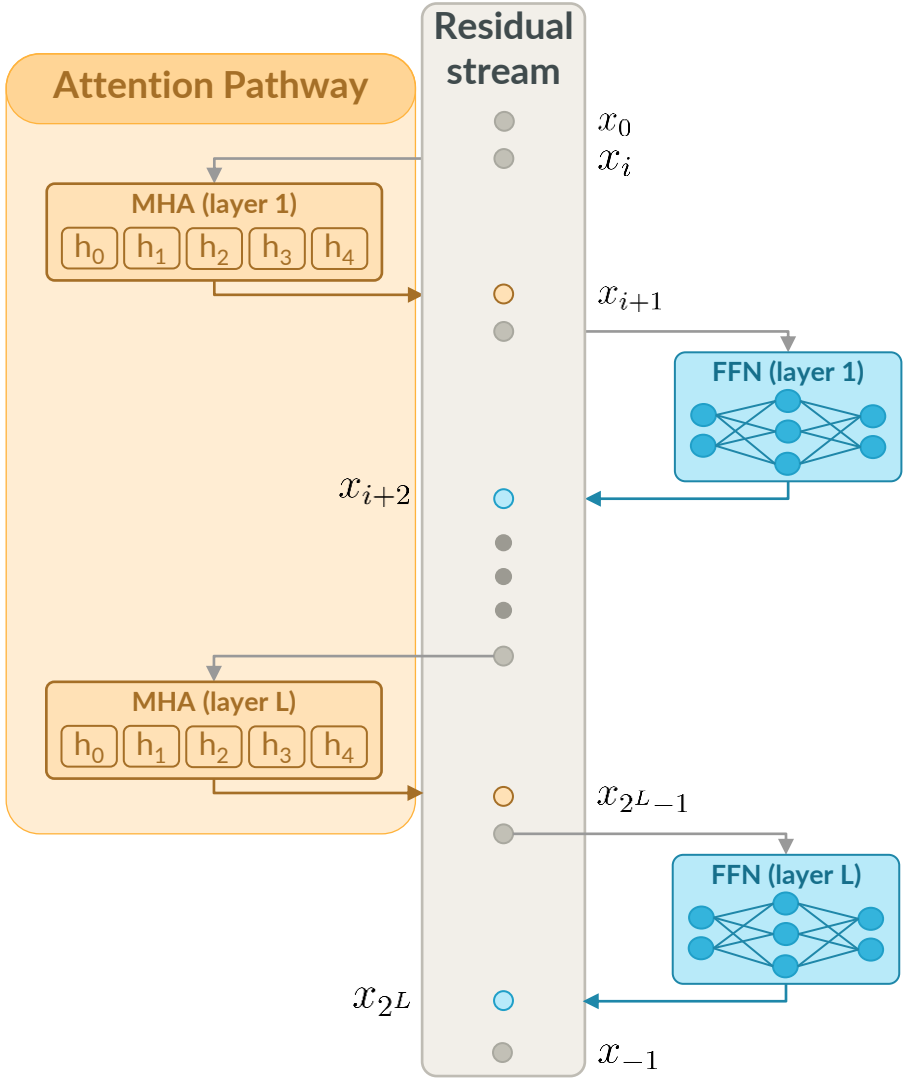
RQ 2. How can we measure the individual contribution of MHA & FFN and their coordination at citation time?

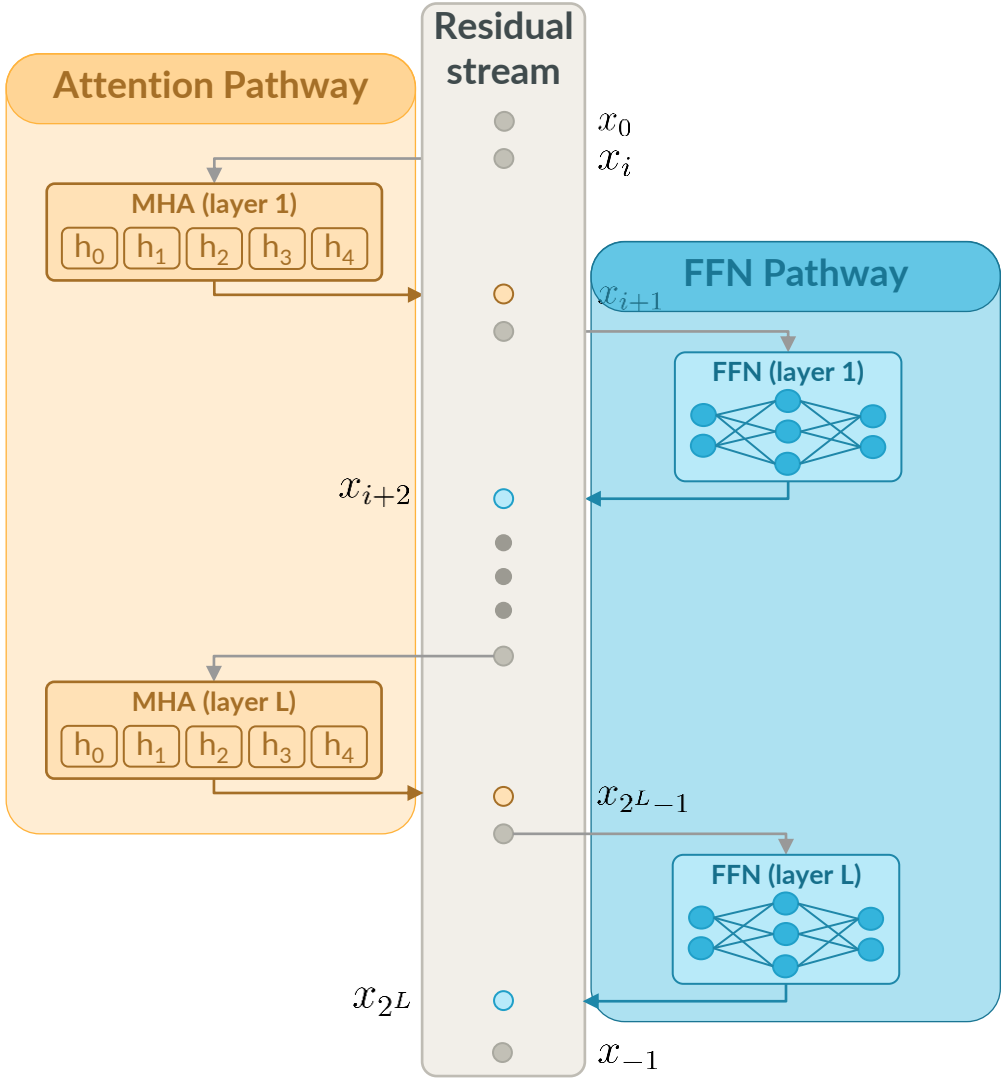

logits
unembed
x_{-1}
x_{i+2}
FFN
x_{i+1}
Multi-Head Attention
h0 h1 h2 h3 h4
h5 h6 h7 h8 h9
x_i
x_0
embed
tokens
One residual block


Residual stream
x_0
x_i
MHA (layer 1)
h0 h1 h2 h3 h4
x_{i+1}
FFN (layer 1)
x_{i+2}
MHA (layer L)
h0 h1 h2 h3 h4
x_{2^L-1}
FFN (layer L)
x_{2^L}
x_{-1}

Reads from the context

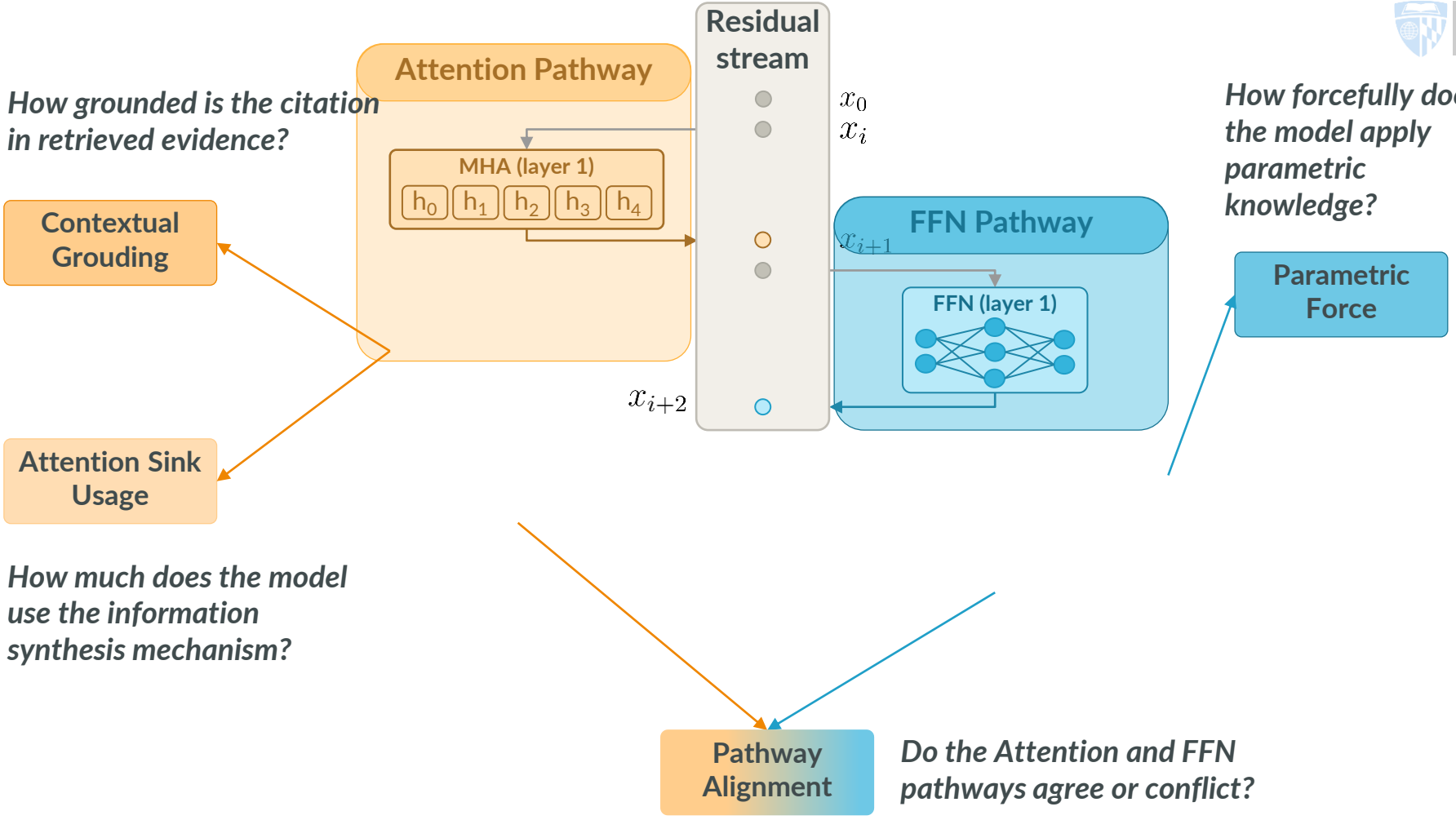
"the reader"





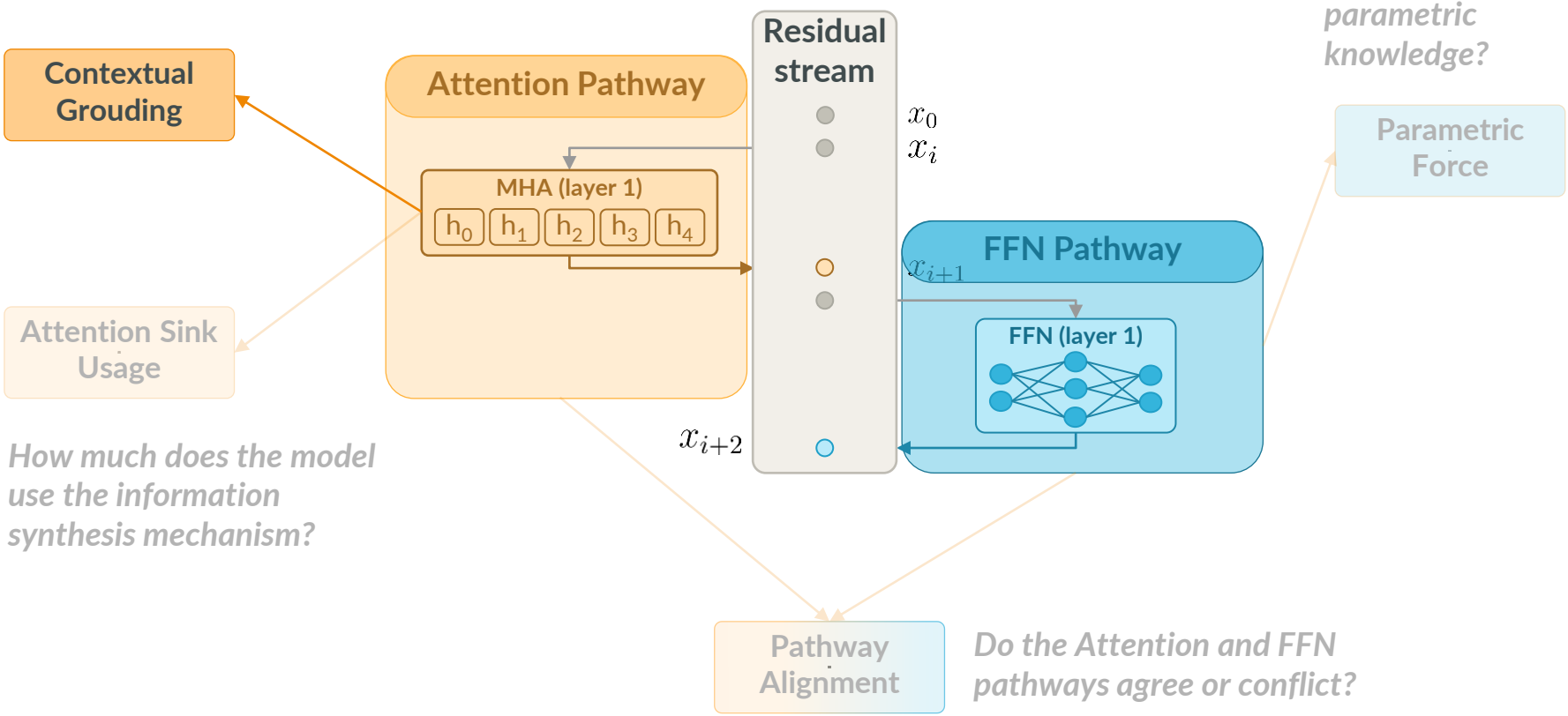
Recalls from parametric memory

"the recaller"


How grounded is the citation in retrieved evidence?
Contextual Grouding
Attention Sink Usage
How much does the model use the information synthesis mechanism?
Attention Pathway
MHA (layer 1)
h0 h1 h2 h3 h4
Residual stream
x_0
x_i
x_{i+1}
x_{i+2}
FFN Pathway
FFN (layer 1)
How forcefully does the model apply parametric knowledge?
Parametric Force
Pathway Alignment
Do the Attention and FFN pathways agree or conflict?

How grounded is the citation in retrieved evidence?

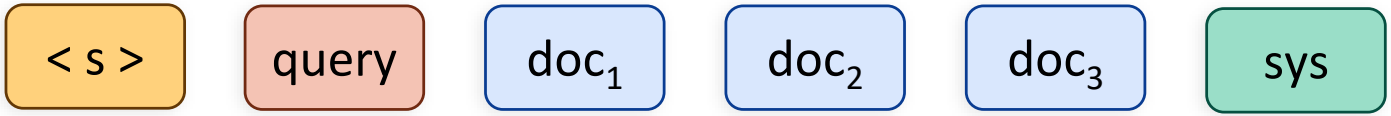
How forcefully does the model apply parametric knowledge?



How much does the model use the information synthesis mechanism?

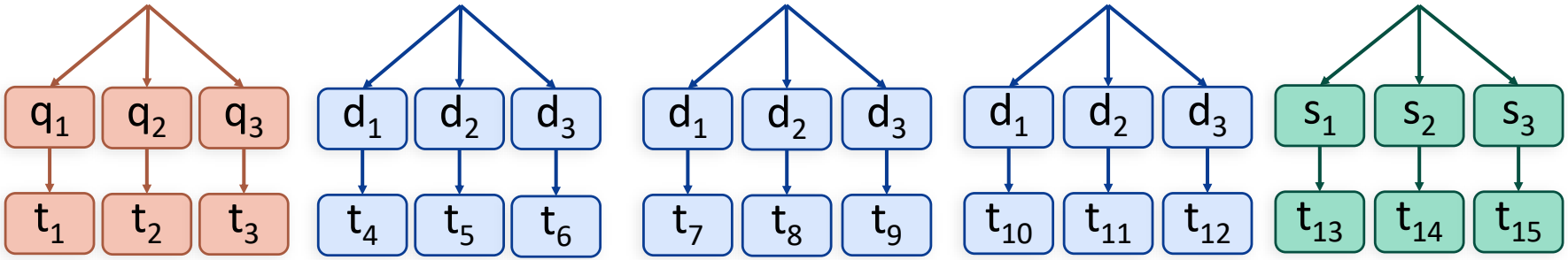
Do the Attention and FFN pathways agree or conflict?

Contextual grounding

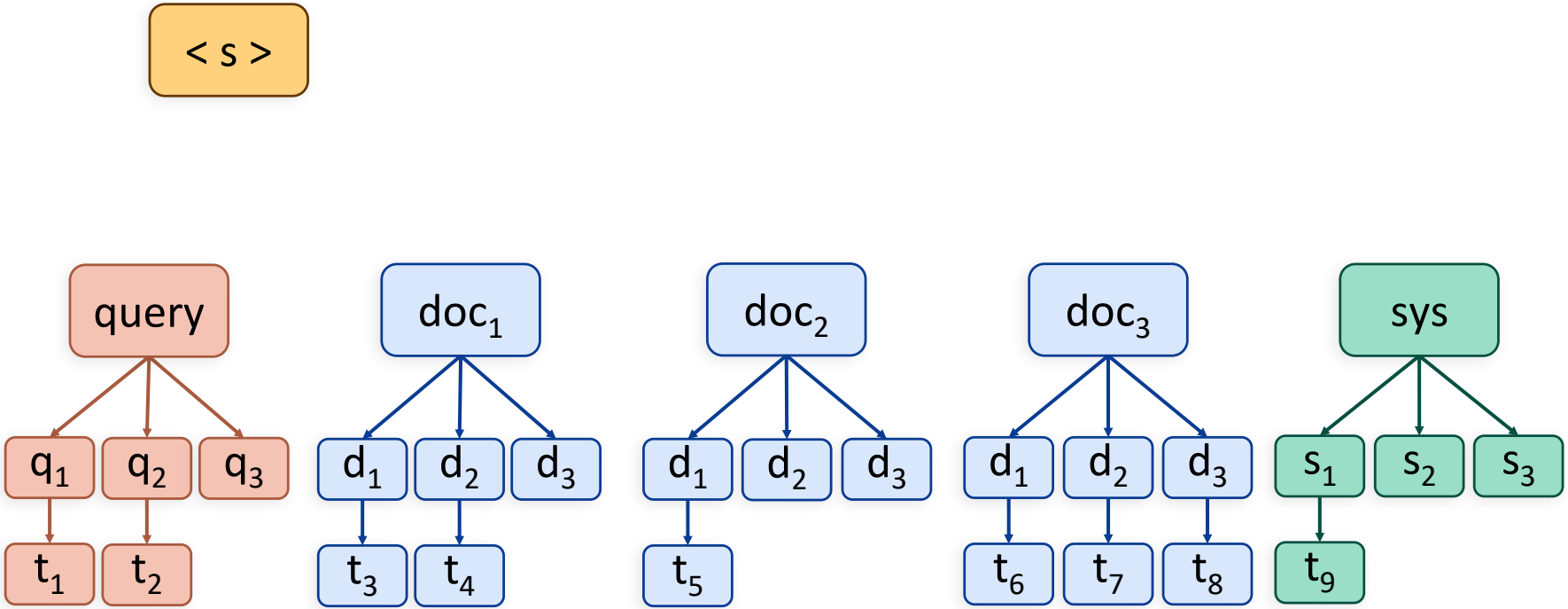


"Write a report about .. based on the following context"

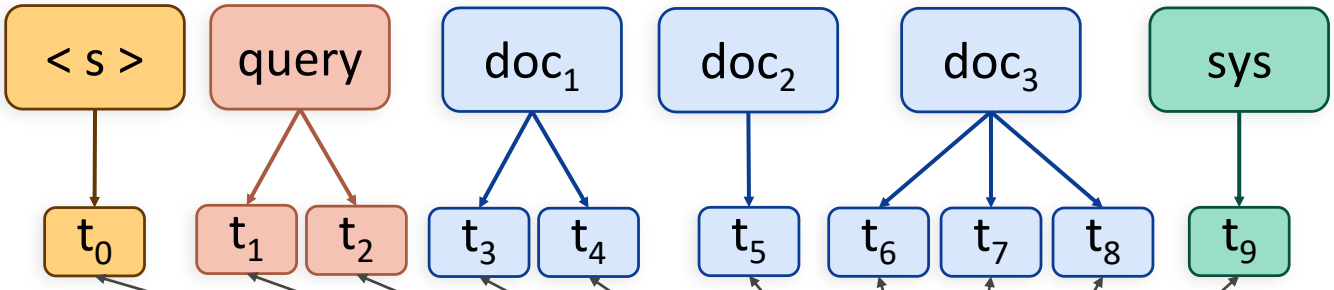
"Provide inline citations for each claim that you make"



Contextual grounding



Contextual grounding



input







gen

claim1

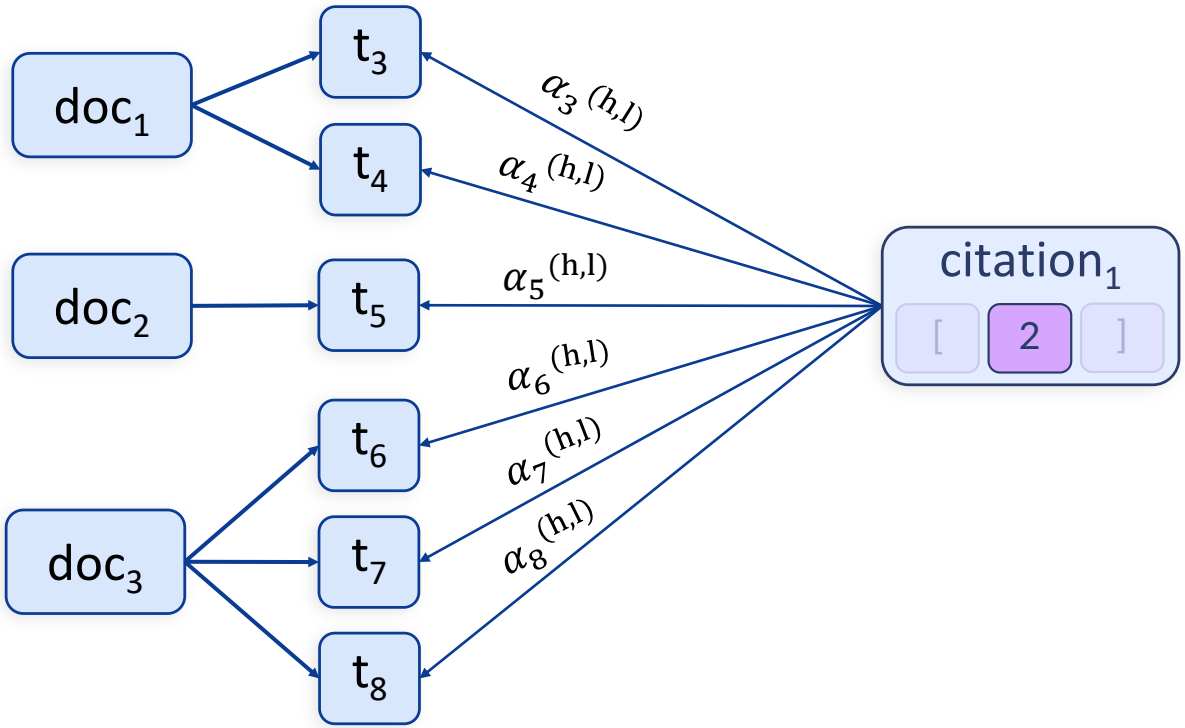
t0 t1 t2 t3 t4

citation1

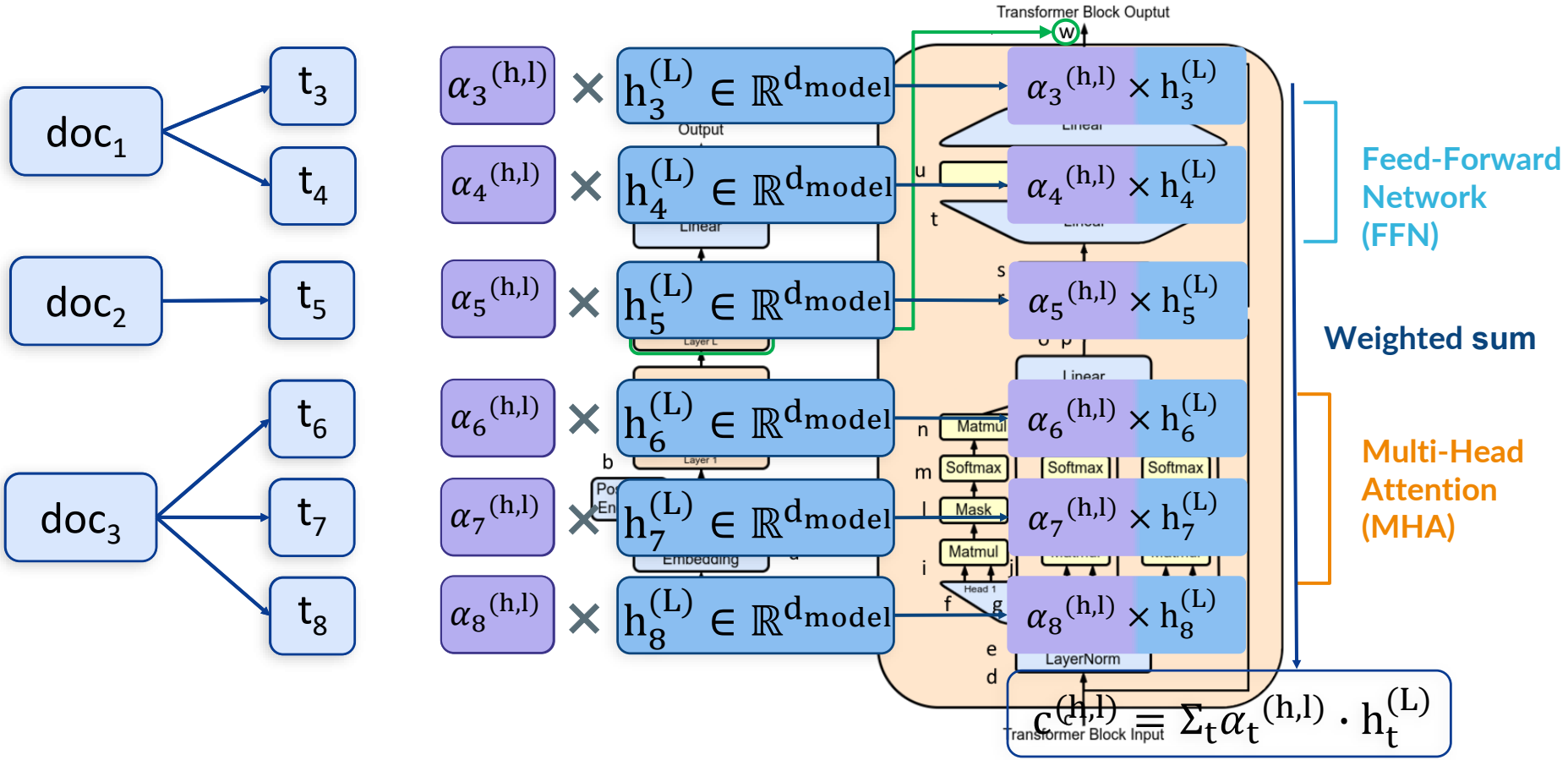
t5 t6 t7

How grounded is the citation in retrieved evidence?

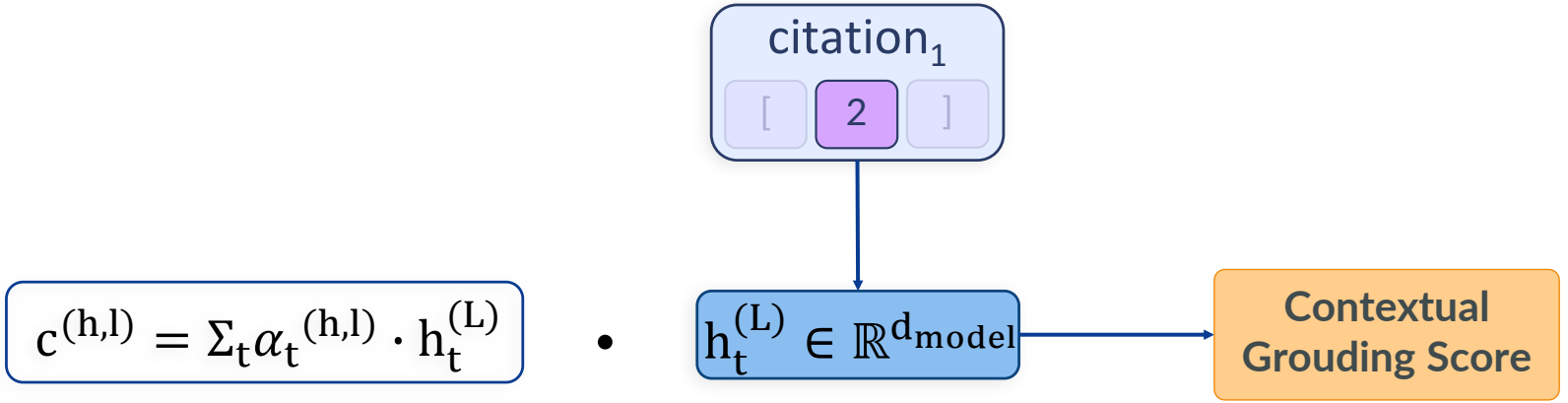
Contextual grounding



Contextual grounding



Contextual grounding

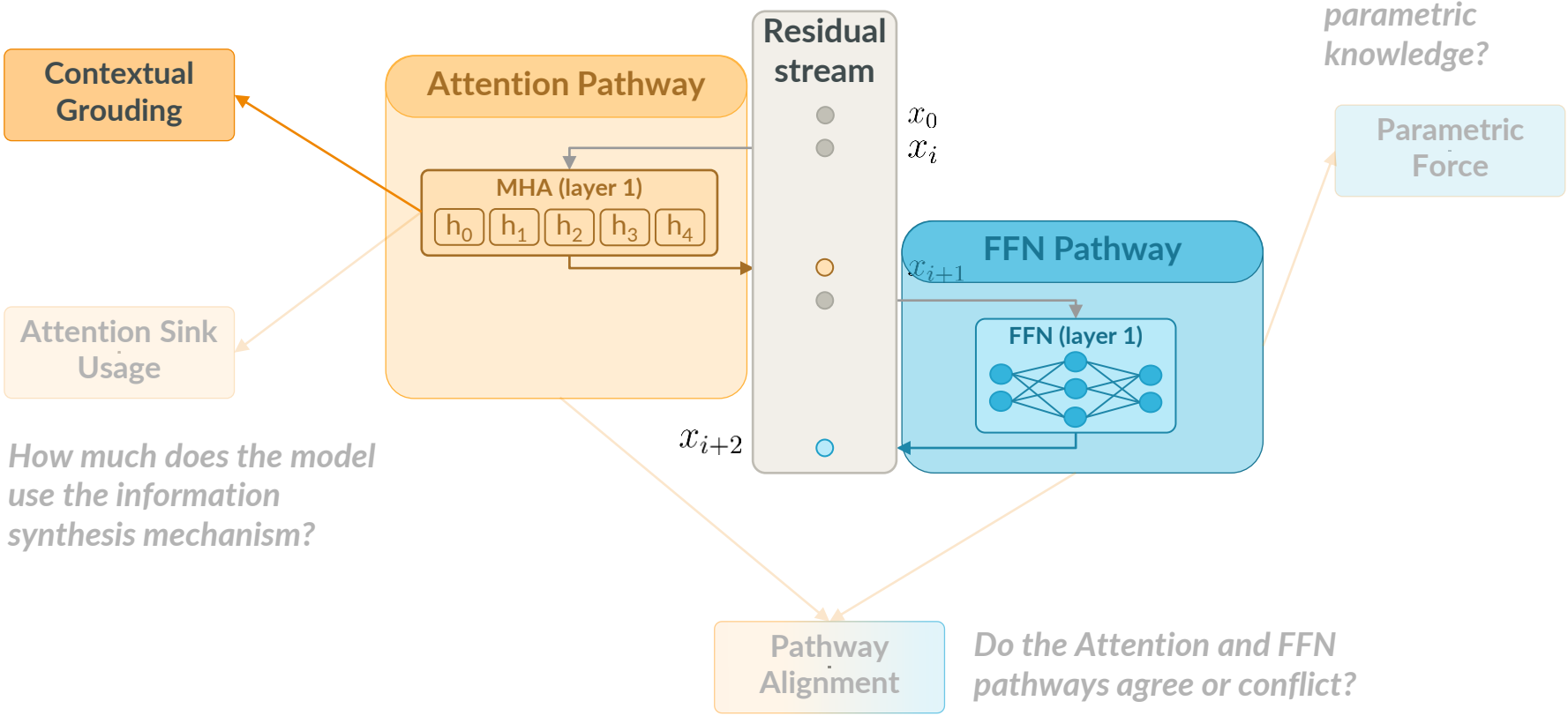


Semantic summary of the evidence the model looked at during citation generation

Is the citation token's own meaning consistent with the evidence it attended to?

How grounded is the citation in retrieved evidence?

How forcefully does the model apply parametric knowledge?

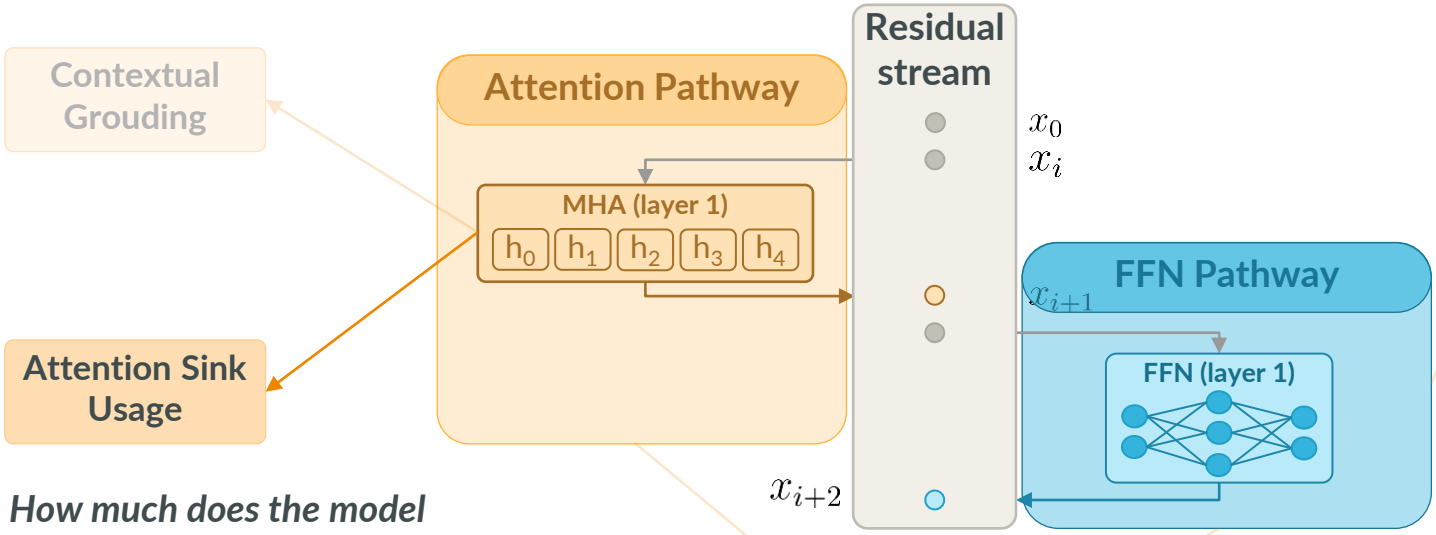


How much does the model use the information synthesis mechanism?

Do the Attention and FFN pathways agree or conflict?

How grounded is the citation in retrieved evidence?

How forcefully does the model apply parametric knowledge?



How much does the model use the information synthesis mechanism?

Pathway Alignment

Do the Attention and FFN pathways agree or conflict?

Why are attention sinks important?

Universal

Found in every modern transformer

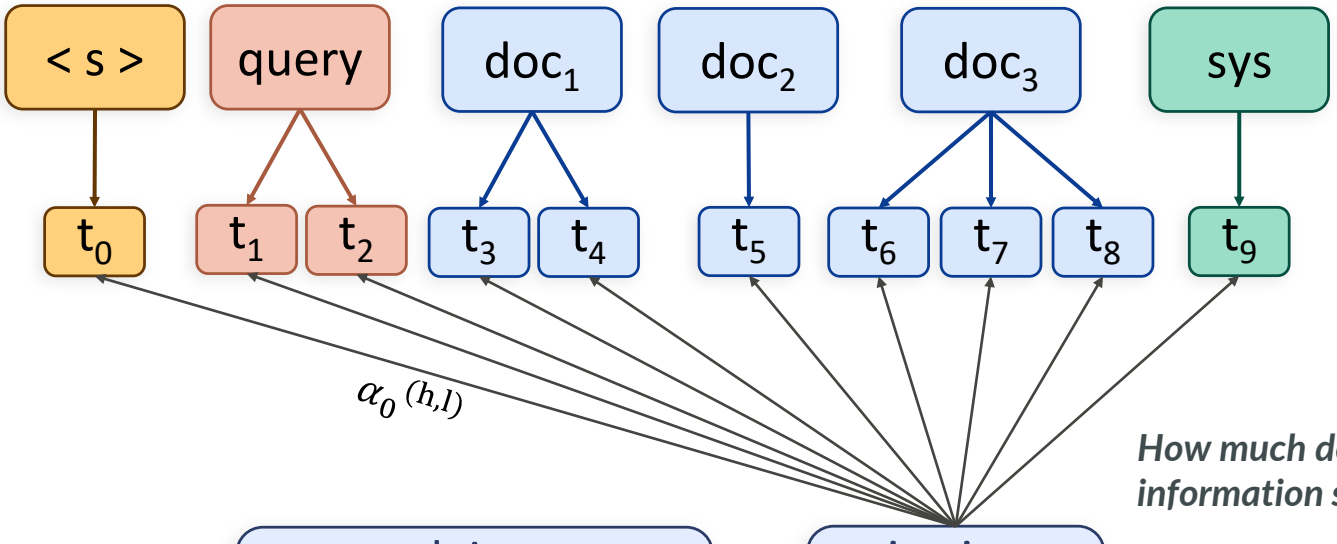
Scale-dependent

Sink rate grows with model scale

Load-bearing

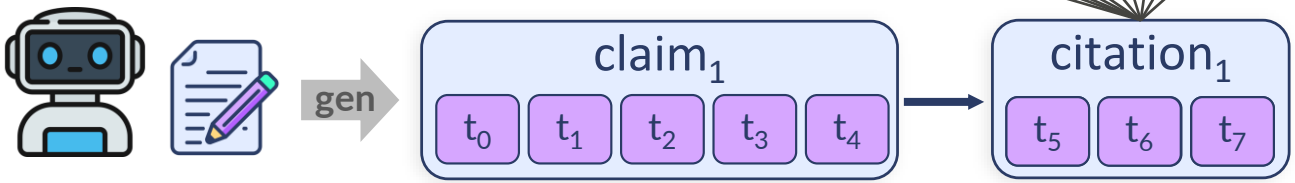
- Over-squashing
- Representational collapse

Attention sink usage



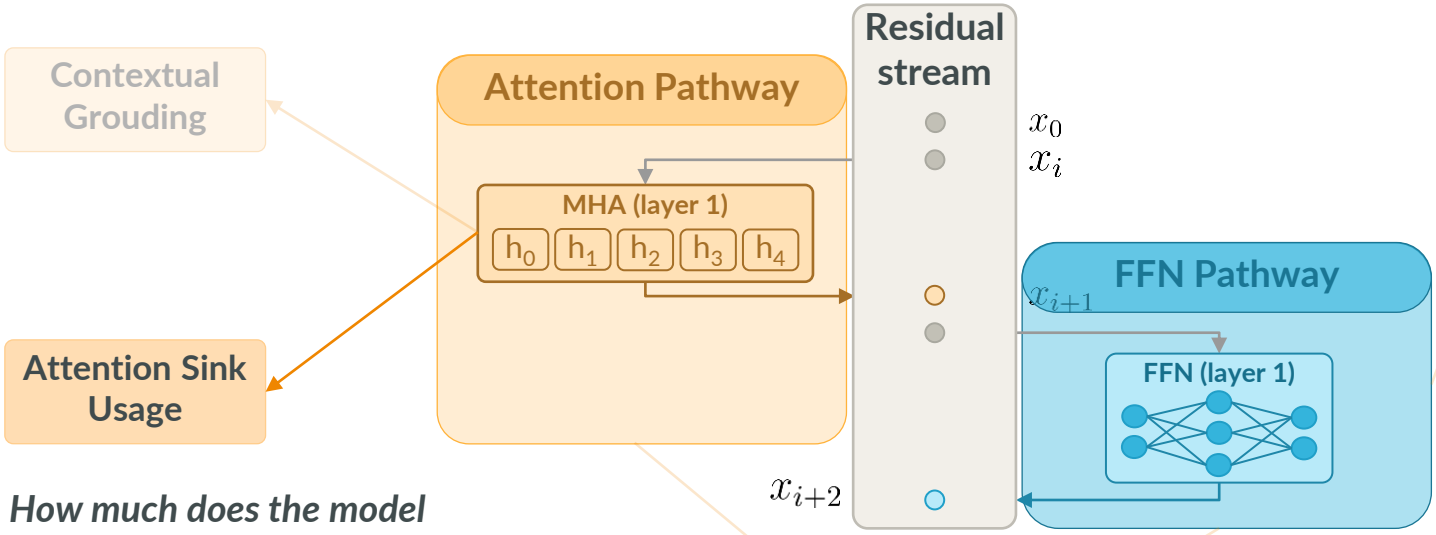
input

How much does the model use the information synthesis mechanism?



How grounded is the citation in retrieved evidence?

How forcefully does the model apply parametric knowledge?



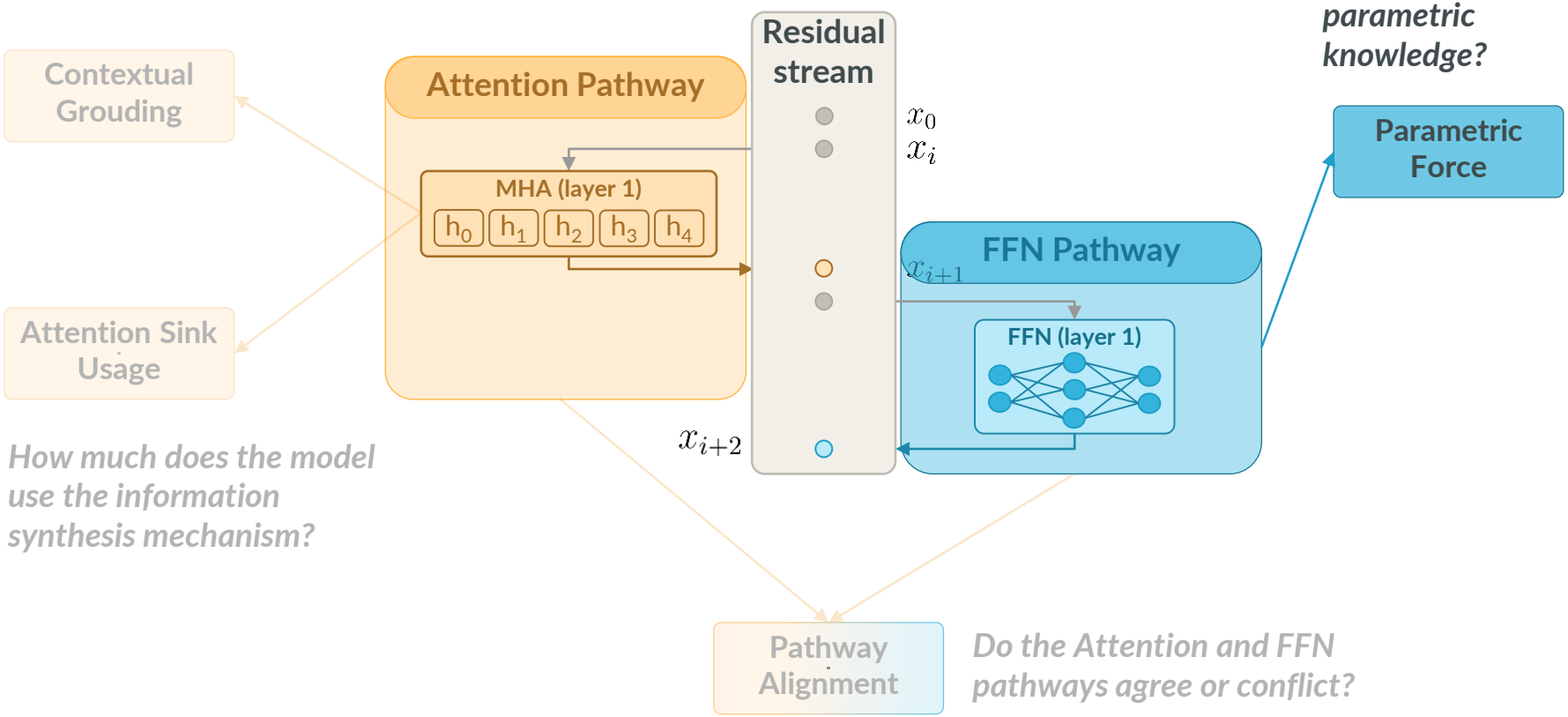
How much does the model use the information synthesis mechanism?

Pathway Alignment

Do the Attention and FFN pathways agree or conflict?

How grounded is the citation in retrieved evidence?

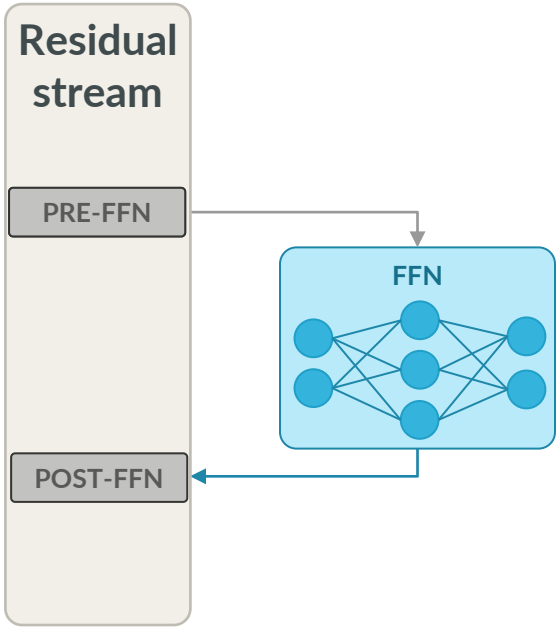
How forcefully does the model apply parametric knowledge?



How much does the model use the information synthesis mechanism?

Do the Attention and FFN pathways agree or conflict?

Parametric force



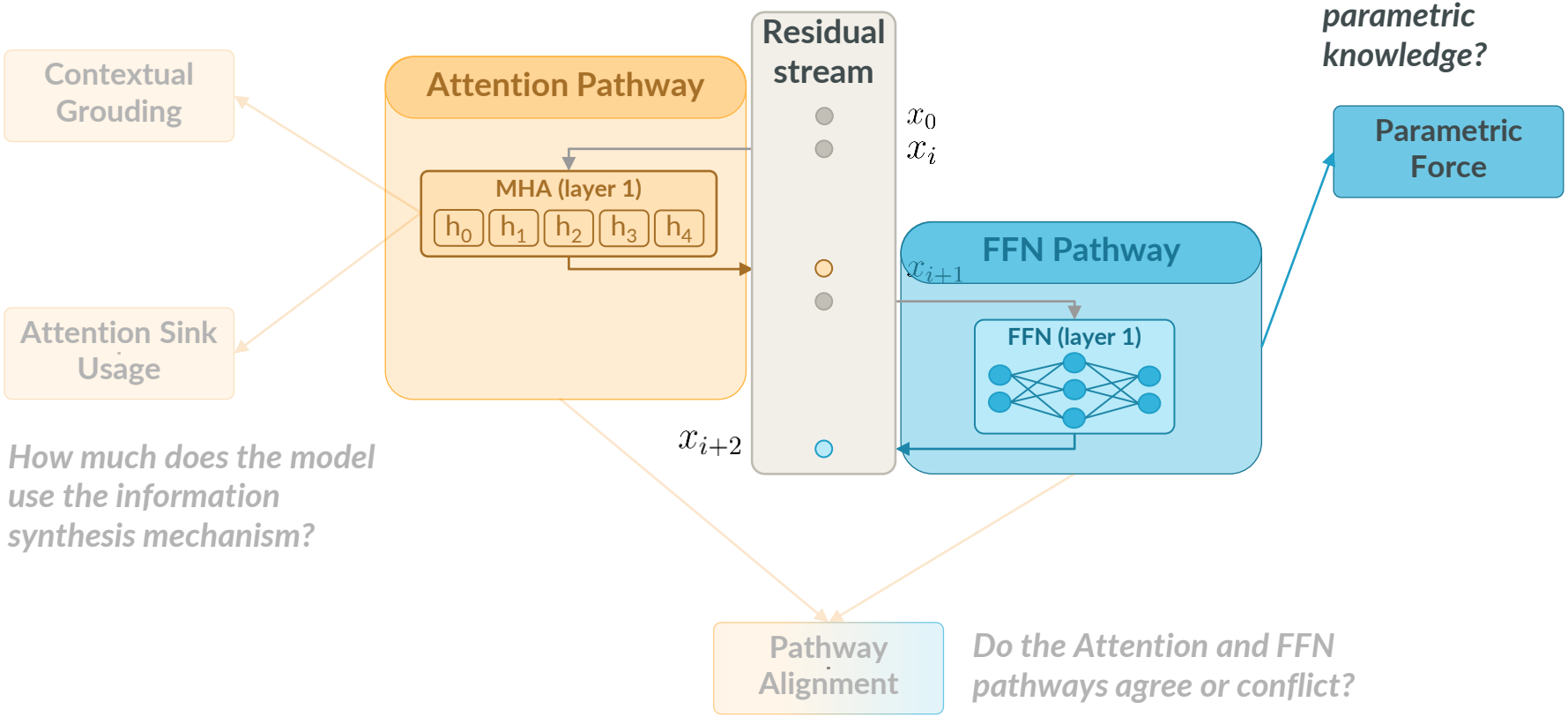
-

$||v_{ffn}||$

Parametric Force

How grounded is the citation in retrieved evidence?

How forcefully does the model apply parametric knowledge?

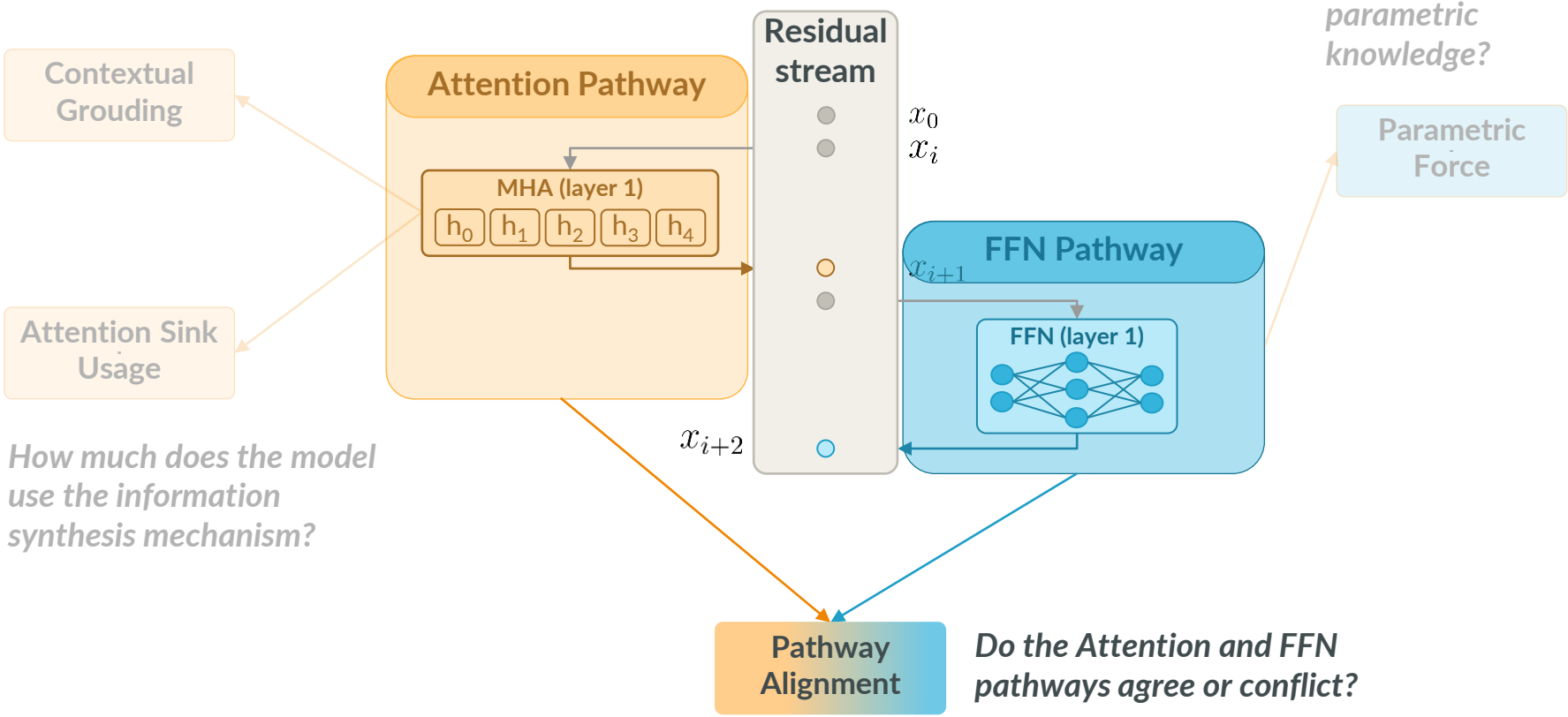


How much does the model use the information synthesis mechanism?

Do the Attention and FFN pathways agree or conflict?

How grounded is the citation in retrieved evidence?

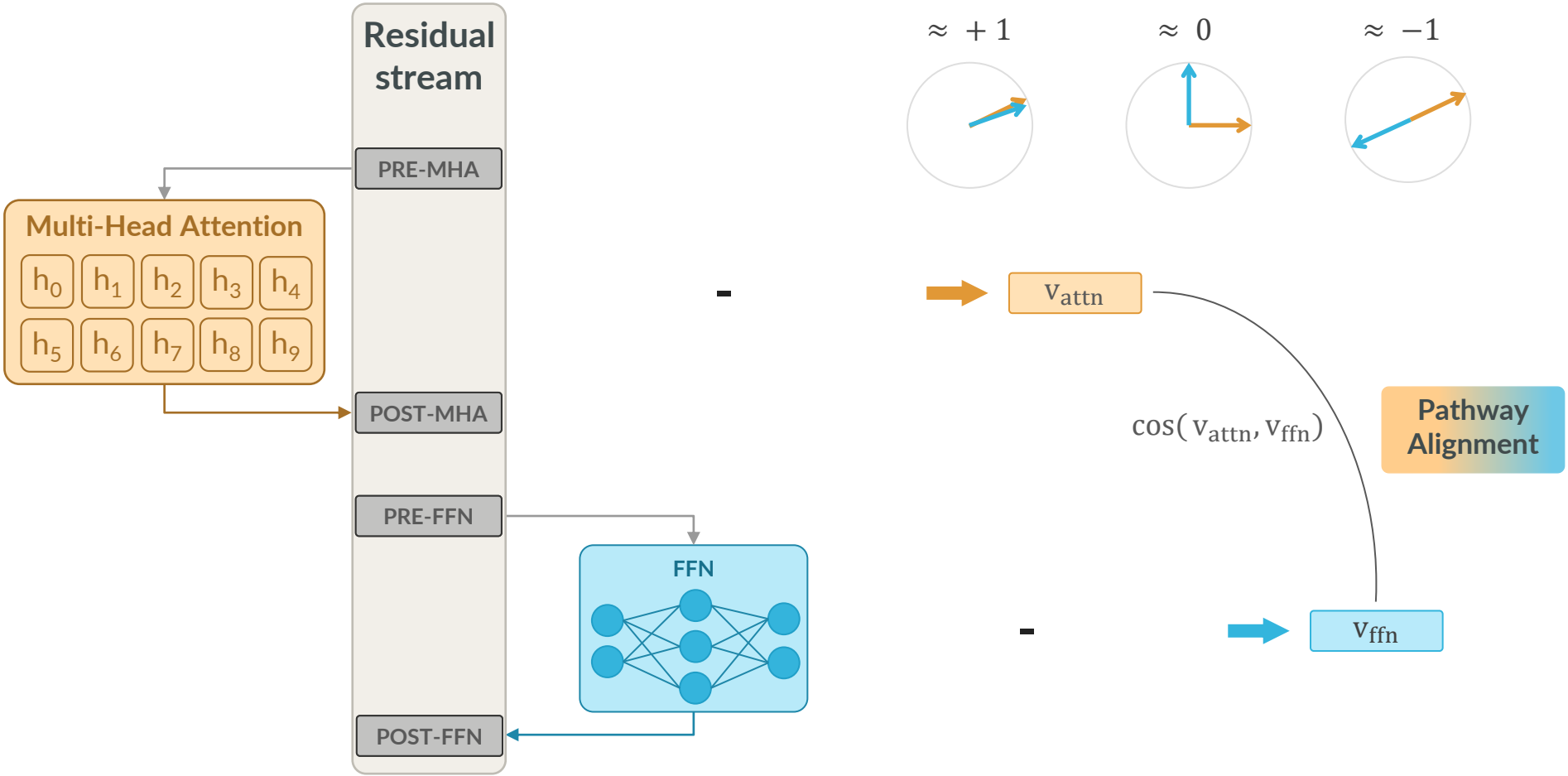
How forcefully does the model apply parametric knowledge?



How much does the model use the information synthesis mechanism?

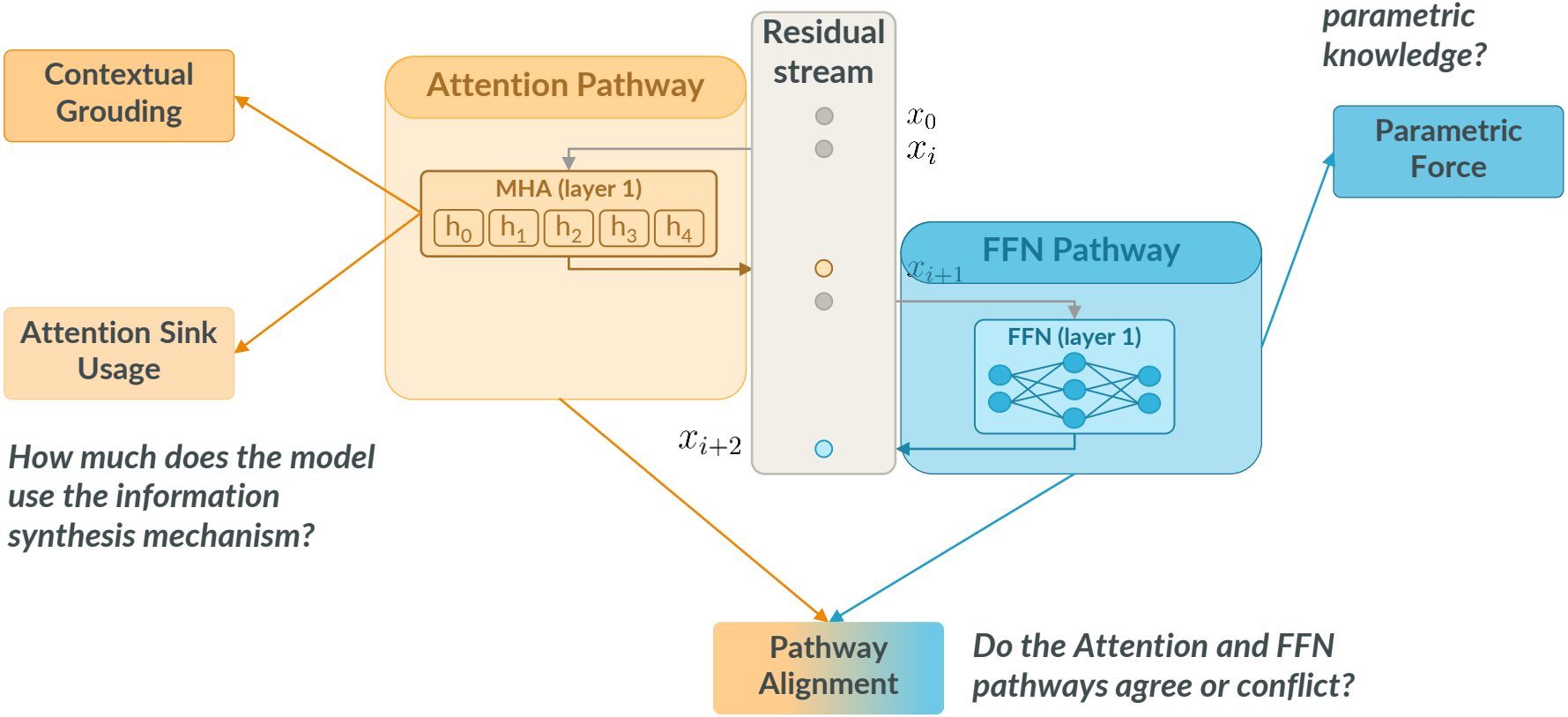
Do the Attention and FFN pathways agree or conflict?

Pathway alignment



How grounded is the citation in retrieved evidence?

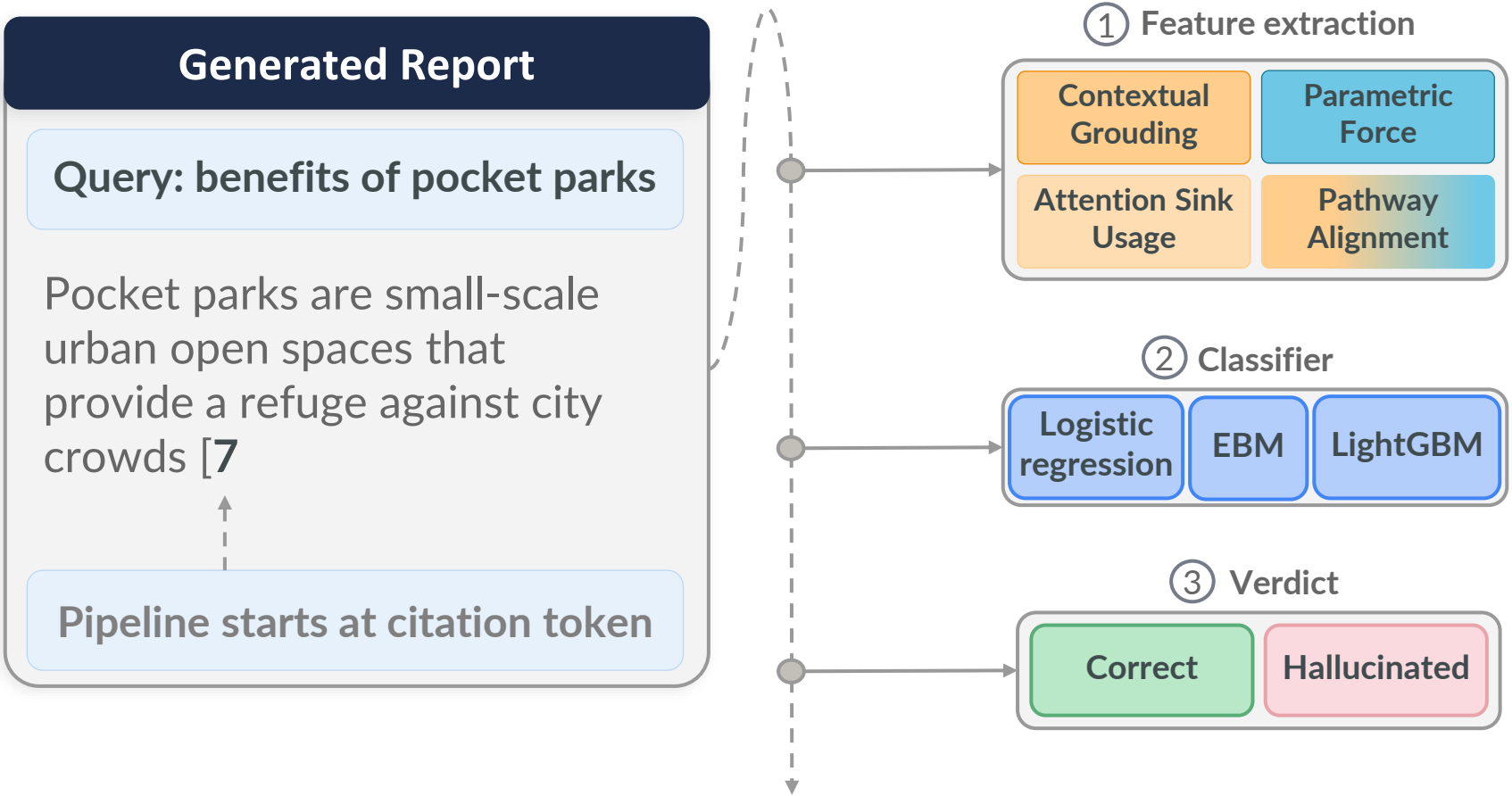
How forcefully does the model apply parametric knowledge?



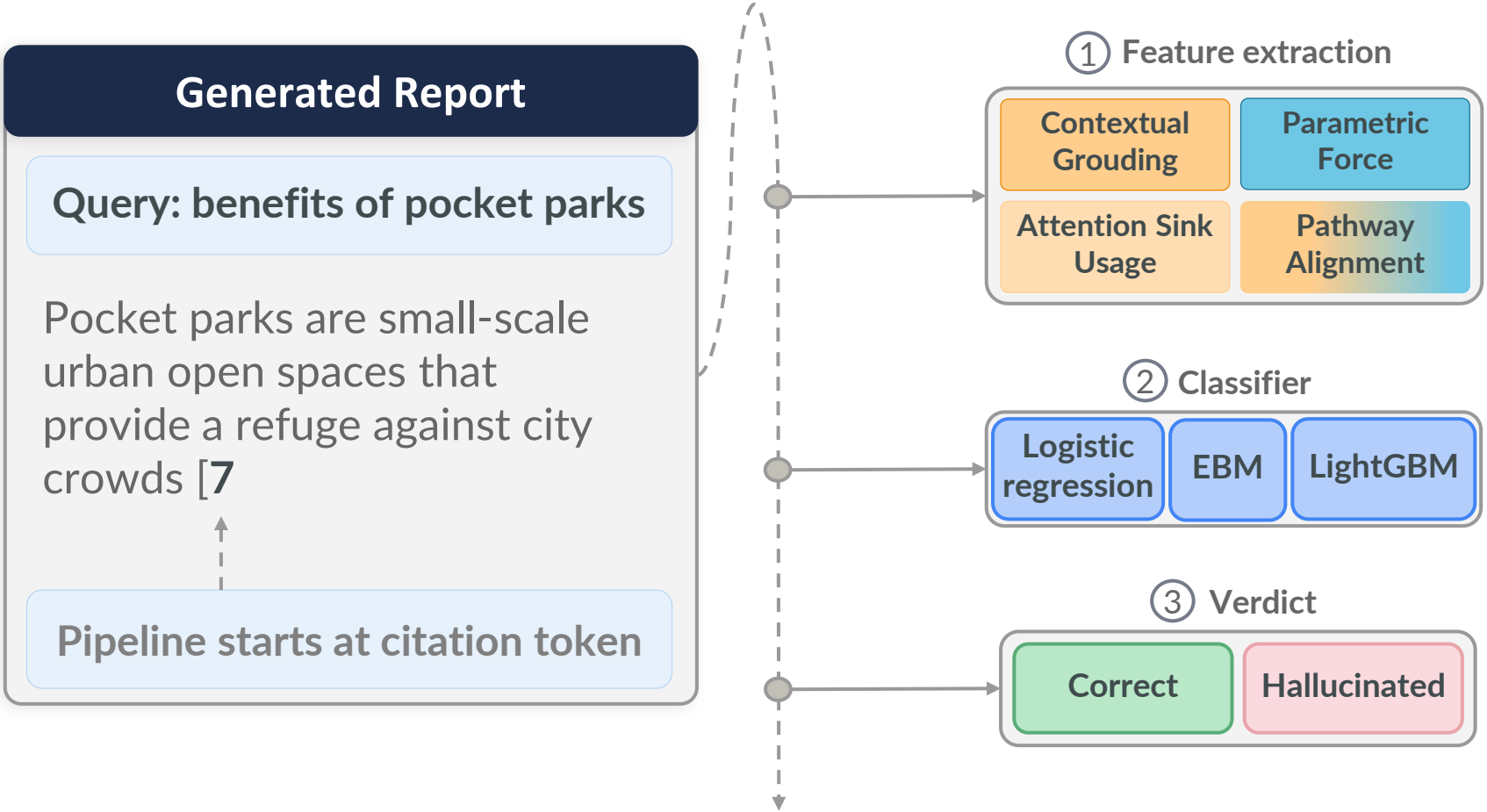
How much does the model use the information synthesis mechanism?

Do the Attention and FFN pathways agree or conflict?

From signals to prediction



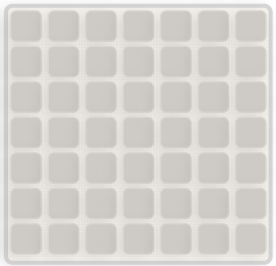
From signals to prediction



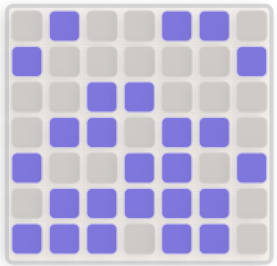
From signals to features

Attention-based

Raw scores
heads x layers



Prune
Keep top-k%



μ across heads
one value per layer



across layers
slope or FFT



One feature per score
4 features per citation token

- Contextual grouding
- Attention sink

Layer-based

one per layer



No head step needed



- Parametric Force
- Pathway Alignment

Experimental Setup

Dataset & Task

TREC NeuCLIR 2024
Long-form report generation
Up to 15 docs per query

10-fold stratified CV
Split at report level
Balanced train / imbalanced test

Labels & Validation

LLM judge (ARGUE)
Llama-3.1-70B-Instruct
~1.5% variation across temperatures

Citation Attestation

Full report eval = *too hard*
Each atomic eval = *tractable*

Break down problem, aggregate only at the end

Models

128k context window

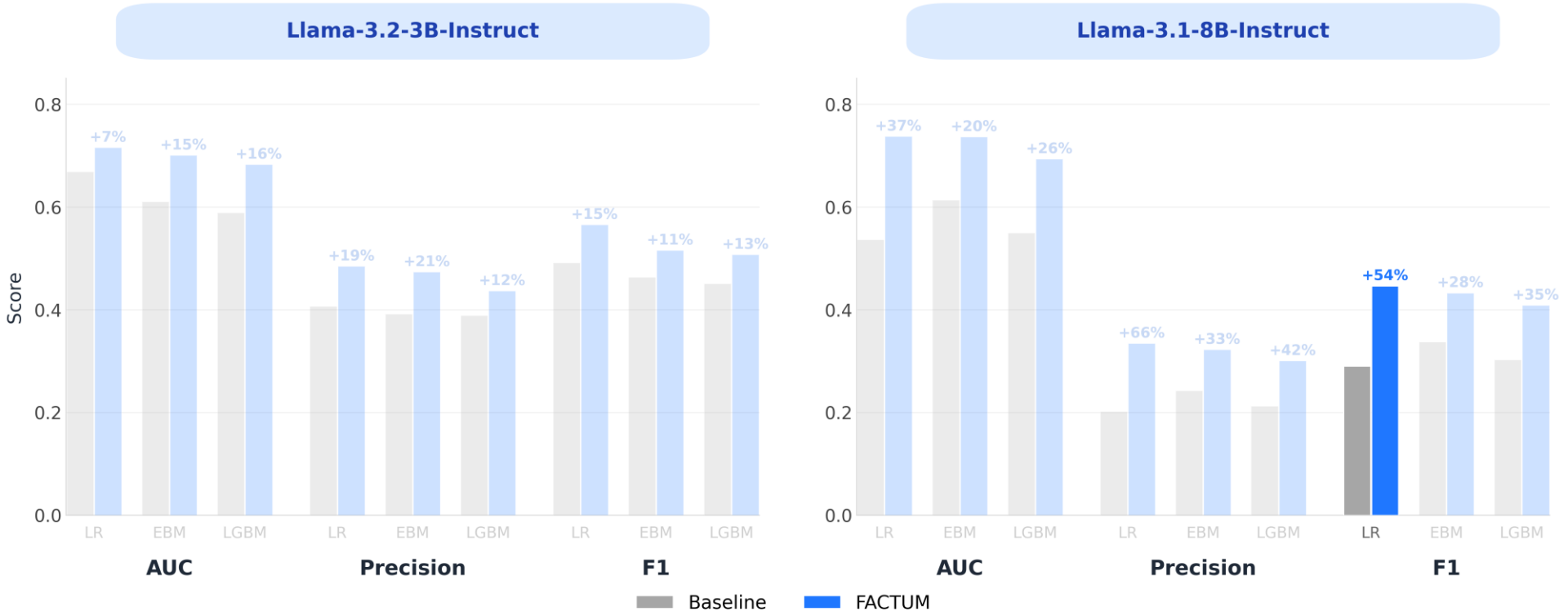
Llama-3.2-3B-Instruct



Llama-3.1-8B-Instruct

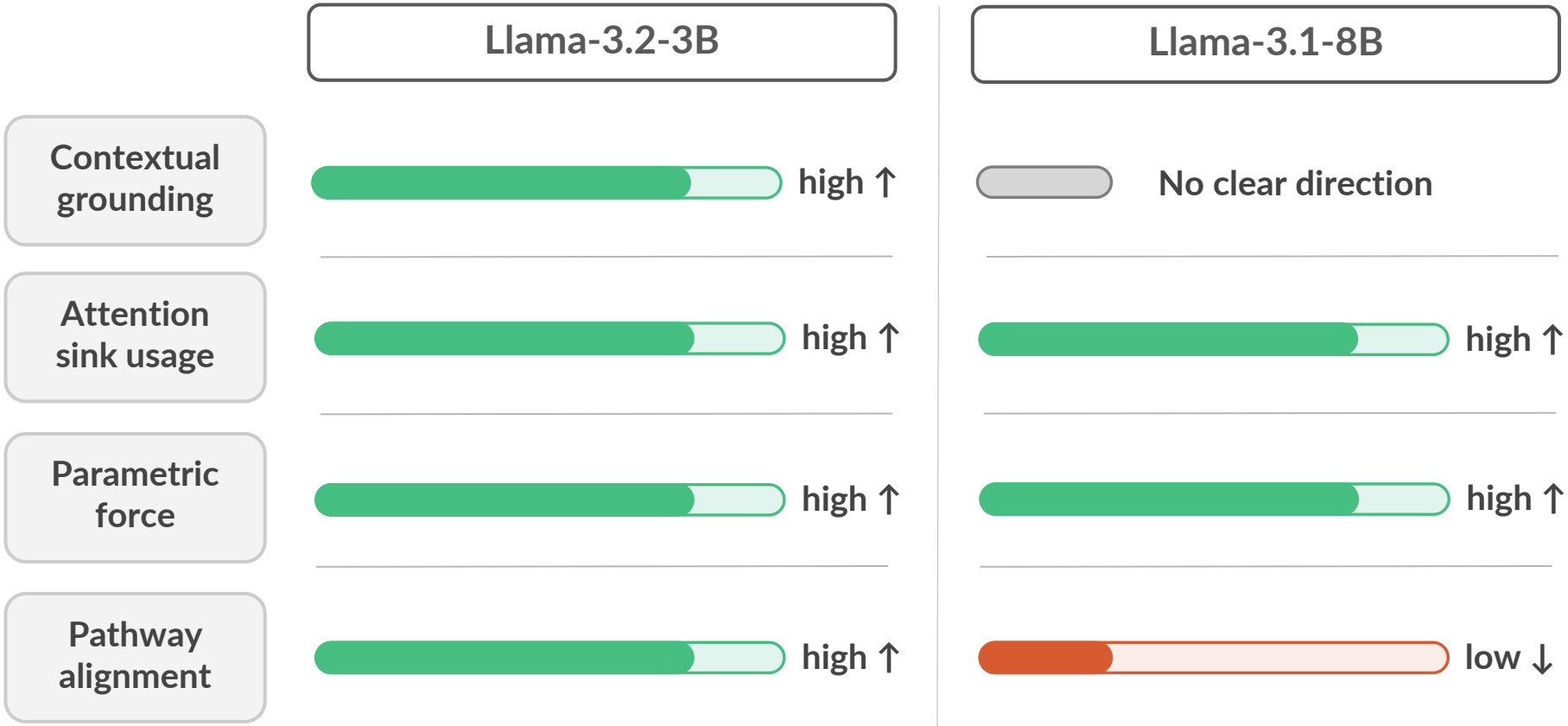


Results


Llama-3.2-3B-Instruct
Llama-3.1-8B-Instruct
Score
0.0
0.2
0.4
0.6
0.8
+7%
+15%
+16%
+19%
+21%
+12%
+15%
+11%
+13%
+37%
+20%
+26%
+66%
+33%
+42%
+54%
+28%
+35%
LR
EBM
LGBM
AUC
Precision
F1
Baseline
FACTUM

What can we do with these model internal signals?

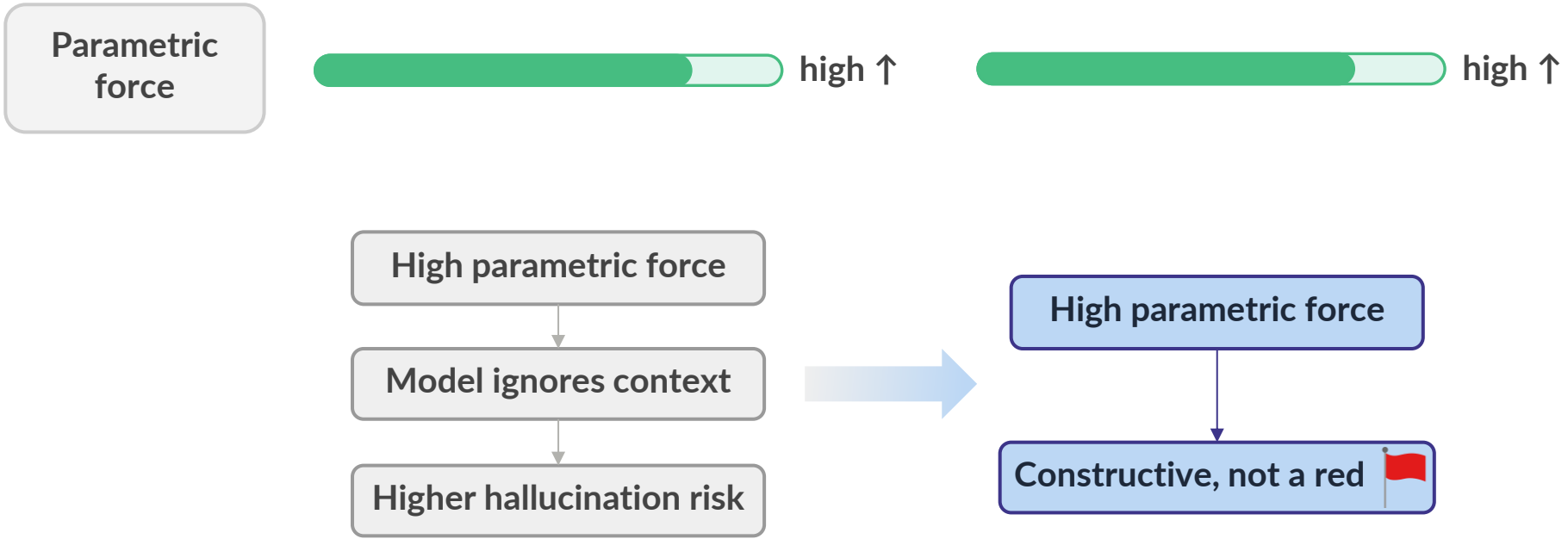
Model internal signals per model: Logistic regression



Challenging Prior Assumptions on Parametric Force



Challenging Prior Assumptions on Parametric Force



Scope & open questions

What FACTUM establishes

A new SoTA for citation hallucination detection
+37% AUC · +66% Precision

A reframing of citation hallucination
high parametric force is constructive

Mechanistic signals black-box methods 👀
same fluent text, but internals tell the difference

Reusable methodology RAG community
4 interpretable scores, applicable to any transformer

What we leave open

Generalisation across model families
Mistral, Qwen, Gemma, MoE are next

From correlation to causation
interventional approaches

From detection to prevention
measurement earlier enables steering

Cross-domain & multi-citation
medical, legal RAG · multi-source agg.

FACTUM: a step toward mechanistic, interpretable guardrails for reliable RAG systems

Beyond the Neuclir dataset

If you know why a citation went wrong, you know where to intervene

Law

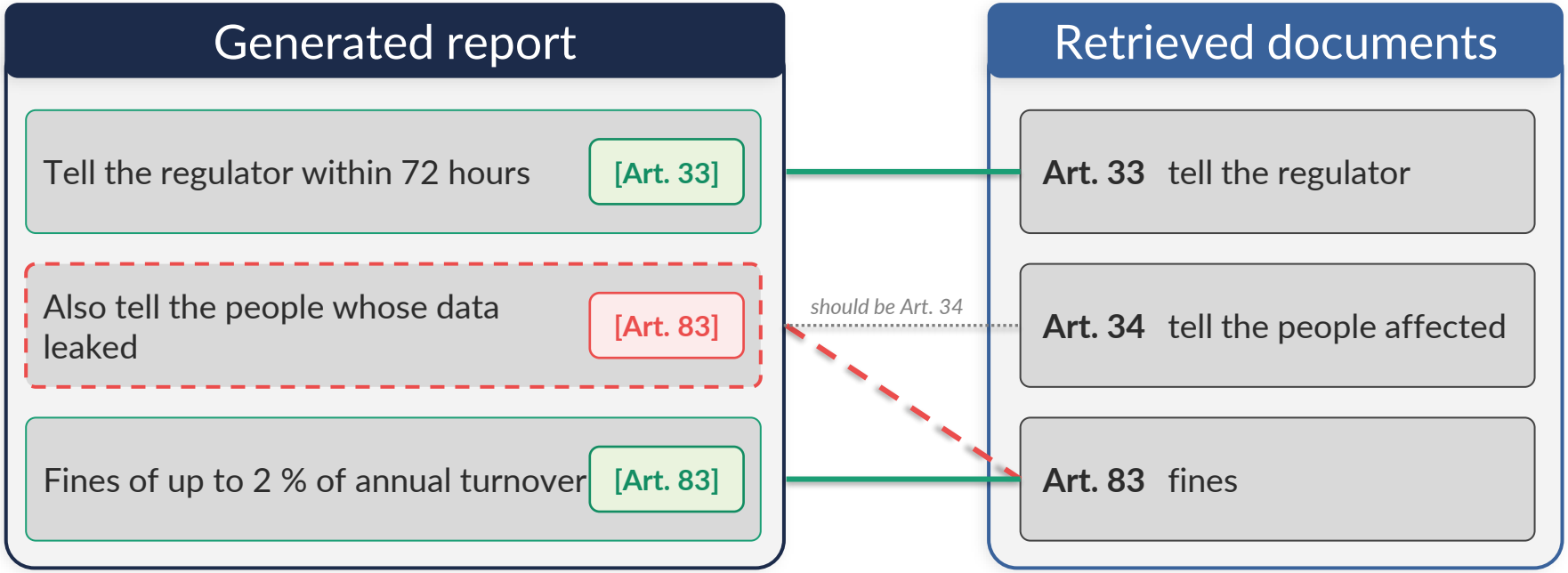
the claim must attach to exactly the right article

Medicine

the dose must come from the right guideline

The same failure in a legal setting

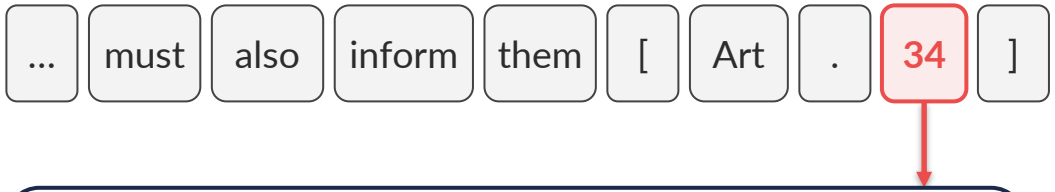
Question: *"What must a company do after a data leak?"*



The citation is the exact token where we read the model's internal state. Nothing in the method changes.

Use model internals to get insight into what went wrong

Generated token stream



Internal state at the citation token

Contextual grounding		low ↓
Attention-sink usage		low ↓
Parametric force		high ↑
Pathway alignment		low ↓

→ **flagged:** "...must also inform them [Art. 83]*"

Legal references are digits

- **Art. 6(1)(f) GDPR** *hierarchical, multi-token*
- **42 U.S.C. § 1983** *title · section*
- **ECLI:NL:HR:2024:1234** *case identifier*
- **Cass. civ. 1re, n° 22-14.345** *French pourvoi number*
- **art. 1240 C. civ. (ex-1382)** *renumbered in 2016*

Multi-token references → aggregate the token scores per reference (paper §4.1). The mechanism is unchanged.

Same dials, different settings

Law

The claim must attach to exactly the right article, literally.

Reader leads, memory only corroborates. Keep attention on the documents and protect the sink over longer texts.

Medicine

Guidelines change, a dose remembered from training may be last year's dose.

Turn memory down at citation time, the reader must dominate.

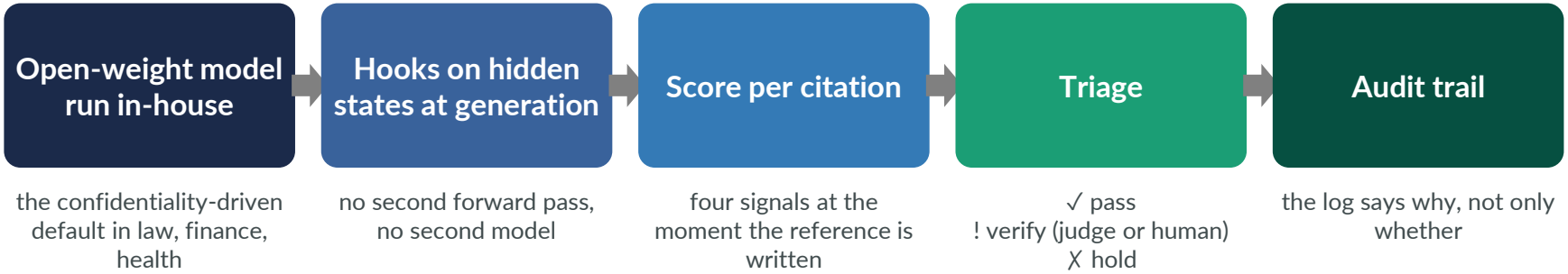
Fine-tuned in-house model

Memory is current and trustworthy, legal or clinical.

Let memory corroborate more, and use alignment to check that memory and text still agree.

The domain decides how much of each signal you want.

How an organisation would use this



Requirements

White-box access, so not API-only models

Per-model calibration, the signature changes with scale and family

A labelled set, cheap with a narrow attestation judge (κ = 0.68 vs. humans)

Why it is cheap

An LLM judge must read every source for every citation

These signals are read from what the model already computed

The expensive verifier sees only the flagged minority, and gets a reason with it

Beyond detection

Narrow the judge

open-ended evaluation is unreliable

→ *decompose into the most verifiable atomic question*

Scale matters

Mechanisms specialize as model size grows

→ *Citation signature changes with scale*

Combine toolkits

Black-box and mechanistic solve different problems

→ *Mechanistic explains and enables repair*

Thank you!

m.s.w.dassen@uva.nl

Beyond detection

Narrow the judge

open-ended evaluation is unreliable

→ *decompose into the most verifiable atomic question*

Scale matters

Mechanisms specialize as model size grows

→ *Citation signature changes with scale*

Combine toolkits

Black-box and mechanistic solve different problems

→ *Mechanistic explains and enables repair*

Questions?

m.s.w.dassen@uva.nl