Philosophical Progress and Improvement Across Time

Introduction

There are many fields, especially the ones that aren't as quantitative, which are branded as naval gazing. In the philosophy book *Phenomonolgy,* Chad Engelland writes: "The various sciences that philosophy has spun off have achieved independent success…while philosophy itself seems unable to take a single step beyond it's starting point." Just looking at philosophy, I would argue that there is evidence that true progress is being made and the field behaves like they can discern good philosophy from bad and that the field is making progress in distilling more and more good philosophy.

I will not argue whether there are objective truths, especially in the humanities. If we are talking about science, it's hard to argue gravity is not an objective truth. But the idea in the humanities that there is no such thing as progress and only changes of (group) opinion guiding creative work, whether that's painting, novels, movies, or philosophy, is quite prevalent. This is not surprising. The fact that for most of history, there are no tools to measure good/bad philosophy, English literature etc. other than the opinions of others, allows that theory to proliferate.

However, with cutting edge machine learning and natural language processing, we now have better tools to quantify different areas of more qualitative fields, and it is now possible to measure suggestive evidence regarding progress in these previously qualitative fields. Focusing on one field, I will argue that the field of philosophy, which has a large contingent sympathetic to the idea that it has made no progress, and even extending this line of thinking alogn post-modern and post-structuralist theory, that there is no such thing as progress towards truths only changes in points of view[1]. Nevertheless, despite prevalence of this extreme view across some of the humanities and even attempts to argue that more quantitative fields in the sciences also don't progress towards truth but just move towards

[1] Although philosophy is the birthplace of post-structuralism/relativism, nowadays philosophy has a smaller contingent of relativists than other humanities fields as well as social sciences that are more qualitative like sociology. Despite that, I am more familiar with the field, the literature mainly consists of texts which are better analyzed than artwork for embeddings, and it has a long history compared to many of the other fields. Thus, I will analyze philosophical works with machine learning, but I imagine the same sort of analysis could be applied to other fields.

points of view[2], philosophy acts in a way which suggests it's members believe that they are making progress and closing in on better understanding of "truths". I use machine learning analysis of texts to find suggestive evidence that the field of philosophy is making progress, improving, and producing at least incrementally agreed upon jointly recognized knowledge that progresses over time.

Although not entirely relevant to this endeavor, I would next like to argue the conditions that are necessary subject matter than humans engage in to improve over time:

1) roughly ideas that are more likely good (by whatever standard good is: movies that people like, basketball moves that will score you more points etc.) are more likely to be adopted that ideas that are less good.
2) When ideas are evaluated against one another, what happened in the past doesn't influence the choice now to such an extent that we don't too much discount a better idea in favor of what we have done in the past.

The second point ensures that no one idea doesn't derail the path of history. There could be a detour, but over time, the field will self-correct back towards solutions that are more optimal. As an example of why this system works, evolution follows these two premises and results in life becoming better adapted to their local environment.

Most fields, even the ones that aren't quantitative, have these two features. Most fields also progress over time. Just because we can't measure progress in story telling the way we can measure GDP, or new scientific discovery, doesn't mean we aren't getting better at story telling. While there are good movies and directors in the past, I would argue we have better formulas to perform engaging movies, whether it's a refined hero's journey, specific camera technique in chase scenes, or the setup of a good joke. Of course, these movies may not be high art like Hitchcock or Orson Welles, because the pressure is towards making movies that are as profitable as possible. Nevertheless, despite the invention of competing mediums for play like the internet or Netflix, I would argue the movie industry has gotten better and better at storytelling in a way that generates profits.

We can say the same for NBA players and basketball training. Despite improvements in both defense and offense, so looking at scores won't reveal improvement, NBA players have better preparation (since Detroit won a championship in 2004, every team now has private planes for game travel), better skills and athleticism (Think Russel Westbrook/Lebron James versus anyone in the 1960s-1970s), and better strategies (more

[2] See, for example, the entire field of scientific relativism which is an attempt among sociologists, philosophers, historians, and others to demonstrate relativism as a guiding principle in the sciences

heavy reliance on three-point shooting). All these things suggest improvement in basketball over time.

I can go on and discuss many qualitative fields, but I want to stick to the main point. Most qualitative fields are improving, and an analysis of philosophy suggests that the field is innovating, making progress overtime, and moving in a way that suggests the field acknowledges certain ideas that were hard to discover are closer to philosophical truths than other ideas.

The most important thing I want to show is that there are some findings accepted as truths, and we build upon them in a directional manner rather than field evolving with changing points of view, but no guiding direction for the content and style of the preeminent texts in the field.

One such philosophical example of improvement, we can talk about how modern analytic philosophy fixes logical flaws in older systems. For example, Aristotelian logic also has significant problems addressed in modern analytic philosophy, for example, Aristotle argues that if all S are P, then some S are P. This is generally true, but the edge case where S doesn't exist means that the statement some S are P doesn't follow[3]. There are many other cases of flaws in ancient logic, and this is just one such example of definite progress. Analytic philosophy is the easiest area to show progress as that was one of the points of analytic philosophy: it was more scientific, more logical, and one could show progress in a more quantifiable manner (new proofs are true no matter when they are discovered). This is just one such, obvious example. However, I will show that progress has been occurring for much before analytic philosophy has been a field and in competing continental philosophy as well in much the same extend as analytical philosophy.

Before I continue, I would like to fully acknowledge that I am a data scientist and not an expert in philosophy, although it has been a hobby for a long time. I welcome any criticism of my claims, especially involving areas of philosophy where I lack expertise and will engage in a faithful manner. At the same time, I will acknowledge that this write up is an example of economical imperialism. However researchers in the humanities are also guilty of the same type of imperialism, attempting to theorize about how science work in with many theories that are accepted but some like scientific relativism that are roundly criticized by scientists.

Scientific relativism argues that no method can determine the superiority of one theory over another, but theories arise out of prevalent points of view or other social (and not physical) constructs at the time. In its most extreme form, it is ludicrous: I don't think the

[3] https://iep.utm.edu/aristotle-logic/

laws of gravity are up to interpretation and change depending on social contexts. Scientific relativism is relevant here because if there are no such things as scientific truths, it is difficult to argue there are such things as scientific progress. Many philosophers and scientists have argued persuasively that scientific progress is occurring (TODO cite) and one only look outside to see the impact progressing science has on society. I will not repeat these defenses, rather I would like to argue even in philosophy where the argument of relativism and lack of progress is even more appealing, data science can unearth evidence of progress being made in the field[4].

Methodology:

While there is a significant amount of statistical analysis that I perform, the main tool is the use of text embeddings. A text embedding is an algorithm that maps words, sentences, or paragraphs into a vector. This vector represents something like the algorithms understanding of what the word means, so vectors that are closer to one another have similar meaning and vice versa. In machine learning, this is established work and most of the field recognizes that this method is a valid and extremely informative way to quantify something like text. A famous example was with the original embedding paper word2vec, which has upwards of 40000 citations: The vector for “king” minus the vector for “man” plus the vector for "woman” is very close to the vector for “queen”, suggesting some sematic understand by the model.

BERT embeddings use the transformer architecture (the architecture 99% of all large language models use) to build even better embeddings and has over 110000 citations. Finally this paper, which is not especially well known because it is well known folk knowledge, points out the LLM embeddings like the Jina AI model that I use, performs the best and better than BERT. One such test that the text embedding capture meaning is if they are predictive of the philosophical content. If I’m writing a book on logic, the style and

[4]Nevertheless, the multidisciplinary critique of science from the humanities and from the sciences to the humanities, helps to sharpen insight and encourage consideration of different views. While I may not agree with all arguments, I believe multidisciplinary back and forth has much to offer all areas of discussion. Gravity may not be a phenomenon that is socially determined, but one can make an argument relativity and quantum mechanics were discovered more quickly due to the social impact of individualism and the revolt against scientific determinism prevalent at the time.

content will be different than if I'm an existentialist writing a novel about the absurdity of life and man's search for meaning.

If the embeddings can predict the subject of a book, that provides evidence that these embeddings are capturing some sort of philosophical content and style. The correlation won't be absolute. Plato writes in a similar style across many different topics so the Gorgias and the Republic may have similar styles but entirely different topics. Meanwhile, Camus and Derrida are both Continental Philosophies and yet have wildly different styles and focus. Additionally, I am taking the content of a single sentence or two and attempting to predict the topic. Many sentences in philosophical works will be somewhat generic. For example, "Philosophy is a noble endeavor" could be a sentence in a book on any topic.

I use the Jina AI embedding model which has good performance on Hugging Face's embedding database, has a setting with relatively low dimensional (32) embeddings, and is not too computationally intensive. All of which are important. The low dimensionality is important because at some point when I run my analysis and my linear regressions, I'm going to overfit. I also don't have tons of GPUs to throw at the problem.

I take all the books I've obtained and use the model to convert the text to embeddings. I embed a couple of sentences to a single representative vector. So, every 2-4 sentences gets its own embedding. The embeddings get thrown into a database with author, book, field, and date as additional data (TODO: link). Performing this analysis, I show that building a prediction engine based on just the embedding in a text has a 50% correlation with the actual topic. We can talk about adjusted R^2 and cluster adjusted R^2s but the data has over 200000 data points and 32 embedding dimensions and so none of these adjustments will change anything about the correlation and R^2. This is reasonable evidence that the predictive power of these embeddings is quite strong. You can find the code here <TODO link code>.

Then I do a couple tests. I use a couple of dimensionality reduction methods – t-SNE, UMAP and PCA. All give similar reasonable results. I think t-SNE and UMAP may work best when the dimensionality is large, and my space is only 32 dimensional.

For the sake of brevity, I will only show the UMAP result. Here is a good description of how UMAP works. Briefly, it tries to maintain the same topological structure from your high dimensional space to the low dimensional space. Points that are clustered will stay clustered to one another and points that are more outliers will stay outliers.

For my analysis, I regress the publication date or a work on the embeddings. This is an attempt to show that there is text that is correlated with dates suggesting at least that tastes have changed from ancient times until now. However, since there are embeddings

more correlated when they come from the same book or author, this intuitively means the degrees of freedom are less than the number of samples since draws are not fully independent. I use Clustered standard errors, which were developed for linear regressions with this issue.

Additionally, I will measure distances between cluster of points. For example, I want to measure the distance between embeddings in ancient texts--written before 219 BCE—compared to embeddings for books written in 1930s+. I do this in two different ways:

1. Measure the average distance between every point in cluster one to every point in cluster two
2. Entropic regularized Wasserstien-2 distance between the two clusters

Wasserstein-2 distance is a metric from optimal transport. The idea is exactly what the name sounds like, suppose you have 10 supply depots and 10 stores. Each store needs supplies from any one depot. The optimal way to supply the 10 stores from the 10 depots is the optimal transport map. If you can measure the distance between any of depots to any of the stores, then the optimal transport distance is the sum of the distances from all the depots to all the stores when transporting optimally. If you use the regular Euclidian distance, this is a Wasserstein-2 optimal transport distance. Since this problem is a complex linear program with many edge cases and difficulty to calculate, entropic regularizations smooth out the edge cases and makes optimization of the distance easier, but at the expense of some inaccuracy after considering the regularization.

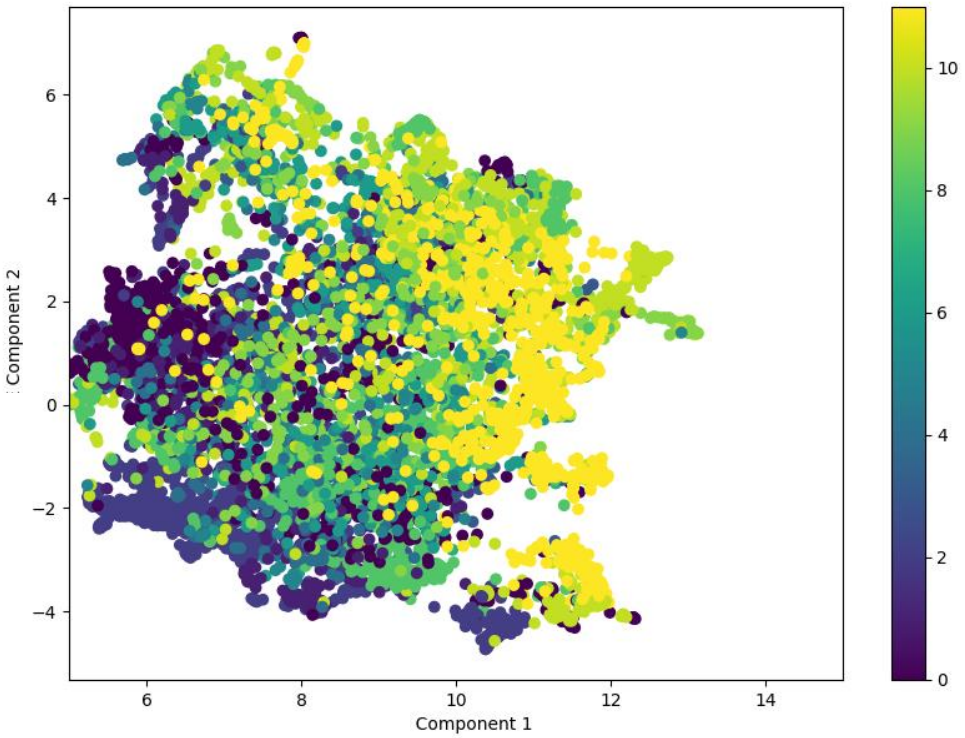
This covers the main methods used in this experiment. Now I will discuss the results that come from this analysis.

Results

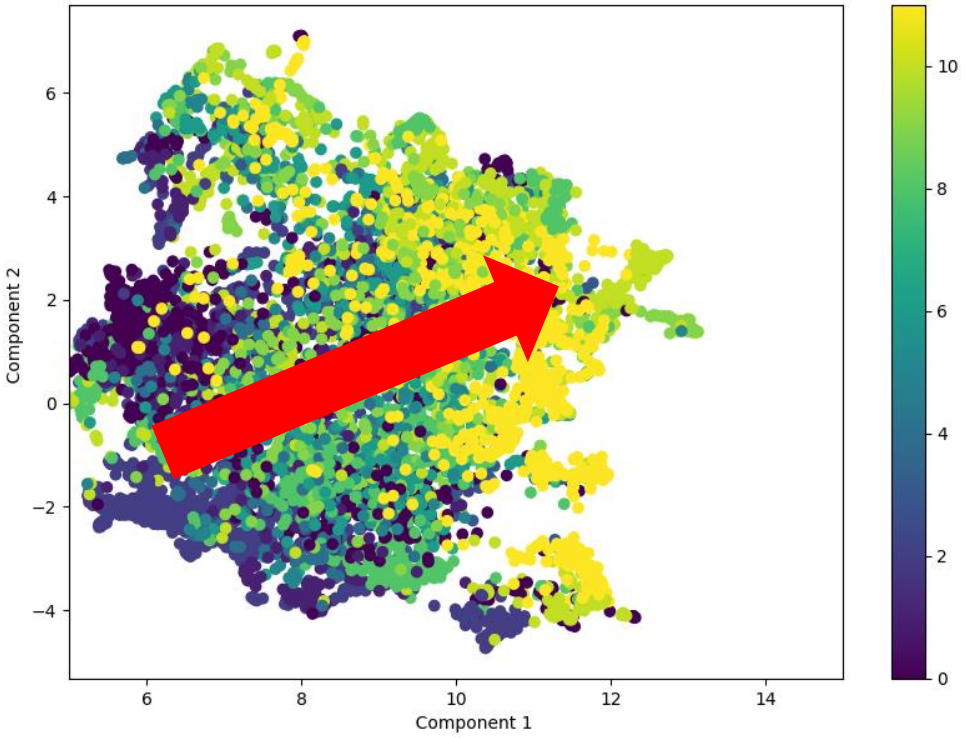
First, I'm going the use the JIRA AI model to embed the documents into word vectors (along with metadata like author, subject...). This data set contains 200000+ vectors. While I do the linear regression and standard deviation analysis on the entire set, for visualization purposes this number is too chaotic. Thus, I average every 25 vectors in each book to reduce the number of datapoints to roughly 8000. This number of points is more manageable from a visualization perspective.

Visualizing the Data

After converting the philosophical texts into embedding I get a dataset that relates these embedding to metadata like date (most important), author, subject, and book. For purposes of visualization, I use UMAP to reduce the dimensionality of the embeddings down to two. Here is the visualization:



The darkest colors indicate the word embeddings for text furthest back in time with progressively lighter colors moving into present day philosophy. You can see a key direction the embeddings move towards as time progresses:



This visualization is not unique. If I use t-SNE with multiple seed values, this trend almost always seems to appear suggesting there is momentum in the style and content of philosophy books. This is the first piece of evidence that progress is occurring in philosophy.

If new philosophy was just a matter of taste, we shouldn't see directionality in the content of philosophy. We should be randomly moving further and closer to ancient philosophy. This directionality only occurs when insights are accepted. We start with Plato and move philosophy in some direction. If those new insights like say from Aristotle are accepted, we can't move back to Plato as those insights have already been mined. The only direction to drive novel insights is away from Plato. Thus, this illustrates that there is an increasing balance of accepted insights over time causing one "direction" of movement away from Plato. I wouldn't read to much into the fact that it's a straight line, but the illustration that it is moving further and further from ancient philosophy is telling. If Aristotle builds off Plato and philosophy only changes due to the caprice of taste, it is no more probable that one returns closer or staying the same distance from Plato, rather than pressure on the content and style of philosophy to move further and further from the ancients. On can argue that this could have been just due to chance, however the likelihood of this happen is extremely

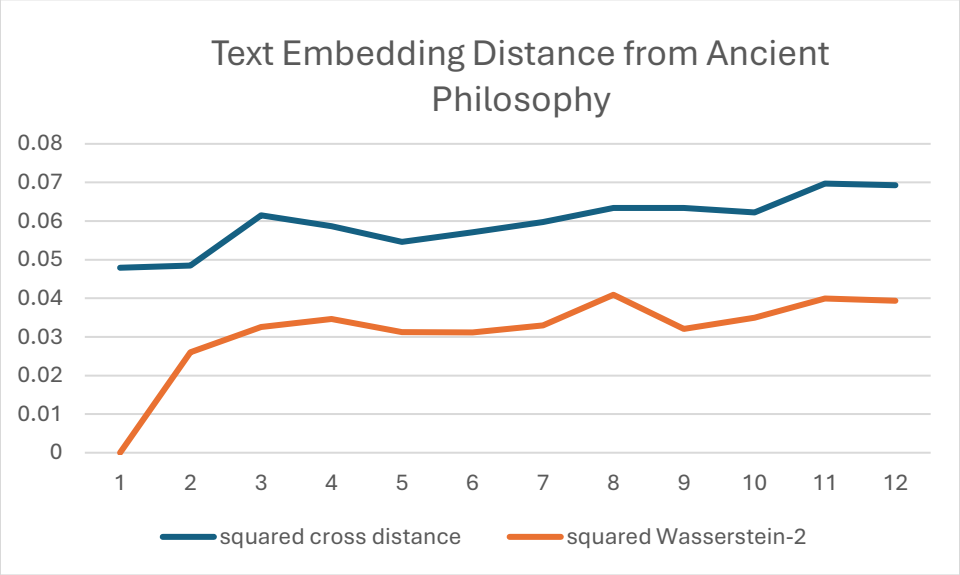
unlikely given the consistent progress of texts moving consistently away from the ancients in semantic embedding space. However, with the next analysis looking at the continual and persistent progress moving further and further away from historical, I can quantify just how unlikely this data would be generated from the random moves regarding changes in taste.

Texts have Become Less Similar to Ancient Philosophy over Time

This part of this analysis is to study just how far different philosophical eras are from the ancients. The first distance I will define is the average distance of every point in one set to every point in another set. I don't know what this is called, but I will call this a cross distance. The second distance is the Wasserstein-2 distance. I split up philosophical texts across 8 bins: The cutoffs are:

$$-\inf, -219, 200, 1400, 1642, 1710, 1796, 1852, 1873.5, 1896, 1910.5, 1930, \inf$$

The buckets contain all the embeddings from any book written between the two cutoffs. As the bin number increases the works get more recent in time. Here is a graph of the distances:



So, the fourth bucket contains all embeddings of text written between 1400 and 1642 and the distance compares all embedding points in a particular bucket to all embedding points in the first bucket. This shows that there is a consistent trend of moving further and further away from ancient philosophy.

If we assume that changes across time of what researchers focus on are simply due to taste, there will be no directionality. Tastes change, but roughly there should be a constant mean which is the center of what philosophy is about and changes in taste shouldn't have a pattern of moving further away from this core. This directionality is indicative of progress.

We can test this effect in a crude way by regressing distance from ancient philosophy using both metrics on the numbers 1 through 12 reflecting increasing values as the texts get more recent. If there is no drift in philosophical content, the parameter on the date term (values 1-12) should be statistically no different than zero. As if we are moving further and further in time, we shouldn't get further and further away from some center. The distance from ancient philosophy should not increase if there is no progress occurring. For cross distance, the probability that the date term is zero or smaller is less than 1% likely (p-value of below 1%) . For Wasserstien-2 distance the date term p-value is 1.2%. Instead of using values 1-12 I also used the midpoints of the bucket endpoint dates with similar results. For my work, see excel document OT notes 1.xlsx.

Additionally I test for stationarity, which argues that not only is the mean constant, but also the standard deviation, and in some definitions all moments are constant. Likewise, if philosophical research has directionality driven by people attempting to find novel insights by building off previous insights, we will be going further and further way from the content and style of ancient texts. This is called non-stationarity. Stationarity is like going into orbit around earth, non-stationarity is like achieving escape velocity and constantly drifting further and further from the earth[5]. This is a stronger criterion than having only a constant mean and using the KPSS test gives similar results. See KPSS.py. I use the KPSS test over the more used Dickey-Fuller test, because for the KPSS test stationarity is the null hypothesis.

Distance from Ancient Philosophy: Analytic Philosophy and Continental Philosophy

If you asked current philosophers which branch of philosophy is best at quantifying progress, I think most would mention analytic philosophy as the most "scientific". Many would have other qualms of the field, but analytic philosophy was constructed so that one defines words precisely and adopts specific logical operations such that the truth/falsity of certain statements is basically logic or math and thus we know for sure, this thing is true,

[5] A random walk in 32 dimensions should increase its variance over time, but it wouldn't be monotonic in increasing its value.

and that thing is false. This is a bit of a caricature of analytic philosophy but gets across the reputation of the field.

Thus, because it uses the scientific method and logic, one can argue that Analytic Philosophy is different than other philosophical branches in its unique ability to make progress. However, this is not borne out in the data. It's not that analytic philosophy makes no progress, but rather that compared to its oft cited counterpart, Continental Philosophy, Analytic Philosophy is not further away from ancient philosophy. In fact, depending on whether you use squared Wasserstein distance or squared cross distance, Continental Philosophy texts could be less like Plato etc. than Analytic Philosophy.

distance to pre 219 BCE philosophy	'Analytic Philosophy'	'Continental'	'Eastern Philosophy'	'Empiricism'	'Epicurean'	'Epistomology'	'Ethics'
OT Distance	0.048993	0.043416	0.039906	0.033125	0.068893	0.026814	0.024091
Cross distance	0.069383	0.071153	0.064562	0.056983	0.079371	0.050816	0.052668

distance to pre 219 BCE philosophy	'History of Philosophy'	'Logic'	'Love and Friendship'	'Metaphysics'	'Philosophy of the Arts'	'Political Philosophy'
OT Distance	0.026697	0.037174	0.025415	0.031365	0.031831	0.024349
Cross distance	0.047982	0.055647	0.041704	0.058708	0.058622	0.05535

distance to pre 219 BCE philosophy	'Pragmatism'	'Pre-Existentialism'	'Psychology'	'Stoicism'	'Transcendentalism'	'Western Religion'
OT Distance	0.039379	0.029342	0.044797	0.026291	0.036457	0.035032
Cross distance	0.06046	0.051342	0.070492	0.046976	0.054086	0.064529

As you can see if using squared Wasserstein-2 distance Analytic Philosophy is further from ancient philosophy and is further than any field except for the Epicureans, which is a small category with very few text embeddings and is likely an outlier. But using cross distance Continental is further than Analytic philosophy and second further to only Epicureanism (which is again an outlier). This shows that date is more important for measuring progress/distance rather than the topic, suggesting that both Analytic and Continental

philosophy are progressing and building off prior work in a systematic fashion. I will note that the Pre-Existentialists (Kierkegaard, Nietzsche and a few other one-off books/authors), did represent a turn back towards Ancient philosophy, however this retreat was small and didn't counteract the trend including all the other work being written up and until that time, and was quickly reversed by the future Continental philosophers who included Sartre and de Beauvoir, as well as "post-existentialists" like Derrida and Foucault.

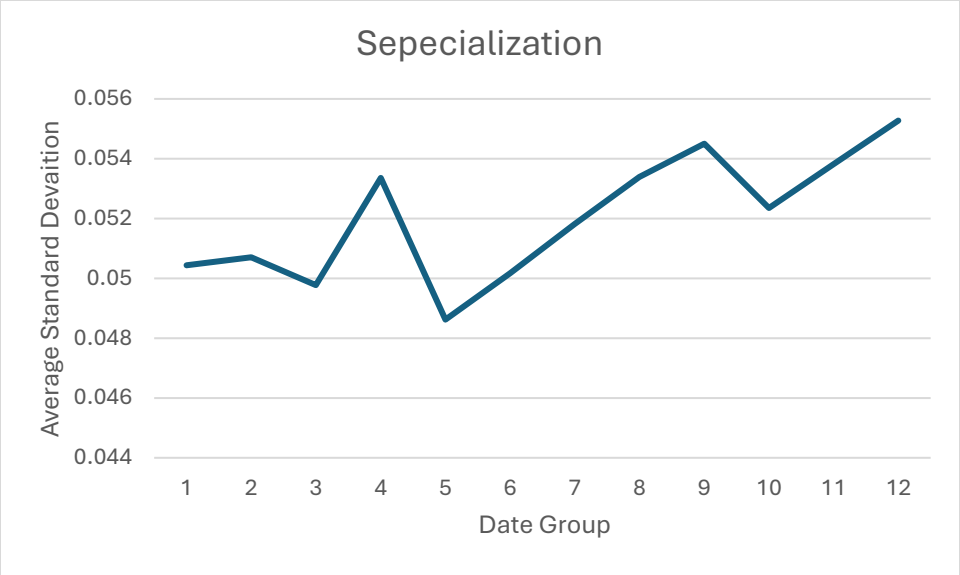
Specialization

The final piece of analysis is about specialization. If the change in content and style of philosophy books over time is due to random preference, we should not see preference towards specialization. I will argue the pressure towards specialization occurs if it gets more difficult to generate novel insights. In addition to moving further and further away from previous content as it gets "mined", one specializes as the general insights applicable to everyone are all acknowledged, and novel insights require better understanding of niches that are unexplored because general knowledge is not good enough.

Specialization is evidence that researchers agree regarding certain general principles, and the specialists specialize because the low hanging fruit is becoming rarer and the lowest hanging fruits are more difficult to pick than in previous generations and can only be picked by people with specialist domain knowledge. Thus, the move towards specialization is indicative of progress.

A vast majority, if not all, fields in the sciences and other fields, where we acknowledge improvement over time, show evidence of specialization, also providing empirical evidence that specialization is a symptom of a field that is progressing.

The analysis is easy. I calculate the standard deviation of each element of my embedding over all 32 dimensions. I then average all the derived standard deviations for each of the 12 date cohorts. Refer to the date cutoffs in the distance section. While the data isn't super conclusive, you can see a general trend of increasing standard deviation of these embedding vectors suggesting that philosophical texts are become less like one another over time and philosophers are specializing:



Conclusion

I have provided suggestive evidence that a field as qualitative as philosophy, one can detect evidence of progress, moving toward greater and greater understanding of what the field seems to consider truths, and building upon and recognizing the important insight of philosophers that came before. Analyzing both the direction that innovation moves towards (and away from), the consistent progress and direction of progress, and the evidence of increased specialization has resulted in suggestive evidence that even a field as subjective as philosophy seems to have fundamental agreements on good and bad philosophy and is progressing towards deeper understanding of what the field acknowledges as "good" philosophy. Ultimately, it would be interesting to apply this kind of text analysis to other fields in the humanities and other applications in philosophy.