How Few-Shot Examples Add Up
A Causal Decomposition of Function Vectors in In-Context Learning

Entang Wang[1] Yiwei Wang[1] Aleksandra Bakalova[1] Michael Hahn[1]

[1]Saarland Informatics Campus, Saarland University

June 14, 2026

Why Study This?

- In-context learning (ICL) lets LLMs infer a task from a few demonstrations without gradient updates.
- Mechanistic interpretability has identified function vectors (FVs) as causal activation directions that can reinstate task behavior.
- But most prior FV work is **dataset-level**: it averages over many prompts.
- This paper asks a more local question:
  - How does a **single few-shot prompt** form its own task vector?
  - How do individual examples combine?
  - What exactly does **contextualization** contribute?

Core claim

Few-shot examples contribute approximately *additively* to a prompt-level task vector, while contextualization improves the vector mainly by *reweighting* examples through Query–Key routing.

Relation to Existing Mechanistic Views

View 1: Additive / superposition

- Examples contribute composable signals.
- Related to induction-style and task-superposition stories.

View 2: Retrieval / routing

- Performance depends on query–key alignment.
- Attention selectively routes information from useful examples.

This paper's synthesis

The prompt-level FV is approximately an **additive composition** of per-example sub-FVs, and contextualization improves it mainly by **changing the routing weights** over those sub-FVs.

Main Contributions

1. Show that an n-shot prompt FV is well approximated by a linear superposition of per-example sub-FVs.
2. Introduce an **uncontextualized** counterfactual to isolate cross-example interactions.
3. Show that ambiguous vs. unambiguous examples induce different FV-head attention patterns.
4. Causally decompose contextualization into Query–Key (QK) and Value (V) pathways.
5. Show that, in most ambiguous settings, QK alignment is primarily **example-led** rather than query-led.

High-level takeaway

ICL here is best understood as **additive task-vector formation + context-dependent attention reweighting**.

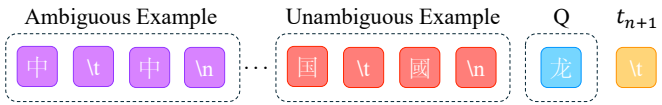
Task Setup and Prompt Format

A task t defines a mapping $g_t : \mathcal{X} \to \mathcal{Y}$.
An n-shot prompt is

$$p_n^{(t)} = \big((x_1, y_1), \ldots, (x_n, y_n), x_{n+1}\big), \qquad y_i = g_t(x_i).$$

In the paper, prompts are formatted as:

$$\underbrace{x_1 \texttt{ \textbackslash t } y_1 \texttt{ \textbackslash n}}_{Ex_1} \cdots \underbrace{x_n \texttt{ \textbackslash t } y_n \texttt{ \textbackslash n}}_{Ex_n} \underbrace{x_{n+1}}_{Q} \underbrace{\texttt{\textbackslash t}}_{t_{n+1}}$$



- The final separator token t_{n+1} is the main readout position.
- The paper studies both normal tasks and deliberately constructed ambiguous tasks.

Prompt-Specific Function Vectors

For a prompt $p_n^{(t)}$, define the prompt-specific FV as the sum of activations from localized FV heads at the final separator token:

$$v_{\mathrm{FV}}(p_n^{(t)}) = \sum_{a \in \mathcal{A}_{\mathrm{FV}}} a\big(p_n^{(t)}, t_{n+1}\big).$$

- FV heads are identified using causal mediation, following Todd et al. and related work.
- Unlike prior work, this paper does **not** average across prompts first.
- That allows prompt-level analyses of composition, attention allocation, and contextualization.

Interpretation

The FV is treated as the prompt's compact, causal representation of the inferred task.

How FV Quality Is Evaluated

To test whether a prompt FV captures the task, the paper injects it into a 0-shot prompt:

$$h_{n+1}^{(\ell)} \leftarrow h_{n+1}^{(\ell)} + \alpha\, v_{\mathrm{FV}}(p_n^{(t)}).$$

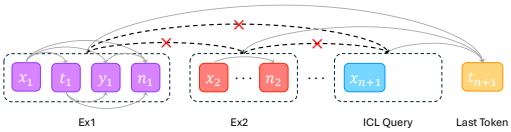
Then measure next-token accuracy on 0-shot queries. The scalar FV quality score is

$$\mathrm{Acc}_{\max}(v_{\mathrm{FV}}) = \max_{\ell \le L',\, \alpha \in \mathcal{A}} \mathrm{Acc}^{(\ell,\alpha)}(v_{\mathrm{FV}}).$$

- Sweep over injection layers and scaling factors.
- Higher $\mathrm{Acc}_{\max}$ means better causal reinstatement of the task.
- This becomes the main quality metric throughout the paper.

Contextualized vs. Uncontextualized Models

- **contextualized**: original model with full cross-component interactions.
- **uncontextualized**: preserve attention *within* each prompt component and into the final token, but remove attention *between* different components.



Why use **uncontextualized**?

It isolates the causal role of **cross-example interactions**. Local example encoding and final-token aggregation stay intact.

Finding 1: Linear Superposition Setup

For each example i, extract a sub-FV v_i by masking attention so the final token can only attend to:

- the i-th example,
- and the query token x_{n+1}.

Then fit the full FV using shared weights across prompts:

$$v_{\mathrm{FV}} \approx \sum_{i=1}^{n} w_i v_i + \varepsilon.$$

- The weights w_i are fit **globally**, not per prompt.
- This makes the test stricter than a trivial per-instance decomposition.
- The appendix also includes null baselines to show the result is non-trivial.

Finding 1: Linear Superposition Results

Results show that OLS reconstructions closely match observed FVs in both **contextualized** and **uncontextualized** settings.

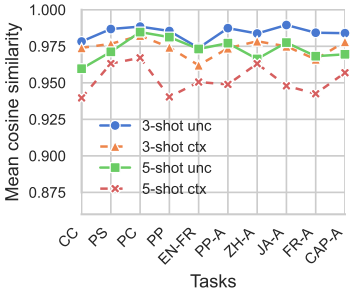
Reported summary:

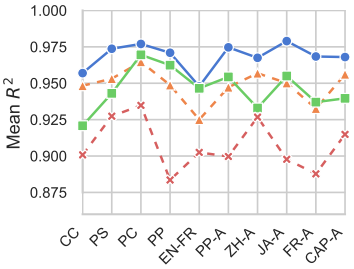
- mean cosine similarity ≥ 0.925,
- mean $R^2 \geq 0.875$.
- causal accuracy recovery $\mathrm{Acc} \geq 97\%$

Interpretation

Example-level task signals combine approximately **linearly** into a prompt-level task vector.

- Additivity survives contextualization.
- But contextualization can still change the *weights* assigned to different examples.





Ambiguous Prompt Design

The paper constructs prompts containing:

- **unambiguous** examples that uniquely identify the mapping (buy -> bought)
- **ambiguous** examples consistent with multiple mappings (let -> let, put -> put)

Design choices:

- ambiguous:unambiguous ratio fixed at $2:1$,
- multiple positional layouts to control recency bias,
- tasks include PP-A, ZH-A, JA-A, FR-A, and CAP-A.

Purpose

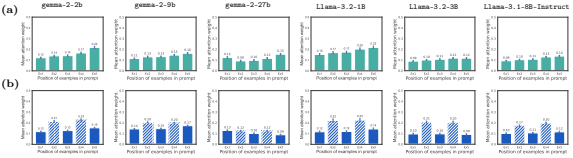
This creates a controlled conflict regime where example selection should matter most.

Finding 2: Unambiguous Examples Are Intrinsically More Attractive

In the **uncontextualized** model, examples cannot contextualize one another. So attention differences reflect intrinsic Query–Key compatibility.

Observations:

- On normal tasks, FV-head attention is largely recency-driven.
- On ambiguous tasks, FV heads systematically assign more attention to unambiguous examples.



Key patching result

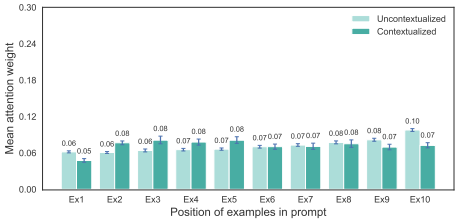
Swapping ambiguous and unambiguous Keys changes the attention pattern accordingly, confirming that example-side Keys causally drive the preference.

Finding 3a: What Contextualization Does on Normal Tasks

On normal tasks, contextualization mostly acts as **positional rebalancing**:

- it weakens strong recency bias,
- it shifts some attention mass toward earlier examples,
- but keeps attention relatively smooth.

The paper reports that normalized attention entropy remains high, indicating flattening rather than sharp selection.



Interpretation

In easy, non-conflicting settings, contextualization mainly redistributes mass, rather than choosing a special subset of examples.

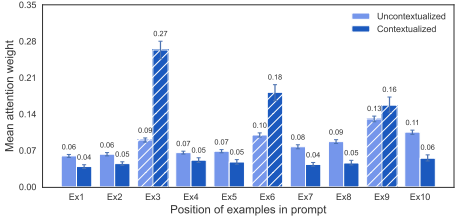
Finding 3b: What Contextualization Does on Ambiguous Tasks

Ambiguous tasks behave differently. Contextualization:

- reduces attention entropy,
- increases concentration on unambiguous examples,
- sharpens QK alignment in a semantically targeted way.

Example from the paper:

- in a 10-shot ambiguous setting, unambiguous examples receive **32%** of FV attention mass in **uncontextualized**,
- and **61%** after contextualization.



Interpretation

Under ambiguity, contextualization becomes a **selection mechanism** that resolves information conflict.

Finding 4: Factorial Decomposition of QK vs. V

The paper treats FV quality as a function of contextualization state in two pathways:

$$F(QK, V), \qquad QK, V \in \{\mathrm{unc}, \mathrm{ctx}\}.$$

This gives four settings:

1. $F(0,0)$: fully **uncontextualized**,
2. $F(0,1)$: uncontextualized QK + contextualized V,
3. $F(1,0)$: contextualized QK + uncontextualized V,
4. $F(1,1)$: fully **contextualized**.

Then use a symmetric Shapley decomposition:

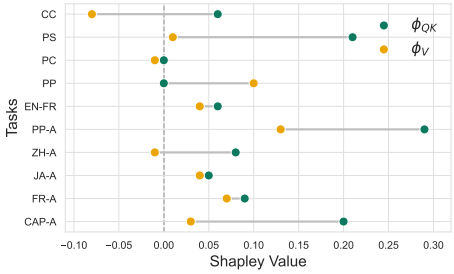
$$\phi_{\mathrm{QK}} = \tfrac{1}{2}\big(F(1,0) - F(0,0)\big) + \tfrac{1}{2}\big(F(1,1) - F(0,1)\big),$$

$$\phi_{\mathrm{V}} = \tfrac{1}{2}\big(F(0,1) - F(0,0)\big) + \tfrac{1}{2}\big(F(1,1) - F(1,0)\big).$$

Finding 4: QK Usually Matters More Than V

Across Gemma and Llama families:

- ϕ_{QK} is usually larger and more consistently positive than ϕ_{V}.
- On ambiguous tasks, QK is active in the vast majority of configurations.
- V effects are heterogeneous: often complementary, sometimes weak, occasionally detrimental.



Interpretation

Contextualization improves FV quality mainly by changing **where attention goes**, more than by rewriting the content of Value vectors.

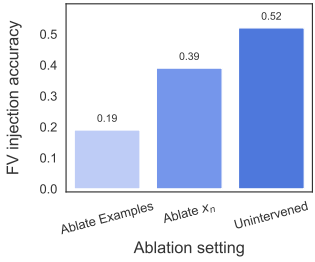
Finding 5: Where Does the QK Signal Come From?

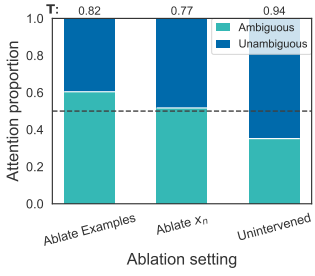
Q patching compares two corruption types:

- **examples-only corruption**: keep query, replace demonstrations,
- x_{n+1}**-only corruption**: keep demonstrations, replace query.

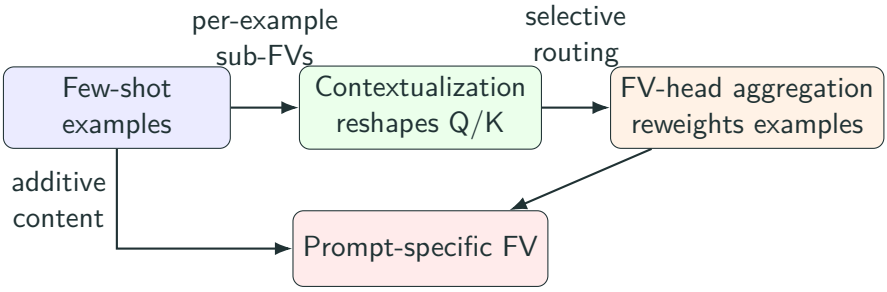
Main result:

- On most ambiguous tasks, routing is predominantly **example-led**.
- Exception: capitalization-ambiguous tasks are more query-led.





Unified Mechanistic Picture



One-sentence summary

ICL in this setting looks like **linear composition of example-level task signals**, refined by **context-sensitive routing** that preferentially selects informative examples.

Discussion and Limitations

Implications

- Bridges additive and retrieval-style mechanistic accounts.
- Suggests some theory models overemphasize query-to-example similarity.
- Offers a natural story for saturation and robustness phenomena.

Limitations

- Focused on short-horizon discrete mapping tasks.
- Value-pathway effects remain less well understood.
- Ambiguity is controlled experimentally, not formalized by a single scalar metric.

Future directions

- richer reasoning tasks,
- longer-horizon autoregressive settings,
- tighter integration with theoretical ICL models,
- formation-execution-readout circuit analysis beyond FV heads.

Thank you!