The Art and Arcana of Scientific Writing

Jason Alan Fries, PhD

BIOMEDIN 212

May 28th, 2025

Outline

- Framing Your Scientific Narrative
- Structuring Your Paper
- Building Writing Habits for Research

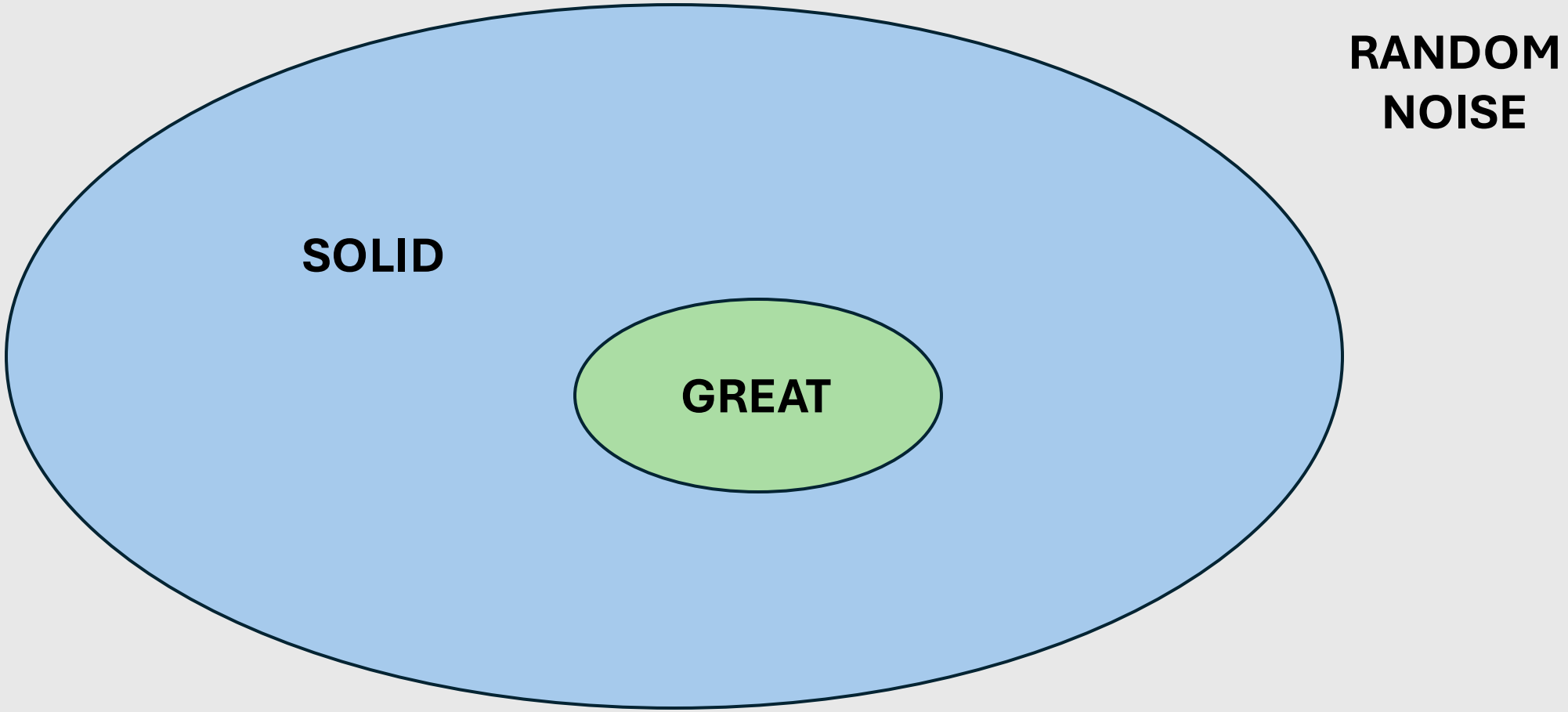
The lecture copies and builds upon lots of excellent material

- Sasha Rush *"How to write an okay research paper"*
- Jennifer Widom "*Tips for Writing Technical Papers*"
- Chris Aberger "*Check Yourself Before You Wreck Yourself*"

A Biased Preamble

- **Research = 50% Good Ideas + 50% Communication**
- Your goal: **de-risk research** and **place smart bets** on where to invest your time
- Executing with **speed, clarity, and rigor** is hard—and a skill to practice
- Success comes from committing to a **repeatable process** for research *and* communication.

Goal: Minimum Viable Paper

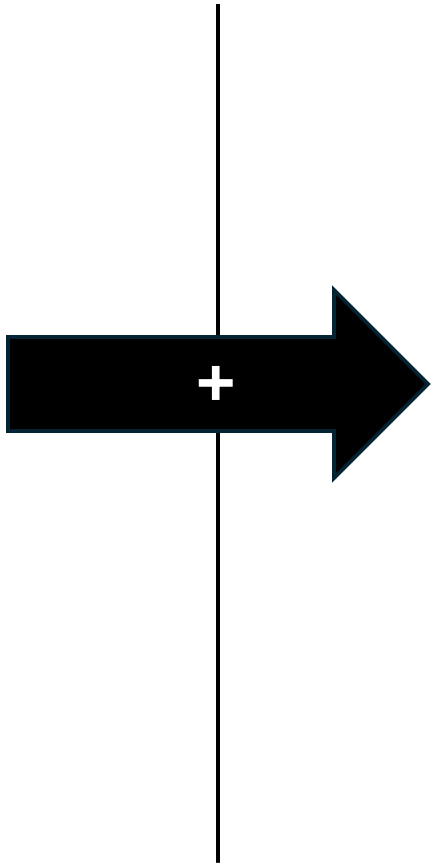


Adapted from: Sasha Rush. *"How to write an okay research paper." YouTube 2024*.

What's the Difference?

SOLID

- Structure
- Coherence
- Clarity



GREAT

- Creativity
- Rhetoric
- Empirical Precision

Adapted from: Sasha Rush. *"How to write an okay research paper." YouTube 2024*.

Framing Your Scientific Narrative

Many Types of Scientific Writing in Research

LITERATURE REVIEW / SURVEY PAPER

POSITION PAPER

GRANT PROPOSAL

DATASET PAPER

ANALYSIS PAPER

APPLICATION PAPER

....

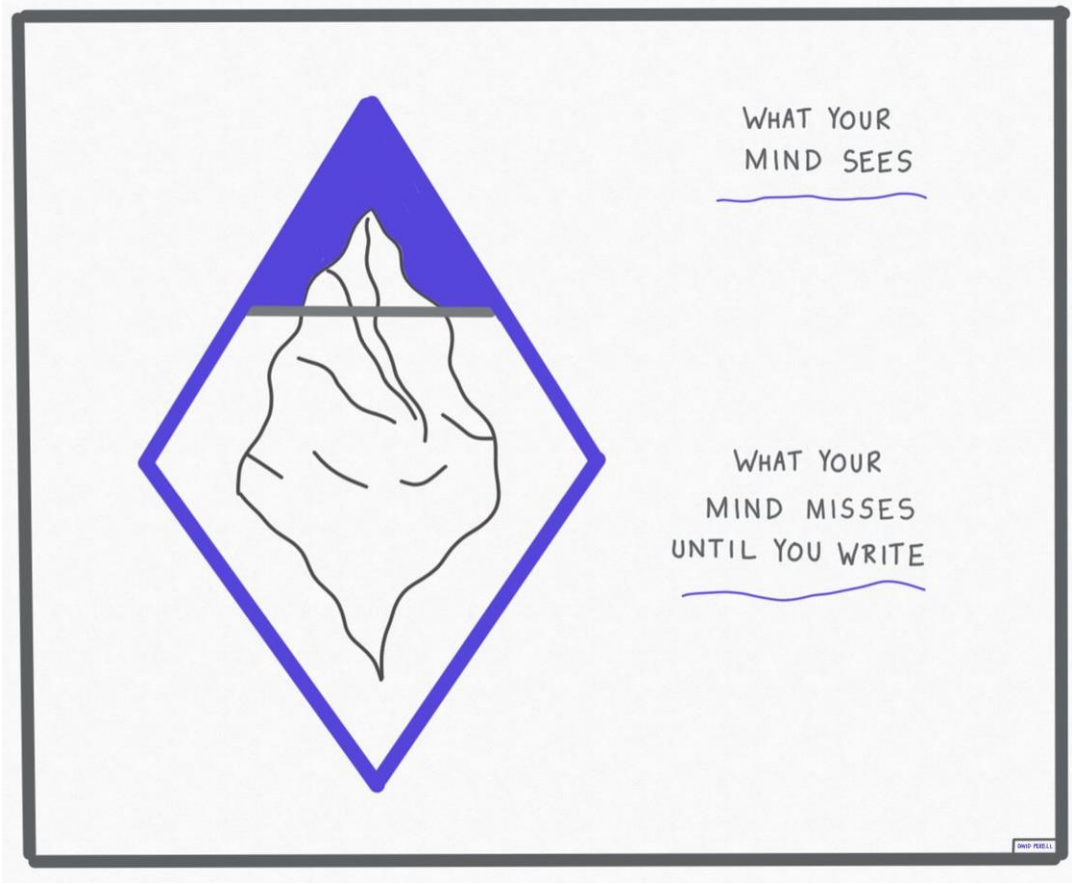
EMPIRICAL ML PAPER

Today's Focus

Who is Your **Audience**?

- Tailor your writing to your audience—informatics, computer science, and medicine each have their own conventions
- Frame your contributions to meet the expectations of your target community and venue
- Choose publication venues strategically: computer-science conferences differ markedly from informatics journals
- **Who do you want to read your research?**

Start **Writing Early**



Credit: David Perell

- You should write and design experiments **at the same time**
- **Do not** treat writing as the **final step of research**
- **Writing is integral to shaping and refining your research questions**

Formulate Your Paper's **Terms-of-Art**

- Papers employ a lot of **jargon / terms-of-art**
- Important to land on **consistent names for concepts**
  - New dataset, method, model?
- **Reduce the cognitive burden on the reader**
- Helpful for all authors **to draft a "living style guide"** for names

Tell a Story with a **Running Example**

- Ground your work in concrete examples
- Use a real case to frame your contribution and strengthen your narrative
- This example influences your framing **shows up in your choice of figures**

Show Not Tell

This is a simple one, but it's surprising how often it is done wrong. Instead of telling someone how great your idea or result is, let your experimental (or theoretical) results do the talking. **Numbers speak louder than words**

In this paper, we present System X and show that it performs well on task Y.

- **Lacks precision**
- "performs well" is open to interpretation
- Fosters false expectations in the reader

In this paper, we present System X and show that it can outperform baseline Z by up to 10x on task Y while remaining within 10% of Z on tasks A, B, and C.

- **Precise statements**
- Claims not open to interpretation

Chris Aberger "*Check Yourself Before You Wreck Yourself*"

Structuring Your Paper

Outline of Paper Sections

Introduction	Related Work Background Methods	Exp. Setup Results Analysis	Conclusion
Significance	**Explanatory**	**Empirical**	**Summarize**

Adapted from: Sasha Rush. *"How to write an okay research paper." YouTube 2024.*

Running Example Paper

Harnessing the Universal Geometry of Embeddings

Rishi Jha **Collin Zhang** **Vitaly Shmatikov** **John X. Morris**
Department of Computer Science
Cornell University

Abstract

We introduce the first method for translating text embeddings from one vector space to another without any paired data, encoders, or predefined sets of matches. Our unsupervised approach translates any embedding to and from a universal latent representation (i.e., a universal semantic structure conjectured by the Platonic Representation Hypothesis). Our translations achieve high cosine similarity across model pairs with different architectures, parameter counts, and training datasets.

The ability to translate unknown embeddings into a different space while preserving their geometry has serious implications for the security of vector databases. An adversary with access only to embedding vectors can extract sensitive information about the underlying documents, sufficient for classification and attribute inference.

Research Pitch: Introduces `vec2vec`, the first method to translate text embeddings across model spaces without access to encoders or paired data.

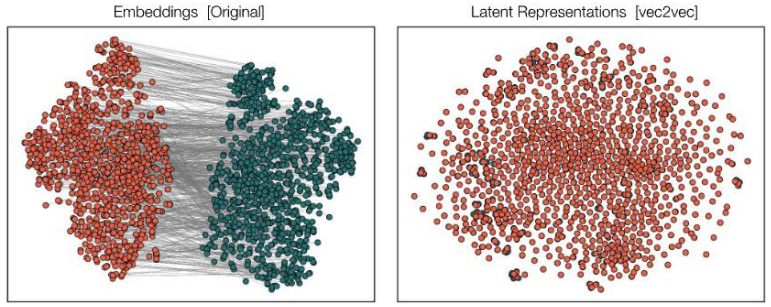
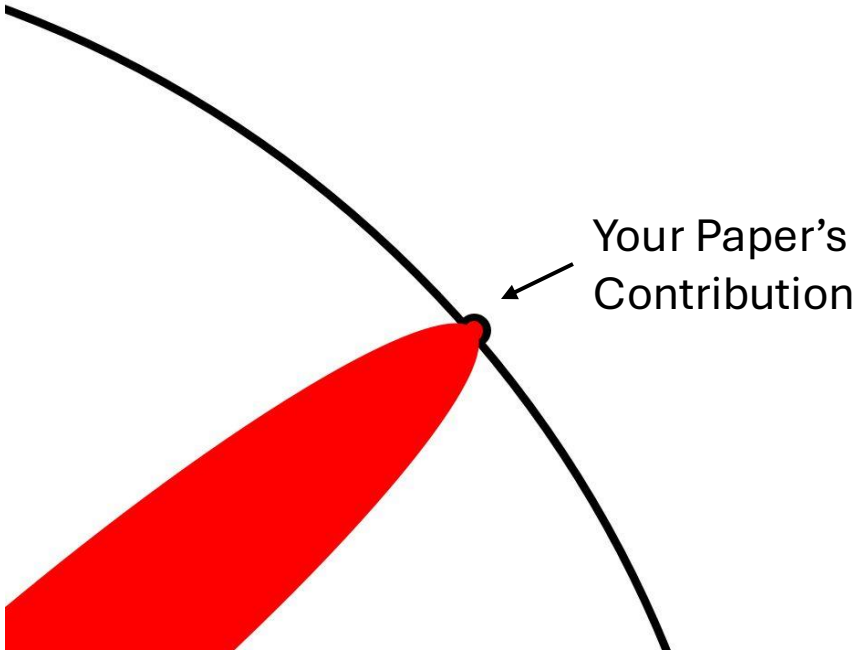


Figure 1: Left: input embeddings from different model families (T5-based GTR [41] and BERT-based GTE [29]) are fundamentally incomparable. Right: given unpaired embedding samples from different models on different texts, our model learns a latent representation where they are closely aligned.

Implications for the security of vector databases

Introduction: Landing Your Research Pitch

Most important goal: articulate how you are
pushing the frontier of knowledge



Credit: Matt Might

- The hardest section to write
- After reading the intro, **most readers will develop a prior to accept or reject**
- You will likely **write this section last**

Introduction: Landing Your Research Pitch

- The introduction should **stay as an outline** until the rest of the paper is drafted
- Define the **minimal area of contribution** you can defend
- This must be novel and justified as novel
- **The writing challenge is restricting focus to only your contribution and significance**
- **Keep It Simple (KIS principle)**

For technical papers, I'm a fan of the "Stanford InfoLab's patented five-point structure for Introductions"

1. **What is the problem?**
2. **Why is it interesting and important?**
3. **Why is it hard?** (e.g., why do naive approaches fail?)
4. **Why hasn't it been solved before?** (Or, what's wrong with previous proposed solutions? How does mine differ?)
5. **What are the key components of my approach and results?** Also include any specific limitations.

“Pull Figure”

Jha et al. “Harnessing the Universal Geometry of Embeddings.” Preprint 2025.

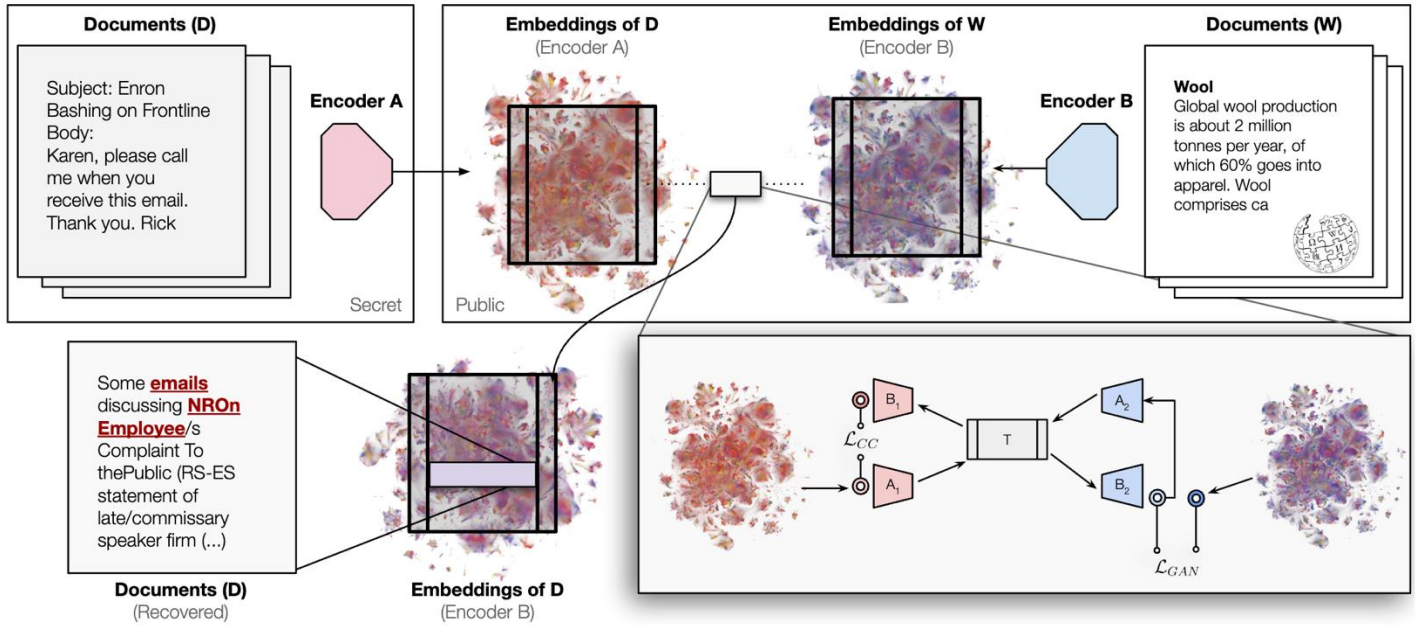


Figure 2: Given only a vector database from an unknown model, `vec2vec` translates the database into the space of a known model using latent structure alone. Converted embeddings reveal sensitive information about the original documents, such as the topic of an email (pictured, real example).

- Main figure of the paper – **akin to a visual abstract**
- Iteratively refine this figure as you write, just like the intro!
- Should capture the **core idea of the paper**
- A great pull figure helps readers build an immediate intuition for the paper

Running Example

Ground Truth: "Subject: Enron Bashing on Frontline \n Body:..."
Generation: "Some emails discussing NROn Employee/s Complaint To thePublic ..."

Ground Truth: "Subject: Trades for 3/1/02 \n Body: \n John , \n The following trades..."
Generation: "... future transactions may await John G..."

Ground Truth: " The following expense report is ready for approval..."
Generation: " The upcoming expense statement from YYYY MM Dec..."

Figure 6: Examples of inversions that infer entities and content .

Corporate emails are the **running examples** used in this paper

Related Work: Goals

- Clarify the contributions and novelty of your work
- Use existing work to **contrast and sharpen your contributions**
- **Setup and motivate baselines**

Common Mistakes

- This is not an exhaustive literature review
- Avoid **simple book-keeping** (list of papers sort-of related papers)

Related Work: Goals

Advice

- Create a spreadsheet with **~3 facets of your work**
- Collect papers that bucket under each facet
  - Not work your paper users -- **attempts at reaching a similar goal**

	A	B	C	D
1	Facet	Paper Title	Year	Contrast
2				
3				
4				
5				
6				
7				
8				

Related Work: Example

Facets

Contrasts

Baselines

Representation alignment. Similarities between representations of different neural networks are investigated in [24, 28, 52, 4, 17]. CCA [36], SVCCA, [45], CKA [21], ICA [55], and GUI-based systems such as embComp [15] compare embeddings from different subspaces. Methods like [33, 39, 34, 42] harness representation similarity for zero-shot stitching, substitution, and multi-modal adaptation. All rely on some amount of paired data, which is difficult to reduce [5]. Our method does not just measure similarity, we learn how to *translate* representations across spaces without any paired data.

Unsupervised transport. The problem of unsupervised transport has been studied for images [16, 32, 61], words [54, 9, 14, 8], and natural language sequences [46, 25, 1, 3, 56]. Our method builds on these works, which often employ a combination of cycle-consistency and adversarial loss. [48] proposes a solver for matching small sets of embeddings between different vision-language models. Our method goes well beyond matching by taking unknown embeddings and *generating* matching embeddings in the space of another model.

Embedding inversion. An emerging line of research investigates decoding text from language model embeddings [49, 27, 37] and outputs [38, 6, 58]. `vec2vec` helps apply these to unknown embeddings, without an encoder or paired data, by translating them to the space of a known model.

Bridging modality gaps. Previous work has noted an inherent "gap" between image- and text-based models [30] and proposed various ways to unify the modalities [50]. Some approaches feed image embeddings directly into language models [20, 53, 10, 31], while others generate captions from image embeddings [35] or even from text embeddings themselves [37]. [11] introduces a shared embedding space that integrates inputs from multiple modalities, including text, audio, and vision. In contrast, our post-hoc approach directly translates between representations and complements these systems by enabling inputs from a wide variety of embedding models.

Background: Goals

- The "`import statement`" to make your contribution compile for the reader
- The imports must be relevant to understanding your work
- **Introduce core notation here** (i.e., problem formulation, canonical forms)
- **Reduce cognitive burden for the reader**

Common Mistakes

- Do not introduce unnecessary concepts that are not directly used by your contribution
- If you are doing a technical paper, avoid the impulse to overly "math up" your paper
- **Example:** Including notion for multi-headed attention when just using a Transformer

Background: Example

Problem Formulation

Consider a collection of embedding vectors $\{u_1, \ldots u_n\}$, for example, a dump of a compromised vector database, where each $u_i = M_1(d_i)$ is generated by an unknown encoder $M_1 : \mathbb{V}^s \to \mathbb{R}^{d_{M_1}}$ from an unknown document d_i. We cannot make queries to M_1 and do not know its training data, nor architectural details. Our goal is to extract any information about the documents d_i.

Core Underlying Assumption

The Strong Platonic Representation Hypothesis. Our hope that unsupervised embedding translation is possible at all rests on the stronger version of the Platonic Representation Hypothesis [17]. Our conjecture is as follows: *neural networks trained with the same objective and modality, but with different data and model architectures, converge to a universal latent space such that a translation between their respective representations can learned without any pairwise correspondence.*

Background: Example

"Sign Posting" Figure

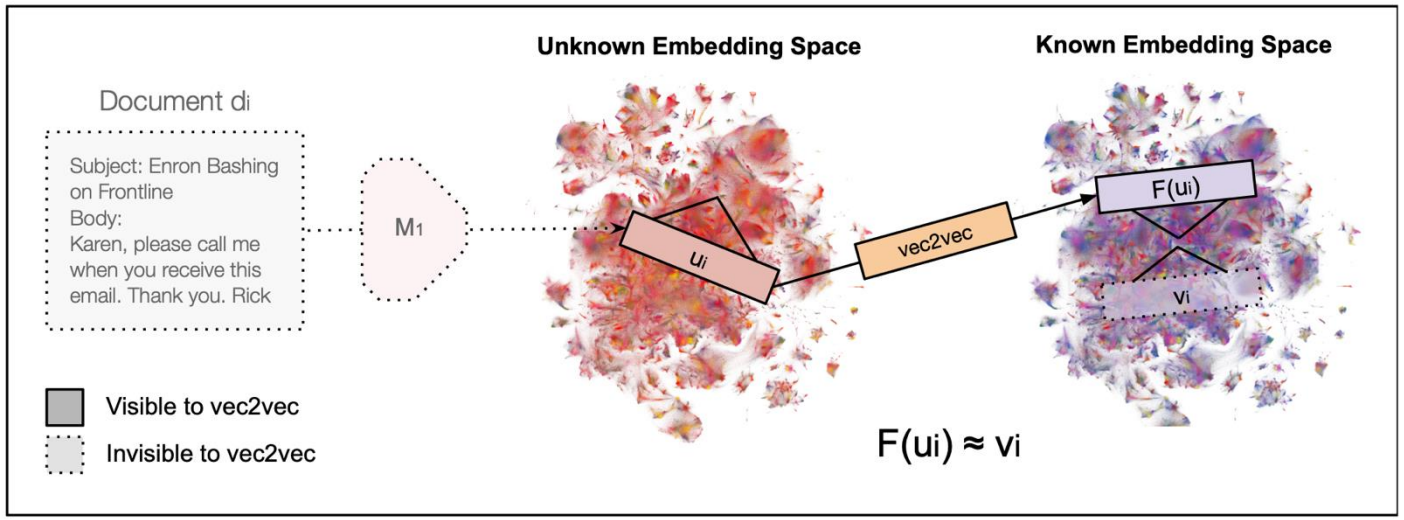


Figure 3: **Unsupervised embedding translation.** With access to only $u_i = M_1(d_i)$, vec2vec seeks to generate a translation $F(u_i)$ that is close in M_2's embedding space to the ideal embedding $v_i = M_2(d_i)$ without access to d_i, v_i, M_1, or M_2.

- Re-iterates core idea
- Provides a visual intuitive to reinforce the mathematical framing
- **Orientes the reader on where the paper is heading**

Methods: Goals

- How would you **explain your contribution explain to your past self?**
- **Do not introduce any new core notation –** employ what you setup in the **Background** section
- **Common Problem:** the methods section grows too big. Keep minimal necessary formulas, concepts, etc.
- **This is an expensive part of the paper for readers to grok!**

Methods: Example

1. **Architecture**
2. **Optimization**
    1. Adversarial loss
    2. Generator
        1. *Loss Terms*

This takes ~1.2 pages of space in Jha et al. 2025

Reconstruction enforces that an embedding, when mapped into the latent space and back into its original embedding space, closely matches its initial representation:

$$\mathcal{L}_{\text{rec}}(R_1, R_2) = \mathbb{E}_{x\sim p}\|R_1(x) - x\|_2^2 + \mathbb{E}_{y\sim q}\|R_2(y) - y\|_2^2.$$

where p and q are distributions of embeddings sampled from M_1 and M_2, respectively.

Cycle-consistency acts as an unsupervised proxy for supervised pair alignment, ensuring that F and G can translate an embedding to the other embedding space and back again with minimal corruption:

$$\mathcal{L}_{\text{CC}}(F_1, F_2) = \mathbb{E}_{x\sim p}\|F_2(F_1(x)) - x\|_2^2 + \mathbb{E}_{y\sim q}\|F_1(F_2(y)) - y\|_2^2.$$

Vector space preservation (VSP) ensures that pairwise relationships between generated embeddings remain consistent under translation [40, 57]. Given a batch of B embeddings $x_1, ..., x_B$ and $y_1, ..., y_B$, we sum their average pairwise distances after translation by both F_1 and F_2:

$$\mathcal{L}_{\text{VSP}}(F_1, F_2) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\Big[\|M_1(x_i)\cdot M_1(x_j) - F_1(M_1(x_i))\cdot F_1(M_1(x_j))\|_2^2 + \|M_2(y_i)\cdot M_2(y_j) - F_2(M_2(y_i))\cdot F_2(M_2(y_j))\|_2^2\Big]$$

Terms-of-art and consistent notation

Experimental Setup: Goals

- **Very standard recipe for this** section of the paper: **Metrics, Data, Baselines**
- **This is usually the easiest section to write**
- First version of this section may be long
- You may eventually move subsections to the appendix

Experimental Setup: Example

Datasets

Datasets. We use the *Natural Questions (NQ)* [23] dataset of user queries and Wikipedia-sourced answers for training (a 2-million subset) and evaluation (a 65536 subset). To evaluate information extraction, we use *TweetTopic* [2], a dataset of tweets multi-labeled by 19 topics; a random 8192-record subset of *Pseudo Re-identified MIMIC-III (MIMIC)* [26], a pseudo re-identified version of the MIMIC dataset [18] of patient records multi-labeled by 2673 MedCAT [22] disease descriptions; and a random 50-email subset of the *Enron Email Corpus (Enron)* [19], an unlabeled, public dataset of internal emails of a defunct energy company.

Baselines

`vec2vec` is the first unsupervised embedding translator, thus there is no direct baseline. As our **Naïve** baseline, we simply use $F(x) = x$ to measure geometric similarity between embedding spaces. The second (pseudo)baseline is **Oracle-aided optimal assignment**. It assumes that candidate targets are known and is thus strictly easier than `vec2vec` and the Naïve baseline. We solve optimal assignment, $\pi^* = \arg\min_\pi \sum_{i=1}^{n} \cos(u_i, v_{\pi(i)})$, via either the Hungarian, Earth Mover's Distance, Sinkhorn, or Gromov-Wasserstein algorithms, choosing the best-performing solver per experiment.

Metrics

Given (unknown) embeddings $\{M_2(d_j)\}_{j=0}^{n}$ ordered by decreasing cosine similarity to $F(u_i)$, let r_i be the rank of the correct embedding $v_i = M_2(d_i)$. To measure quality of F, we use three metrics. **Mean Cosine Similarity** measures how close translations are, on average, to their targets. **Top-1 Accuracy** is the fraction of translations whose target is closer than any other embedding. **Mean Rank** is the average rank of targets with respect to translations. The ideal translator F^* achieves mean similarity of 1.0, top-1 accuracy of 1.0, and mean rank of 1.0. Recall that a random alignment corresponds to a mean rank of $\frac{n}{2}$. Formally,

Models

Model	Params (M)	Backbone	Year	Dims	Max Seq.
[41] gtr	110	T5	2021	768	512
[44] clip	151	CLIP	2021	512	77
[51] e5	109	BERT	2022	768	512
[29] gte	109	BERT	2023	768	512
[60] stella	109	BERT	2023	768	512
[13] granite	278	RoBERTa	2024	768	512

Table 1: Embedding models used in our experiments.

This takes ~1.2 pages of space in Jha et al. 2025

Results: Goals of Tables & Figures

- 1:1 mapping of experiments to table/figure
- Compare like-to-like in tables
- **Be consistent.** Use the same notation and terms-of-art
- Structure so that each table/figure has a clear story

Common Mistakes

"A Paper Is Not Like a Sitcom; There Should Be No Big Reveal at the End"

Don't save your best result for a big reveal at the end. The best results should come first.

Experiment Result Ordering

- Micros (Ablation Study)
- Macros (End-to-End Comparison)

Experiment Result Ordering

- Macros (End-to-End Comparison)
- Micros (Ablation Study)

Results: Examples of Tables & Figures

First Result Figure

Performance clear across all 5 models

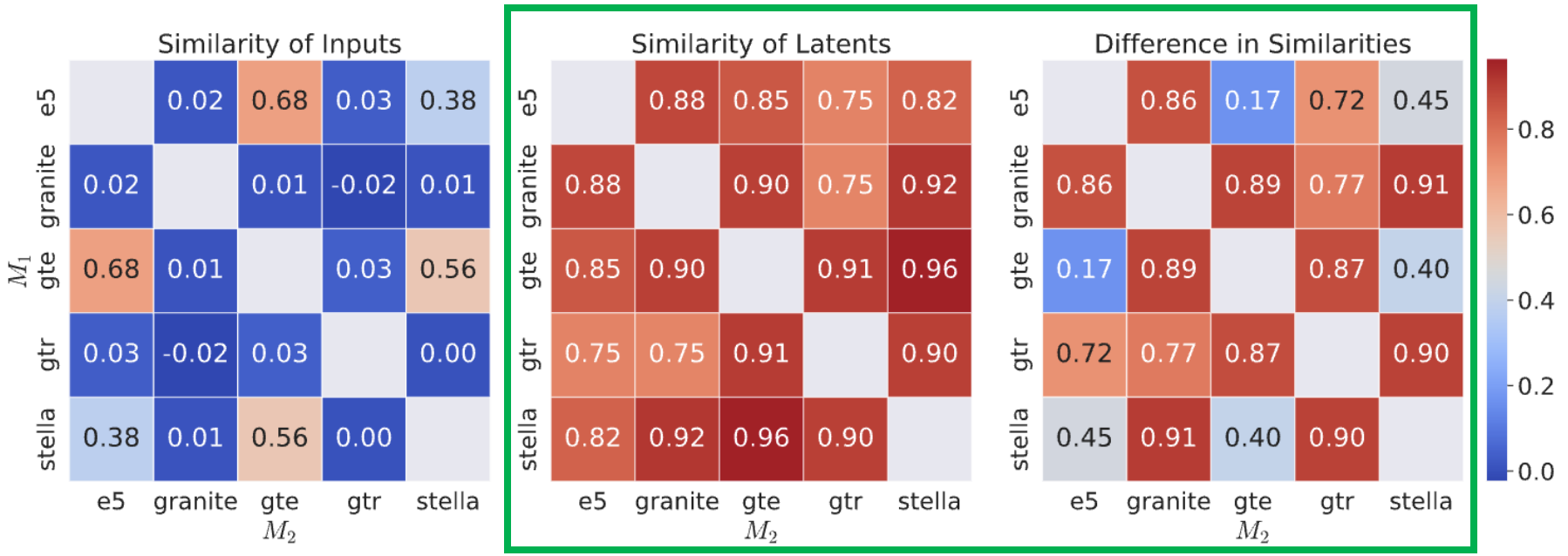


Figure 4: Pairwise cosine similarities of input embeddings (left) and their `vec2vec` latents (middle) across different embedding pairs. The absolute difference between the heatmaps plots is on the right. All numbers are computed on the same batch of 1024 NQ texts.

Jha et al. 2025.

Results: Examples of Tables & Figures

1:1 Mapping of Figure/Table to Text

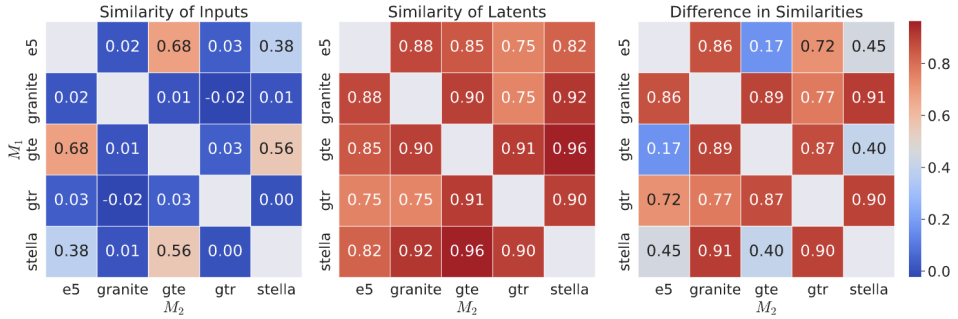


Figure 4: Pairwise cosine similarities of input embeddings (left) and their vec2vec latents (middle) across different embedding pairs. The absolute difference between the heatmaps plots is on the right. All numbers are computed on the same batch of 1024 NQ texts.

- Reiterate your interpretation of the results -- readers may disagree with your conclusions!
- Employ clear 1:1 mappings to make claims and structure transparent and unambiguous

vec2vec **learns a universal latent space.** vec2vec projects embeddings $M_{1,2,\ldots}$ into a shared latent space via compositions of input adapters $(A_{1,2,\ldots})$ and a shared translator T. Figure 4 shows that even when the embeddings $u_i = M_1(d_i)$ and $v_i = M_2(d_i)$ are far apart (*i.e.*, have low cosine similarity), their representations in vec2vec's latent space are incredibly close: $T(A_1(u_i)) \approx T(A_2(v_i))$. Figure 1 visualizes this (via two-dimensional projections) for vec2vec trained on gte and gtr embeddings: the embeddings are far apart, but their latents are *nearly overlapping*.

Results: Examples of Tables & Figures

Terms-of-art and consistent notation

Metrics and directionality clear

		vec2vec			Naïve Baseline			OA Baseline	
M_1	M_2	cos(·) ↑	Top-1 ↑	Rank ↓	cos(·) ↑	Top-1 ↑	Rank ↓	Top-1 ↑	Rank ↓
gran.	gtr	**0.80** (0.0)	**0.99**	**1.19** (0.1)	-0.03 (0.0)	0.00	4168.73 (9.2)	0.00	4096.75 (9.2)
	gte	**0.87** (0.0)	**0.95**	**1.18** (0.0)	0.01 (0.0)	0.00	4088.58 (9.2)	0.00	4069.91 (9.2)
	stel.	**0.79** (0.0)	**0.98**	**1.05** (0.0)	0.01 (0.0)	0.00	4208.26 (9.2)	0.00	4096.75 (9.2)
	e5	**0.85** (0.0)	**0.98**	**1.11** (0.0)	0.02 (0.0)	0.00	4111.60 (9.2)	0.00	4096.75 (9.2)
gtr	gran.	**0.81** (0.0)	**0.99**	**1.02** (0.0)	-0.03 (0.0)	0.00	4169.76 (9.2)	0.00	4096.66 (9.2)
	gte	**0.87** (0.0)	**0.93**	**2.31** (0.1)	0.04 (0.0)	0.00	4080.92 (9.2)	0.00	4079.92 (9.2)
	stel.	**0.80** (0.0)	**0.99**	**1.03** (0.0)	0.00 (0.0)	0.00	4198.78 (9.2)	0.00	4096.69 (9.2)
	e5	**0.83** (0.0)	**0.84**	**2.88** (0.2)	0.03 (0.0)	0.00	4082.84 (9.2)	0.00	4066.42 (9.2)

Jha et al. 2025.

Results: Examples of Tables & Figures

Both baseline methods completely fail vs. their method

		vec2vec			Naïve Baseline			OA Baseline	
M_1	M_2	$\cos(\cdot)\uparrow$	Top-1 ↑	Rank ↓	$\cos(\cdot)\uparrow$	Top-1 ↑	Rank ↓	Top-1 ↑	Rank ↓
gran.	gtr	**0.80** (0.0)	**0.99**	**1.19** (0.1)	-0.03 (0.0)	0.00	4168.73 (9.2)	0.00	4096.75 (9.2)
	gte	**0.87** (0.0)	**0.95**	**1.18** (0.0)	0.01 (0.0)	0.00	4088.58 (9.2)	0.00	4069.91 (9.2)
	stel.	**0.79** (0.0)	**0.98**	**1.05** (0.0)	0.01 (0.0)	0.00	4208.26 (9.2)	0.00	4096.75 (9.2)
	e5	**0.85** (0.0)	**0.98**	**1.11** (0.0)	0.02 (0.0)	0.00	4111.60 (9.2)	0.00	4096.75 (9.2)
gtr	gran.	**0.81** (0.0)	**0.99**	**1.02** (0.0)	-0.03 (0.0)	0.00	4169.76 (9.2)	0.00	4096.66 (9.2)
	gte	**0.87** (0.0)	**0.93**	**2.31** (0.1)	0.04 (0.0)	0.00	4080.92 (9.2)	0.00	4079.92 (9.2)
	stel.	**0.80** (0.0)	**0.99**	**1.03** (0.0)	0.00 (0.0)	0.00	4198.78 (9.2)	0.00	4096.69 (9.2)
	e5	**0.83** (0.0)	**0.84**	**2.88** (0.2)	0.03 (0.0)	0.00	4082.84 (9.2)	0.00	4066.42 (9.2)

Jha et al. 2025.

Analysis: Anticipate Reviewer Requests

- *Can you compute bootstrap confidence intervals?*
- *Can you report subgroup performance?*
- *Can we do an error analysis on experiment X?*
- *Can you run another baseline Y?*

Common Mistakes

“You Don’t Get Credit for What You Didn’t Do”

You cannot write your way around why you didn’t set up an experiment.

“Where is the Ablation Study?”

Ablation experiments break down and isolate the benefit of every contribution you make in the paper. They provide important context for understanding the benefits and limitations of your contribution.

Discussion: Wrapping Up Your Contributions

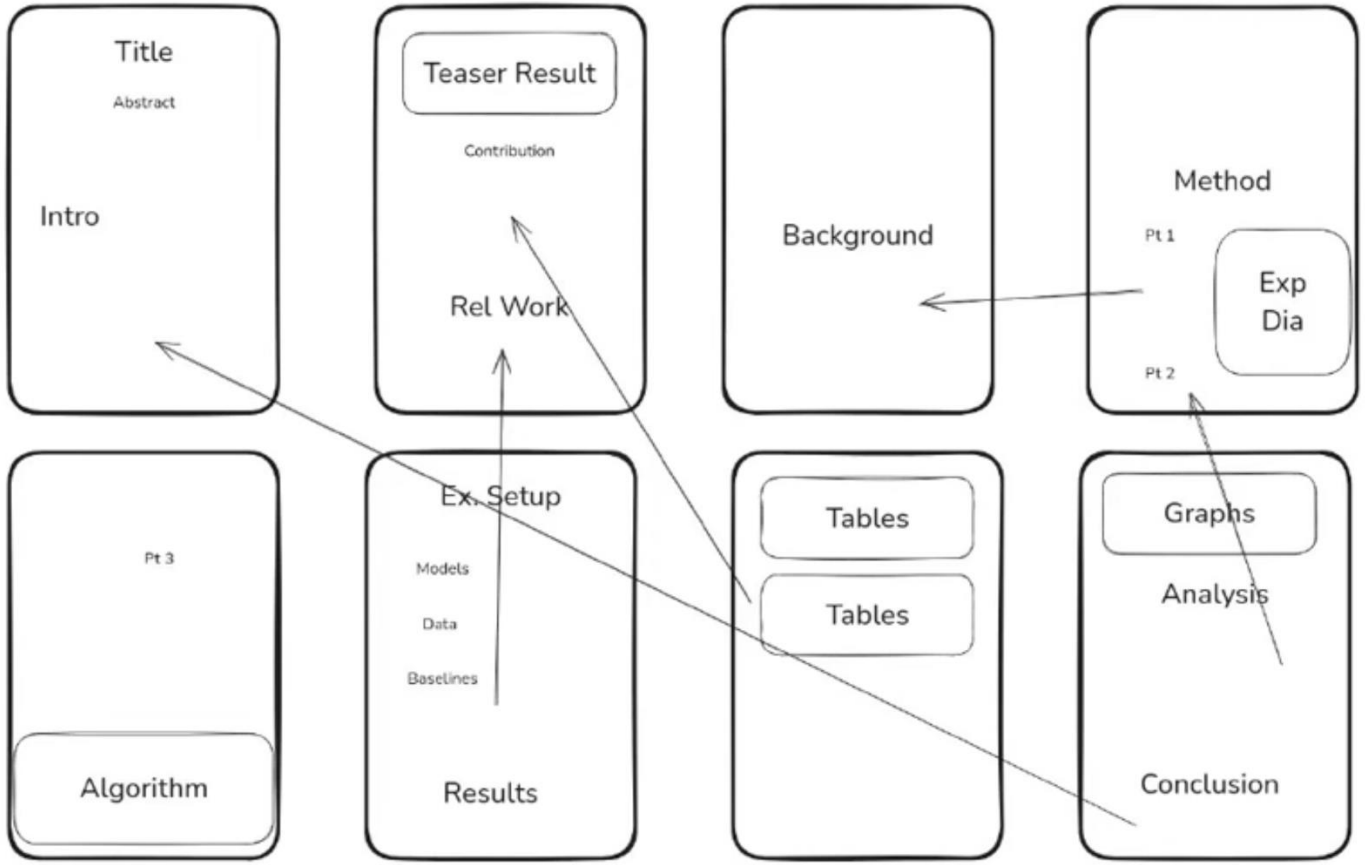
- **Repeat your core claims and contributions**
- Discuss the implications of your findings
- What are any essential limitations?

> The Platonic Representation Hypothesis conjectures that the representation spaces of modern neural networks are converging. We assert the Strong Platonic Representation Hypothesis: the latent universal representation can be learned and harnessed to translate between representation spaces without any encoders or paired data.
>
> In Section 5, we demonstrated that our `vec2vec` method successfully translates embeddings generated from unseen documents by unseen encoders, and the translator is robust to (sometimes very) out-of-distribution inputs. This suggests that `vec2vec` learns domain-agnostic translations based on the universal geometric relationships which encode the same semantics in multiple embedding spaces.

Mapping **Structure** and **Page Utilization**

- Useful to **see the global structure** of a paper
- How much space is used for each section?
- How do sections connect?
- **Sketch this out on a whiteboard**!

Use this structure to build in **sign posts** for the reader



Adapted from: Sasha Rush. *"How to write an okay research paper." YouTube 2024*.

Discussion: Common Mistakes

Telling instead of showing

- Relying on vague claims without concrete evidence, examples, or visuals

Following the order of discovery instead of narrative clarity

- Presenting ideas as they were developed rather than in a reader-friendly logical flow

Using inconsistent terminology

- Switching terms for the same concept, making the paper harder to “compile” in the reader’s mind

Checklists

Lots of great recipes for sanity checking your writing

Check Yourself Before You Wreck Yourself

Chris Aberger

CS Paper Checklist

Yue Zhao

1. 🎯 Title and Abstract

- [ ] 1.1 Title is ≤ 15 words. Check for generic phrasing (e.g., "A Novel Framework...," which conveys little information) and overly narrow focus (which may reduce the paper's audience)—aim for concise but informative.
- [ ] 1.2 Title clearly reflects both the **problem** and the **solution**, and includes at least one technical keyword (e.g., jailbreak, OOD detection, graph learning).
- [ ] 1.3 Title avoids rare or ambiguous abbreviations. Terms like LLM, AI, and ML are acceptable in CS venues, but avoid abbreviations like AD (which could refer to advertisement or anomaly detection).
- [ ] 1.4 Abstract includes at least four key components: (1) problem/task definition, (2) proposed method or idea, (3) main results, and (4) broader impact or significance (some may be combined).
- [ ] 1.5 Abstract avoids undefined abbreviations and vague descriptors (e.g., "important," "novel," "state-of-the-art" without context).
- [ ] 1.6 Bonus: Abstract includes at least one concrete, quantitative result or insight to make the work stand out. For instance, "our method achieves 11.2× acceleration in test-time inference for jailbreak detection."

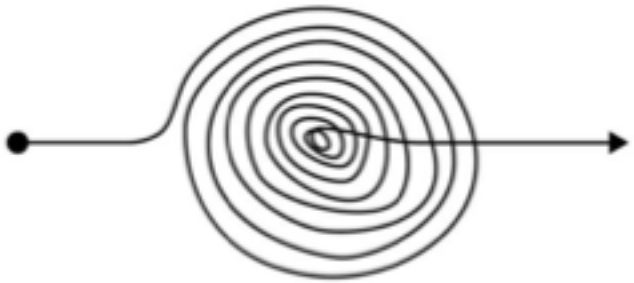
2. 📚 Introduction

- [ ] 2.1 The main problem or task is clearly defined within the first two paragraphs.
- [ ] 2.2 Motivation includes either (a) real-world use cases or (b) citations to prior work—ideally both.
- [ ] 2.3 The introduction ends with a brief overview of the proposed method and its name.
- [ ] 2.4 Contributions are explicitly itemized (e.g., "(1) first framework for ..., (2) new dataset for ..., (3) extensive evaluation on ...").

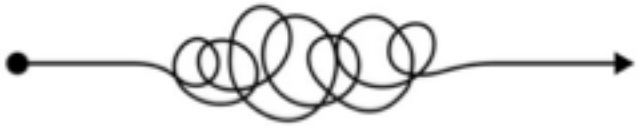
https://github.com/yzhao062/cs-paper-checklist

Build Writing Habits for Research

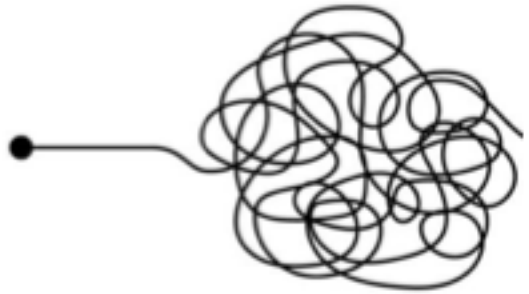
Writing and Research are **Iterative**



The Hope



The Reality



The Danger

Create Enough Space to **Reach Flow State**

- Writing requires **cognitive focus** and time to ramp up
- Clear enough time in your schedule to **reach flow state**
- Minimize distractions!
  - **Example**: I like writing in coffee shops
- Put your phone in the other room

Flow State, In-the-Zone, Lazer-Focused, etc.

When someone is completely immersed in a task or activity

Create a **Project Timeline** with **Milestones**

- Always **draft milestones with clear deadlines**
- 3 weeks before the final deadline all core experiments should be complete

Sala, Varma, Ré. "What is a Paper. An Exploration." Hazy Research. 2019.

Outline Swaps

Once you have a sketch of a paper, seek feedback from your peers to **make suggestions for improvement, find bugs, and otherwise stress-test your idea**.

Elevator Pitch

- The initial research outline presented to the lab before any work is done
- Central claim or hypothesis and contribution
- Anticipate possible criticisms (what is this work doing different?) and think about your rebuttal

Experiments (~3 max)

Keep a **Research Journal**

- Papers are **receipts for research**, but they don't capture the process
- Research journals **reflect how the sausage is actually made**

Research Log - Synthetic Instructions

Author: Jason Alan Fries

Code: https://github.com/som-shahlab/synth-instruct

Sep 25, 2024

TLDR

Only 29% (36/124) of synthetic instructions pass naive quality filters. This doesn't even look at the quality of response or diversity of instructions.

Note: denominator is not the same across all rows due to fast/opportunistic labeling of columns.

- **84% (104/124)** Plausible Clinical Instructions
- **7.6% (7/92)** Missing context (vague, missing part of conversation, etc)
- **40% (48/119)** EHRs | **51.2% (61/119)** | Patient **8.4% (10/119)** N/A
- **31.5% (24/76)** are Simple QA-style questions

Goals

- Manually annotate a subset of ~300 example responses, using schema outlined yesterday
- Understand distribution of errors.

Protocol:

- Make excel spreadsheet and label according to quality review schema (see yesterday's TLDR)

Discussion: Keep a **Research Journal**

Outline an agenda for the day and a time budget

- What research question(s) do I want to answer today?
- When do I cross off a question?
- **Requires focused time or *maker time* (Flow State)**

Maintain **Clear Provenance** for Results

Why do we care?

- Easier to write your manuscript
- You can work backwards if something goes wrong!

Make it **Easy for Others to Help Out**

Your may want to **onboard excited new students to help** on your project or **get feedback on sections-in-progress**

Experiments

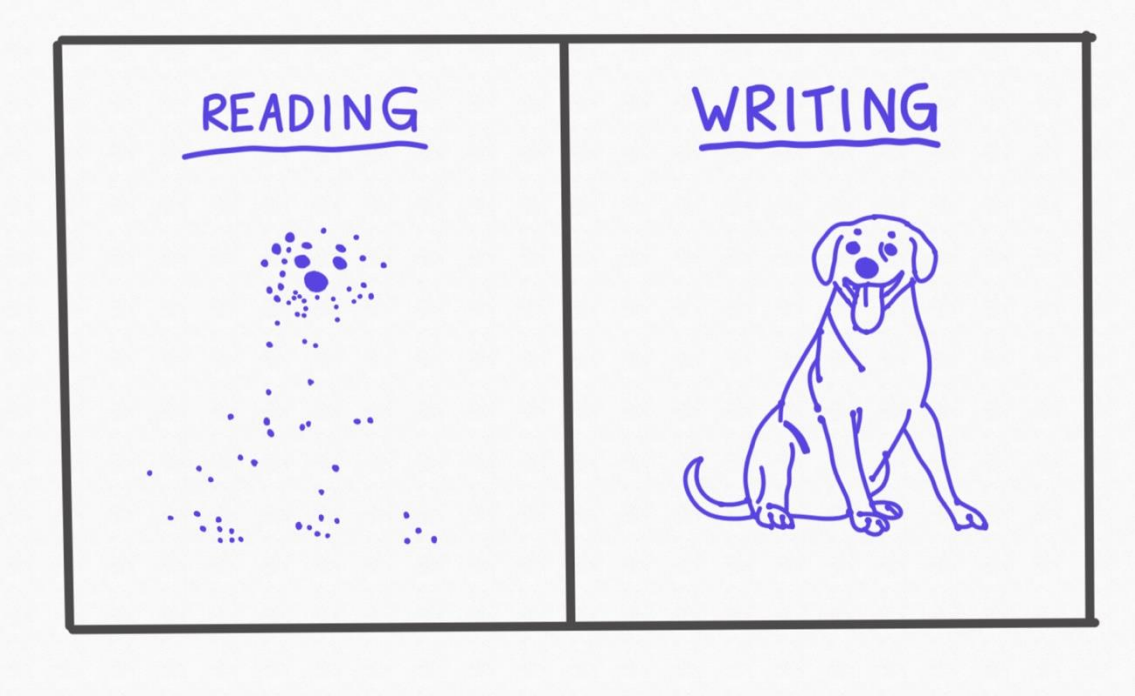
- Sometimes you **need a mercenary to help** with a specific piece of code or experiment.
- If your code and experimental setup is super messy, this makes it very difficult to help you.

Writing Feedback

- Keep you pitch and intro outline up-to-date!
- Any written content makes it easier to provide actionable feedback

Closing Thoughts

- Scientific writing is challenging – **don’t’ get discouraged**!
- Rely on your peers for frequent writing swaps and feedback



Be a **Good Reader**

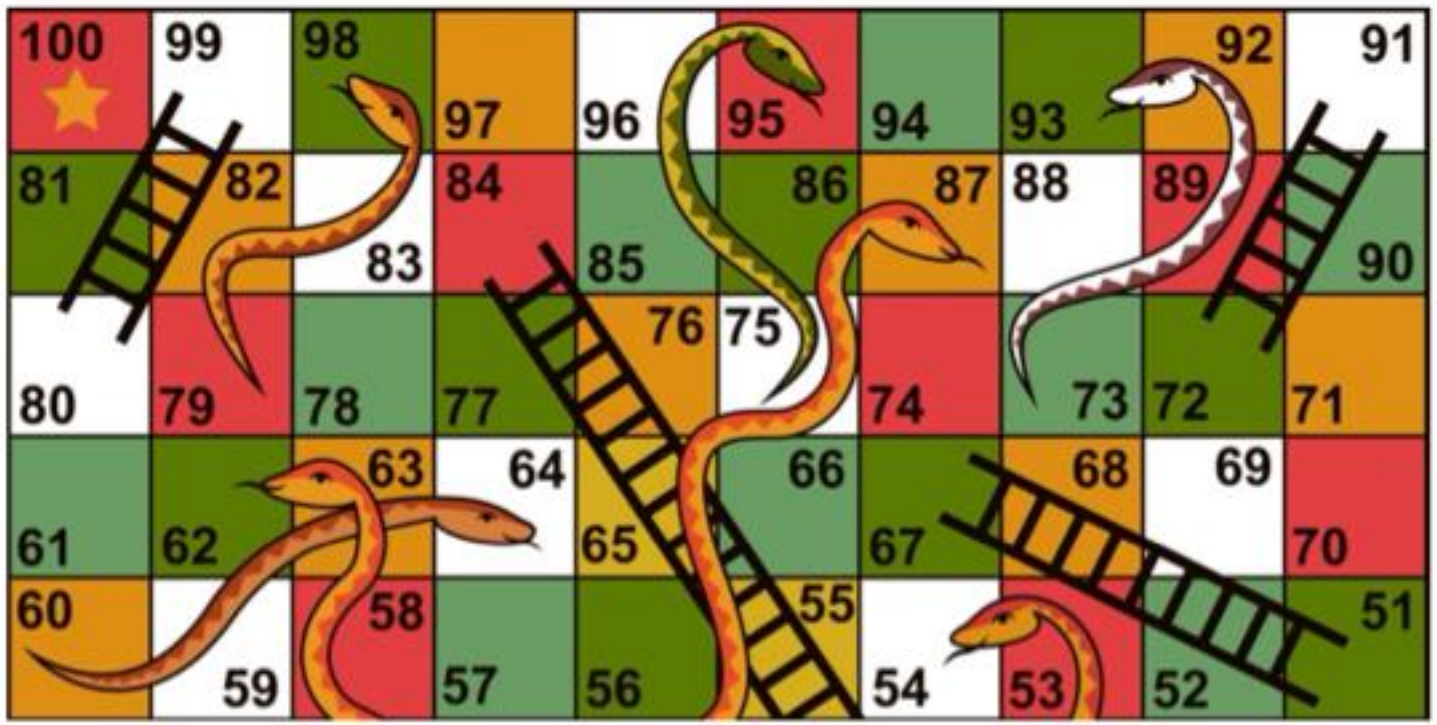
- Don’t be mean, but **don’t be too nice**. The goal is actionable feedback for authors
- Be generous in reading, but honest about gaps and confusing points

Questions & Discussion

jfries@stanford.edu

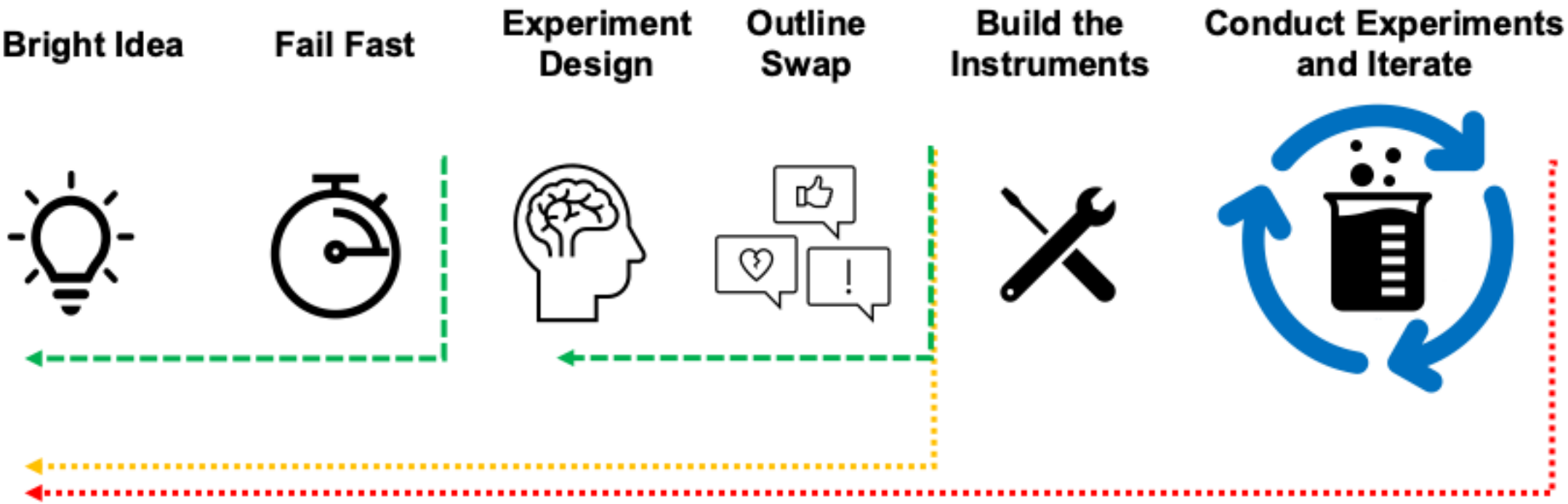
Appendix: Designing Meaningful Experiments

Research is a Bit Like Snakes & Ladders



How do you
Build Ladders
and
Avoid Snakes?

Research Project Lifecycle



Core Purposes

De-Risk Your Research Ideas – **Avoid the Snakes!**

Strengthen Your Findings

Strong Inference (Platt 1964)

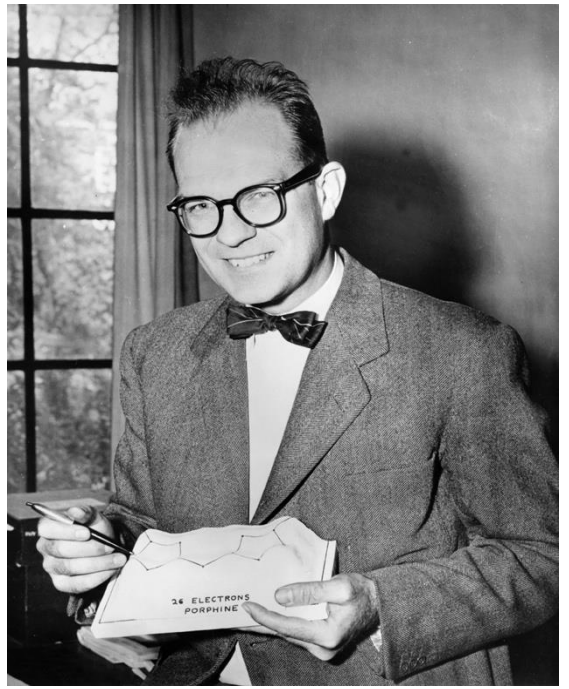
A structured and systematic approach to scientific problem-solving. It proposes that progress in science is best achieved through a method that **continuously narrows down hypotheses**.

Formulation of multiple hypotheses: Develop several plausible explanations rather than relying on just one.

Critical experiments: Design experiments to decisively rule out competing hypotheses.

Iterative process: After rejecting certain hypotheses, form and test new ones in a continuous cycle to refine understanding.

"The difference between the average scientist's informal methods and the methods of strong-inference users is somewhat like the difference between a gasoline engine that fires occasionally and one that fires in a steady sequence."



John R. Platt
1918 - 1992

The Importance of **Failing Fast**

- **Design a short experiment (~1 week)**
- Ask yourself if this problem is worth investing the next 2-3 months.
- The goal is to quickly determine success or failure.
- Don't worry about coding elegance or complex setups—focus on a single Jupyter notebook to test the core idea.
- **You may need several iterations of rapid experiments before landing on a solid direction.**

The Experiment

Most papers have **one central hypothesis** and **several ancillary, supporting hypotheses**

Hypothesis:
A testable prediction about the outcome of an experiment.
Proxy:
A substitute measure for a variable that cannot be directly observed.
Protocol:
The detailed plan outlining your experiment setup, datasets, and baselines.
Expected Results:
A prediction of the anticipated outcome based on the hypothesis.
Results:
The actual findings from the experiment that guide further iterations.
This happens after your first iteration!

The Experiment

From Ce Zhang in Chris Re's HazyResearch Lab

Hypothesis: Caffeine makes graduate students more productive.

Proxy:

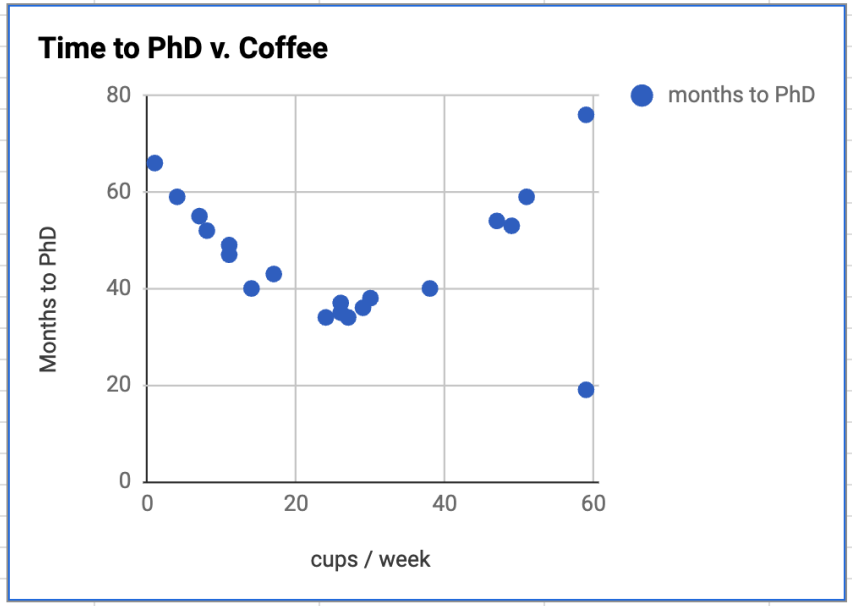
- Productivity: the time it takes to complete their PhD
- Caffeine Intake: the number of cups of coffee a student drinks (we do not count soda or tea)

Protocol: Out of the 100 students in our school, have them report the mean cups of coffee they drink each week. Also, record the date they start their PhD and the date they get their PhD. Only consider students who have left the school with a PhD. Note that if a student is very productive, but dies of caffeine toxicity days before they finish their PhD, they will not be counted

The Experiment

From Ce Zhang in Chris Re's HazyResearch Lab

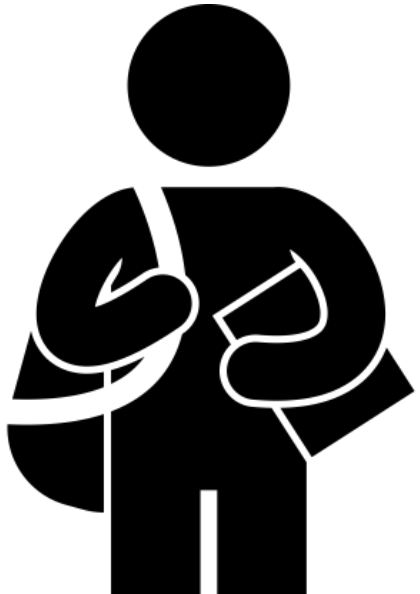
Expected Result: We expect that students without caffeine take a long time to get their PhD; students who drink a lot of caffeine finish their PhD very quickly. But drinking too much caffeine would cause caffeine toxicity and would length the time to get a PhD due to psychosis. In short, we expect a convex curve as attached.



Result: The results match our expected results. We observe that the knee of the curve is at approximately 30 cups of coffee per week with a PhD in 36 months; this seems to be the optimal amount of coffee to drink to be a productive graduate student.

Discussion: The Experiment

MIMIC 4 Feature Experiment from Lawrence Guo

From Jenifer Widom's Tips for Writing Technical Papers

- It's easy to do "hokey" or meaningless experiments, and many papers do.
- It's easy to craft experiments to show your work in its best light, and most papers do."
- What should experiments measure?