A Refutation of "On the Power of Foundation Models"

Anonymous Authors

March 2026

Abstract

We analyze the mathematical core of "On the Power of Foundation Models" [Yuan, 2023] and show that its central claims do not follow from its own definitions. The paper's first **fatal** step is its "induced category" construction. That category is not determined by the learned encoder $f : D \to \mathcal{H}$; it is determined by an auxiliary rule k_f that may reconstruct hom-sets post hoc and is even permitted to call f^{-1}, although f need not be injective. In the language-model example, the situation is worse: the construction also relies on decoding machinery that the paper had explicitly removed from the definition of the foundation model.

The latter theorems then trade on a representation theorem that is never proved. The paper silently replaces the learned feature map by the Yoneda embedding $h_{\mathcal{C}} : \mathcal{C} \to \mathcal{C}^{\wedge}$, reverses variance in a key displayed formula, and writes several expressions that are simply ill-typed. Its prompting theorem is false as stated: under the paper's own notion of "solving a task", representability is not necessary, because a non-representable presheaf can be pointwise isomorphic to a representable one. Its downstream-task semantics are also too weak to model semantic correctness, since correctness is defined only up to isomorphism in **Set** and therefore forgets labels, masks, and structured outputs. The fine-tuning theorem compares a different task class and reduces to the ordinary Kan extension of a functor along Yoneda. The multimodal theorem is itself ill-typed, assumes the crucial alignment conclusion in its hypothesis, and still does not yield a realization theorem for generation.

After these unsupported modeling steps are removed, what remains are standard categorical facts about representable presheaves, Kan extension, and full embeddings. Those facts are correct. **They are not a theory of the power of foundation models.**

1 Introduction

The paper under discussion [Yuan, 2023] advances a strong thesis: once pretraining is idealized, category theory can characterize what a foundation model can and cannot do under prompting, fine-tuning, and multimodal transfer. The intended program has five steps:

1. define a category "induced" by a pretrained model;
2. identify the learned representation with the Yoneda embedding of that category;
3. characterize prompting by representability;
4. characterize fine-tuning by extension along Yoneda;
5. transfer structure across modalities through feature-aligned functors.

The categorical facts in steps 3–5 are not controversial. The failure is at step 1 and, even more decisively, at the unproved passage from learned features to Yoneda coordinates in step 2. The paper never establishes that a trained encoder computes, or even faithfully represents, the presheaves required by the later arguments. Once this missing bridge is removed, the central theorems cease to be theorems about learned representations and collapse into ordinary statements about an abstract category $\mathcal{C}$.

Our objections are mathematical, not stylistic. We do not dispute the Yoneda lemma, Kan extension, or the behavior of full embeddings. We dispute the paper's attempt to derive claims about foundation models from those facts. The induced-category construction is non-canonical. The language-model example does

not instantiate the stated definition. The downstream-task semantics collapse correctness to cardinality. The prompting theorem is false under the paper's own definitions. Several displayed formulas are ill-typed, and the multimodal section assumes the key alignment statement rather than proving it.

2 The induced-category construction is non-canonical and does not justify the learning claims

The paper defines a "foundation-model-induced category" from a feature map $f : D \to \mathcal{H}$ together with an auxiliary function

$$k_f : \mathcal{H} \times \mathcal{H} \to \mathbf{Set},$$

where k_f is declared "data-oblivious" but is explicitly allowed to call both f and f^{-1} as oracles. This does not do the conceptual work the paper assigns to it.

2.1 The category is not induced by the encoder alone

Even if one grants the definition, the induced category is not determined by f. It is determined by the pair (f, k_f), and k_f is largely unconstrained. If f is injective and $\mathcal{C}$ is any small category on the data domain, one may define

$$k_f(f(X), f(Y)) := \mathrm{Hom}_{\mathcal{C}}(X, Y)$$

on the image of f, extend k_f arbitrarily away from that image, and inherit composition and identities from $\mathcal{C}$. The same encoder f has now been made to "induce" the category $\mathcal{C}$. Since $\mathcal{C}$ was arbitrary, the paper has not extracted a canonical category from the model or from the pretext task. It has postulated one through k_f.

This defect is not cosmetic. The paper repeatedly speaks as though the pretraining objective determines a structural category that captures what the model has learned. The definition proves nothing of the sort. At most, it shows that one can retrofit some category to the feature space after the fact.

2.2 The appeal to f^{-1} is illegitimate, and the language-model example leaves the stated model class

Nothing in the framework requires f to be injective. If f is many-to-one, then f^{-1} is not a function and cannot be used as the paper proposes. Yet the language-model construction explicitly recovers raw sentences from the feature vectors in order to test prefix relations.

The problem is even sharper than that. Earlier in the paper, the "foundation model" is defined to be only the feature map, with the output head factored away as an implementation detail. But in the language-model theorem, k_f is allowed not merely to recover the original sentence but also to "run f multiple times" to compute a distribution of next sentences. That is no longer the feature map alone; it smuggles decoding machinery back into the definition after the paper had removed it. The example therefore does not instantiate the framework it is supposed to justify.

2.3 The narrative about learning an "ideal category" is content-free

The paper next argues that each intermediate encoder f_t induces a category $\mathcal{C}_{f_t}$, so training can be interpreted as learning categories, and there is an "ideal category" that best matches the pretext task. But once the induced category is non-canonical, that conclusion has no content. The paper defines no objective on categories, no notion of approximation between categories, and no criterion under which one induced category is the one "best matching" a pretext task. The existence of *some* category attached to each f_t does not show that training learns a meaningful categorical structure.

3 The paper never proves that learned features are Yoneda features

The paper repeatedly treats the following objects as though they were interchangeable:

1. a learned feature map $f : D \to \mathcal{H}$;
2. an auxiliary hom-recovery rule $k_f : \mathcal{H} \times \mathcal{H} \to \mathbf{Set}$;
3. the Yoneda embedding $h_{\mathcal{C}} : \mathcal{C} \to \mathcal{C}^{\wedge}$.

They are not interchangeable. A feature map lands in an ambient representation space. A rule k_f is extra structure supplied by the theorist. The Yoneda embedding is a canonical functor internal to a category. The induced-category definition yields, at most, hom-sets in an abstract category built from (f, k_f). It does *not* show that the feature vector $f(X)$ *is* the presheaf $h_{\mathcal{C}}(X)$ or that the model computes hom-sets in the presheaf category $\mathcal{C}^{\wedge}$.

This missing representation theorem is the central mathematical gap in the paper. Everything later depends on it. Without it, the paper is not deriving claims about a trained encoder. It is simply switching from the encoder to Yoneda by fiat.

3.1 Variance is reversed

The paper defines the Yoneda embedding by

$$h_{\mathcal{C}}(X) = \mathrm{Hom}_{\mathcal{C}}(\,\cdot\,, X),$$

a contravariant presheaf. But it later writes, in effect,

$$f(X)(Y) = k_f(f(X), f(Y)) = \mathrm{Hom}_{\mathcal{C}}(X, Y).$$

If $f(X)$ is being identified with $h_{\mathcal{C}}(X)$, then the correct evaluation is

$$h_{\mathcal{C}}(X)(Y) = \mathrm{Hom}_{\mathcal{C}}(Y, X),$$

not $\mathrm{Hom}_{\mathcal{C}}(X, Y)$. One can switch to the covariant Yoneda embedding $X \mapsto \mathrm{Hom}_{\mathcal{C}}(X, \cdot)$, but then the rest of the paper must be rewritten accordingly. As written, the variance is wrong.

3.2 Several displayed equations are ill-typed

A key displayed equation in the prompting section is

$$T(X) \simeq \mathrm{Hom}_{\mathcal{C}^{\wedge}}(h_{\mathcal{C}}(X), T) \triangleq k_f(f(X), T).$$

Under the paper's own definitions, this is ill-typed. The rule k_f takes two feature vectors in $\mathcal{H}$. But T is a presheaf in $\mathcal{C}^{\wedge}$, not a point of $\mathcal{H}$. The paper supplies no interpretation map from presheaves into feature space, yet proceeds as though one had been constructed.

The same defect reappears in the large-language-model specialization, which compares a task functor T_P directly with a feature vector $f(P)$. Again, one side is a presheaf and the other is an element of $\mathcal{H}$. That is not a theorem. It is a type error.

4 The prompting theorem is false under the paper's own task semantics

The paper defines a downstream task as a presheaf $T \in \mathcal{C}^{\wedge}$ and declares that the model "solves" T when, for every input $X \in \mathcal{C}$, the output is *isomorphic as a set* to $T(X)$. This definition is already much too weak for machine learning: in **Set**, all singleton sets are isomorphic, so a classifier that returns the wrong

label but still outputs a singleton would count as perfectly correct. For segmentation, the definition forgets which pixels are foreground and remembers only the cardinality of the selected set. The semantics erase label identity and structure.

But the same weak notion of correctness also makes the central prompting theorem false. The theorem claims that prompting solves exactly the representable tasks. That is wrong under the paper's own definition of "solve." Non-representable presheaves can still be pointwise isomorphic to representable ones.

4.1 An explicit counterexample

Let $\mathcal{C}$ be the one-object category whose endomorphism monoid is the group $C_2 = \{e, s\}$. Then $\mathcal{C}^\wedge$ is the category of right C_2-sets. The representable presheaf $h_{\mathcal{C}}(*)$ is the regular right C_2-set on two elements.

Now let T be the trivial right action of C_2 on a two-element set $\{0, 1\}$. Then T is *not* representable, because the regular action is not isomorphic, as a C_2-set, to the trivial action. Nevertheless, there is only one object in $\mathcal{C}$, so pointwise we have

$$T(*) \cong h_{\mathcal{C}}(*)(*)$$

as plain sets: both are two-element sets. Under the paper's definition of solving a task, a prompt representing $*$ therefore solves T perfectly, even though T is not representable.

This is not a corner case. It exposes the exact logical error in the paper's proof. From

$$h_{\mathcal{C}}(P) \ncong T$$

one cannot conclude that there exists some object Y with

$$h_{\mathcal{C}}(P)(Y) \ncong T(Y).$$

Failure of natural isomorphism is stronger than failure of pointwise set isomorphism. The proof confuses the two, and the theorem is false as stated.

5 The fine-tuning theorem is not a comparison with prompting; it is just Yoneda plus unlimited downstream capacity

Even setting aside the missing encoder-to-Yoneda bridge, the fine-tuning theorem does not say what the paper claims.

First, it quantifies over a different class of tasks. In the prompting theorem, tasks are presheaves

$$T \in \mathcal{C}^\wedge = [\mathcal{C}^{op}, \mathbf{Set}].$$

In the fine-tuning theorem, however, tasks are covariant functors

$$F : \mathcal{C} \to \mathbf{Set}.$$

Those are not the same objects. The paper therefore does not prove that fine-tuning can solve a larger subset of the *same* task family than prompting. It silently changes the family being quantified over.

Second, the proof is simply the standard extension of a set-valued functor along the Yoneda embedding. Abstractly, that is correct. But it becomes a theorem about learned representations only after one proves that the pretrained model has actually realized Yoneda coordinates. That step is absent. So the actual content of the theorem is far weaker than advertised: if one already had a lossless Yoneda encoding of objects and an arbitrarily expressive downstream learner with unlimited data and optimization power, then one could compute any set-valued functor of those objects. That is a statement about Yoneda, not a theory of foundation models.

6 The multimodal theorem is ill-typed, assumes the hard part, and still does not yield generation

The multimodal section contains three separate problems.

6.1 The theorem is written with a type mismatch

The theorem introduces a functor

$$\hat{F} : \mathcal{C}^\wedge \to \mathcal{B}^\wedge$$

and then concludes that

$$\mathrm{Hom}_{\mathcal{A}^\wedge}(\hat{F}(h_{\mathcal{C}}(X)), \hat{F}(h_{\mathcal{C}}(Y)))$$

recovers $\mathrm{Hom}_{\mathcal{C}}(X, Y)$. But $\hat{F}(h_{\mathcal{C}}(X))$ lies in $\mathcal{B}^\wedge$, not in $\mathcal{A}^\wedge$. The expression is therefore ill-typed unless one inserts a restriction functor along the inclusion $E : \mathcal{A} \hookrightarrow \mathcal{B}$. The proof later identifies $h_{\mathcal{A}}(FX)$ with $h_{\mathcal{B}}(FX)$ as though they were the same presheaf, which they are not. At best,

$$h_{\mathcal{A}}(FX) \cong E^* h_{\mathcal{B}}(FX),$$

where E^* denotes restriction along E.

6.2 The theorem assumes the key alignment statement in the hypothesis

The paper defines "feature-aligned" precisely by the condition

$$\hat{F}(h_{\mathcal{C}}(X)) \cong h_{\mathcal{B}}(F(X)) \qquad \text{for all } X \in \mathcal{C}.$$

But this already states the crucial object-level alignment that the theorem later advertises. Once one assumes that every representable source presheaf is sent to the correct representable target presheaf, the preservation of hom-sets is largely ordinary categorical bookkeeping. The theorem does not explain how multimodal training achieves this alignment. It assumes it.

6.3 The reconstruction step is false as stated

The paper writes expressions involving a global inverse

$$h_{\mathcal{B}}^{-1} : \mathcal{B}^\wedge \to \mathcal{B}.$$

No such inverse exists in general, because Yoneda is fully faithful but not essentially surjective: most presheaves are not representable. One can invert Yoneda only on its representable image, and even then one still needs a separate realization map from a target-category object to a concrete image, audio clip, or video. The paper does not provide that theory.

This is exactly why the CLIP application does not go through. The text simply asserts that the CLIP network is invertible and then treats alignment as generation. Neither move is justified. A dual-encoder alignment model is not a generator. Without an explicit decoder, sampler, or realization theorem, the multimodal section does not prove the ability to synthesize novel images.

7 Even the paper's own examples collapse under its definitions

The contrastive example is especially revealing. There the paper defines

$$\mathrm{Hom}_{\mathcal{C}}(X, Y) = \{f(X) - f(Y)\}.$$

Every hom-set is therefore a singleton. Up to isomorphism of categories, this is just the chaotic category on the data points: there is exactly one morphism between any ordered pair of objects. The geometric content of the feature space has disappeared into the names of singleton arrows. Under the paper's own task semantics in **Set**, representable tasks in such a category are correspondingly trivial.

This matters because the paper repeatedly presents its induced categories as rich structural objects learned by pretraining. In its own worked example, the categorical structure collapses almost completely. The later claims are then being carried not by category theory, but by informal intuitions that were never formalized.

8 What survives after the unsupported claims are removed

Once the unsupported modeling steps are stripped away, the valid residue is familiar category theory:

1. representable presheaves are precisely those naturally isomorphic to $h_{\mathcal{C}}(X)$ for some X;
2. any set-valued functor extends along the Yoneda embedding;
3. full embeddings preserve hom-sets and therefore preserve structure in the ordinary categorical sense.

Those facts are standard and correct. What is missing is exactly what would be required for a real theory of foundation models:

1. a canonical category associated to a given pretraining objective and encoder;
2. a theorem showing that the learned representation is fully faithful, Yoneda-like, or otherwise adequate for the later categorical manipulations;
3. a task semantics rich enough to encode actual labels and structured outputs rather than bare cardinalities;
4. a realization theory mapping aligned target objects to concrete generated samples.

Without these ingredients, the paper does not answer its stated question.

9 Conclusion

The paper's mathematical burden lies precisely where its arguments fail. Its induced category is non-canonical and may depend on an unavailable inverse f^{-1} and on decoding operations that the framework had excluded. Its central bridge from learned features to Yoneda coordinates is never proved. Its prompting theorem is false under its own definition of task solving. Its key equations are ill-typed and reverse variance. Its fine-tuning theorem changes the class of tasks and reduces to ordinary Yoneda extension. Its multimodal theorem is itself ill-typed, assumes the hard alignment statement in its hypothesis, and still does not furnish a realization theorem for generation.

For these reasons, *On the Power of Foundation Models* does not provide a valid mathematical theory of the power of foundation models. It repackages standard categorical facts in machine-learning language while omitting the theorem that would connect those facts to learned representations.

References

Yang Yuan. On the power of foundation models. In *International conference on machine learning*, pages 40519–40530. PMLR, 2023.