
Article:
Cantone, G.G. orcid.org/0000-0001-7149-5213, Zheng, E.-T. orcid.org/0000-0001-8759-3643, Tomaselli, V. et al. (1 more author) (2025) Estimation of disciplinary similarity with large language models. Scientometrics. ISSN: 0138-9130

https://doi.org/10.1007/s11192-025-05385-0

Scientometrics
https://doi.org/10.1007/s11192-025-05385-0

Estimation of disciplinary similarity with large language models

Giulio Giacomo Cantone[1,3] · **Er-Te Zheng**[2] · **Venera Tomaselli**[3] · **Paul Nightingale**[1]

Received: 25 November 2024 / Accepted: 7 July 2025

Abstract
The parameter that captures the similarity among disciplinary categories is a key quantity of many measures of interdisciplinarity. This study evaluates the feasibility of using large language models to estimate this parameter rather than using traditional methods based on citational networks among disciplines. An experimental procedure tested the precision, agreement, resilience, robustness, and explainability of estimates from OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude. The experiment collected a sample of 228 similarity matrices among two disciplinary taxonomies, for a total of 16,200 sampled estimate values. The experiment concludes that Gemini reaches precise estimates, comparable to traditional methods. ChatGPT stands out only for its superior resilience when dealing with semantically trivial changes in how disciplines are described. Claude resulted in a balanced profile. While rarely in full agreement, all three models undertake the estimation task sufficiently well.

Keywords ChatGPT · Claude · Google Gemini · Interdisciplinarity · Estimation · Similarity

✉ Giulio Giacomo Cantone
prgcan@gmail.com

Er-Te Zheng
ezheng1@sheffield.ac.uk

Venera Tomaselli
venera.tomaselli@unict.it

Paul Nightingale
p.nightingale@sussex.ac.uk

[1] Science Policy Research Unit, University of Sussex, Jubilee Building, Falmer Campus, Brighton & Hove, UK

[2] School of Information, Journalism and Communication, University of Sheffield, The Wave, 2 Whitham Road, Sheffield S10 2AH, UK

[3] Department of Economics and Business, University of Catania, Palazzo delle Scienze, Corso Italia, 55, 95129 Catania, Italy

Published online: 12 August 2025

Introduction

Interdisciplinarity can be defined as the integration of "information, data, technique, tools, perspective, concepts, and/or theories" from two or more separated bodies of knowledge to improve understanding or problem solving (Committee on Facilitating Interdisciplinary Research, 2005). Interdisciplinarity also involves collaboration across research areas, cross-citing unrelated scientific journals, and other related phenomena. In particular, IDR has been consistently linked to innovation and technological applications in industry, given its capacity to solve practical problems (Bromham et al., 2016; D'Este et al., 2019; Haeussler & Sauermann, 2020; Larivière & Gingras, 2010; Leahey, 2016; Leahey & Barringer, 2020; Porter & Rafols, 2009; Rafols et al., 2012; van Rijnsoever & Hessels, 2011). Given this widely believed view, the study on the organisation of interdisciplinary research (IDR) has developed an established status as a *research programme* within the field of quantitative studies of science, with clear research pathways. Research questions as: "How to measure interdisciplinarity?", "Does IDR lead to higher scientific impact?", "Is an interdisciplinary education better?" or "Is there a bias against IDR in peer review or funding?", have been explored extensively by dedicated scholars.

Measuring interdisciplinarity has always been a key issue. Given the elusive definitions of IDR, its assessment needs robust measures. Indeed, a series of studies (Fontana et al., 2020; Wang & Schneider, 2020; Zwanenburg et al., 2022) have shown that indicators of interdisciplinarity may not be as mutually coherent as the research community has previously assumed. The findings of a substantial number of studies may overly depend on how the authors define the model of measurement, creating a danger of selective reporting of only those methods that produce outcomes that favour preferred outcomes (Cantone, 2024). Nevertheless, as suggested by Mugabushaka et al. (2016), the pluralism of different quantitative measures of IDR may instead simply reflect the evolution and improvement of theorisation about operative definitions of IDR.

Within the theory of measurement of IDR, a key parameter is the similarity between two disciplinary categories. Already in the original conceptualisation of (Stirling, 2007) and (Porter & Rafols, 2009), the set of references became the canonical metadata used to quantify of the influence of disciplines over a unitary or collective body of research. For example, if half of the references in an article cite journals of Economics, it is deduced that the article recognises Economics for around half of its inspirations. In other words, the metadata of the references tells that such an article must have something to do with the category Economics (Avila-Robinson et al., 2021; Huang et al., 2021; Leydesdorff, 2005; Mutz, 2022; Rousseau et al., 2019; Thijs et al., 2021). The established "Diversity" framework recognises that an article has several inspirations when it cites many disciplines (Variety) in equal proportion (Balance). The third parameter of the formulas of diversity is the similarity (Disparity) among disciplines. If the categories, i.e., the disciplines, are too conceptually adjacent, then the disciplinary diversity of the article is typically tuned down. Intuitively, an article linking two social sciences is less interdisciplinary than one linking a social and a natural science.

The model of measurement based on the references is well-known, reliable, and sensible, and this leads to methods to assess the similarity of a disciplinary taxonomy based on citational networks. Nevertheless, references are only an indirect measure of the effective semantic (*ergo*, textual) content of scientific articles, and an alternative measurement method already employs classifications based on advanced automated techniques from Artificial Intelligence (AI) for Natural Language Processing (Cantone, 2024).

While AI has indeed been employed to classify documents, it has been largely untouched as a source of knowledge about the parameters of similarity among disciplines.

This study is aimed at filling this gap, introducing an experimental protocol that surveys many statistical properties and other qualities of the large language models (LLMs) as a source of numeric values for the similarity parameters among disciplinary categories. LLMs are AIs that are trained to provide useful answers to a wide variety of general questions prompted in natural languages. They are capable of communicating with a human agent, behaving similarly to so-called *chatbots*, but being capable of generalising and generating a much wider range of answers (Mei et al., 2024; Noy & Zhang, 2023; Nejjar et al., 2024; Ray, 2023; Shanahan, 2024). The state-of-the-art of research on LLMs focused on understanding which areas LLMs can replace or complement previous methodologies that require effort or coordination from human agents (Bornmann & Lepori, 2024; Dillion et al., 2023; Gilardi et al., 2023; Jones, 2024; Thelwall, 2024; Zheng et al., 2024). This study stands on the following conjecture: an LLM, asked to provide a numerical estimate for similarity between two disciplines, can return a formally valid response in a number within the unitary range; then must exist a sufficiently detailed question about a finite taxonomy (a 'well-made prompt') capable to generate a whole similarity matrix as its output, such that it approximates sufficiently well results from traditional methods. Finally, while it is rarely easy to understand the criteria used by LLMs to determine the values in the similarity matrix, one could even conjecture that LLM's extensive access to diverse data sources, being capable of estimation, potentially better informed than those derived from citation-based traditional methods.

In other words, a central aim of this study is to assess if estimates obtained by LLM-based methods are not inferior to those achieved through traditional approaches based on citations. Asking questions in natural language does not require computational skill, and has the advantage that it does not require access to a large and potentially very expensive database of citational networks. In this regard, LLM-based methods could be a step towards removing an important entry barrier for researching IDR. However, given the variety of methods and formulas that are used to process matrices of citations into matrices of similarity, both methodologies lack a ground truth parameter to asses their bias, henceforth a strict evaluation of the gain in accuracy of the estimate is impossible to test. Arguably, it could be said that there is no such thing as an 'objective' parameter of similarity, since a precise estimate could still reflect implicit sensitivity to the decisions adopted in the quantification (Marres & de Rijcke, 2020; Rafols, 2019; Stirling, 2023). For this reason, this study will focus on statistical characteristics that define the quality of a sampling procedure, such as the statistical dispersion of estimates among repeated identical trials of queries, or the sensitivity of the output towards alternative spellings of the prompt.

Ideal features for an LLM-based process of similarity estimation are: **Replicability**, the logical and numeric consistency of estimates within and between language models; the **Sensitivity** to semantic changes in the input taxonomy; and the *Explainability* of the model, which is the capacity of an LLM to explain how to relate its internal processes to understandable scientific methodologies. From these principles, five specific statistics can be tested:

1. The *Precision* of the model can be quantified as the inverse of the variance of estimates across identical trial runs as if in parallel. It counts as a test for the *Replicability* of the estimation from the same model. If the same matrix of citations is processed through the same formula, the result ought to be identical. However, the same LLM, at the same

internal parametric state, and on the same state of training, prompted twice with the same prompt for the same taxonomy can provide slightly different estimates. It is important to establish that the variance in estimates is trivial. A potential source of variability in estimates is *temperature*, an internal parameter of the LLM that, once raised, lowers the filtering threshold for source content to contribute to an answer, augmenting the randomness of the answers. Temperature is used in LLMs as a source of randomness to make responses more 'human-like' and less robotic.

2. The *Replicability* of the method also concerns the capacity to reach a logically consistent result independently of which LLM is queried. Different LLMs should be in *Agreement* with each other on the results of any estimation, whereas the statistics can be measured as the difference in the average estimates. These statistics are based on an important epistemological motivation: if different models reach an *Agreement* with low differences on average, it would imply the process of training each LLM draws on evidence from the real world that allows a sufficiently learned LLM to establish reliable scores for the matrix of similarity. In other words, if many LLMs are in *Agreement*, this would count as evidence to support that stable differences (ergo, meaningful parametric similarities) among disciplines exist. On the contrary, if multiple LLMs consistently fail to align in their estimates, it would be reasonable to believe that at least one of them has a bias, even if it is not straightforward to identify which LLM is the biased one.
3. The principle of *Sensitivity* can be synthesised with two simple definitions. *Resilience* is the capacity to differentiate numeric estimates for semantically different categories. A semantic shift in the name of a category should follow a proportionate shift in the estimates of similarity. This capacity is useful because taxonomies of science are mostly outsourced to expert actors who propose names for the disciplinary categories at different levels of granularity.
4. The latter principle holds for the *Robustness* of an LLM-based estimation, which captures the capacity to deal with a shift in the estimates when a category changes its nominal meaning without altering its semantic meaning (e.g., substituting "Computer Science" with "Informatics"). In this case, differences in outcomes should ideally be zero.
5. While LLMs do not offer a way to reproduce their methods, they should at least be capable of outlining the principles behind their assessment of similarity. This is an application of the principle of *Explainability* of an LLM. There is a risk that the software may "hallucinate" an inaccurate response: instead of providing an accurate answer, LLMs sometimes infer an answer from a set of principles that satisfy the query in a generic situation (Farquhar et al., 2024). In such cases, the LLM is trying to appease the agent with a convenient response, but this response could be misleading when attempting to reverse-engineer the LLM's estimation method.

To check to what extent these features hold for LLMs, we developed a prompt and tested it across two taxonomies using three commonly used public LLMs: ChatGPT 4o (CGPT4o), Claude 3.5 Sonnet (C-Sonnet), and Gemini 1.5 Pro (Gemini). We used the taxonomy of five fields of science from the Leiden Ranking (L5F) and a modified version of the disciplinary groups of Clarivate's Journal Citation Report (9S). A control for the temperature parameter has also been included. Given the quality of answers decreased for higher temperatures (which increased randomness), the test of the replicability of the results was exclusively conducted with the temperature parameter at its lowest level, zero. Even with this precaution, some results were unsatisfactory, especially the performance of CGPT4o.

To check applicability, we tested the distribution of the stochastic error of the estimates, and we found that Gemini and Claude outperform CGPT4o in *Precision*; however, the three LLMs are not consistently in *Agreement*. Additionally, we compared estimates for similarity between the categories of 9S among LLM-based methods and citational-based methods, and we concluded that Gemini showed signs to be the most apt LLM for scientometric research, because it consistently approximated traditional estimates based on citations with only a minor bias. To check the *Sensitivity*, we slightly altered the names of the categories. All three LLMs showed good *Resilience*, yet Gemini and C-Sonnet showed excessive *Sensitivity* to trivially altered nomenclatures for the disciplines.

Theoretical background on the measurement of similarity

Similarity is a universal concept covering the semantic space between 'identity' and 'difference'. With these two concepts, similarity shares the features of always being referred to as a pair of objects (X, Y). A singular object can be similar to the average of its group, and a group can be, on average, highly or lowly similar, but all of these features are derived by operations on pairs of elements (Leydesdorff, 2005; Rao, 1982; Tversky, 1977). These characteristics of the concept of similarity are reflected in the two fundamental rules of any operative definition for a z measure of similarity:

- Unitary range:

$$0 < z(X, Y) \leq 1; \forall (X, Y)$$

- Identity of the indiscernibles:

$$z(X, X) := 1 \tag{1}$$

One may be led to think that Eq. (1) implies that

$$z(X, Y) = 1 \iff X \equiv Y \tag{2}$$

but this is not strictly implied in Eq. (1).

From this set of rules, two macro-approaches to advance the definition of similarity emerge.

A structural definition of similarity

A structural definition of similarity tries to capture the underlying essence of a comparison between a pair of objects. It defines the similarity of two objects using a measuring procedure: to see if two objects are identical or different, it involves measuring, element by element, the extent of the congruence between the two juxtaposed objects (Markman & Gentner, 1996; Willett, 2014). This definition is captured by the formula of "Intersection over Union" of two sets:

$$\underset{Jac}{z}(X, Y) := \frac{|(X \cap Y)|}{|(X \cup Y)|} \tag{3}$$

In this formula, X and Y are binary vectors $(\mathbf{x}, \mathbf{y})$ of the presence or absence of features. This is also known as the Jaccard Index.

A limit of any formalisation of the principle of structural similarity is that in many cases, X and Y are not the underlying objects, but the stylised representations of those objects, e.g., vectors, etc. Inevitably, the information encapsulated in X, Y, etc. can be coarse or imprecise, and the format of representing objects as categorical sets can also be limiting. As a result, the structural approach has gradually evolved towards evaluating measures of similarity between vectors. For example, in cases where features i are traced through scales of values, between 0 and 1 for X, i.e., for $x_i \in (0 : 1)$, Eq. (3) can be generalised (under the name Fuzzy Jaccard) as follows:

$$\underset{fJac}{z}(\mathbf{x}, \mathbf{y}) := \frac{\sum_i[\min(\mathbf{x}, \mathbf{y})]}{\sum_i[\max(\mathbf{x}, \mathbf{y})]} \tag{4}$$

As a very general case for $x_i \in \mathcal{R}$ Eq. (3) is extended as follows:

$$\underset{Tani}{z}(\mathbf{x}, \mathbf{y}) := \frac{\sum_i(x_i \cdot y_i)}{\sum_i(x_i^2) + \sum_i(y_i^2) - \sum_i(x_i \cdot y_i)} \tag{5}$$

which can be recognised with the name of Tanimoto Index (Petković et al., 2021; Willett, 2014).

While both Eq. (4) and are generalisations of Eq. (3) and the three equations converge for x_i defined as a binary (presence of absence of the feature), they do not converge for $x_i \in (0 : 1)$, i.e. for fuzzy scales. This may be a reason for the proliferation of alternative structural similarity measures in the literature.

Tversky (1977) tried to comprehend this variety by generalising Eq. (3):

$$\underset{Tver}{z}(X, Y) := \frac{|(X \cap Y)|}{|(X \cap Y)| + \theta_x|X - Y| + \theta_y|Y - X|} \tag{6}$$

The limit of Eq. (6) for $\theta_x = \theta_y = 1$ Eq. (6), converges to Eq. (3), since a union of two sets can be decomposed into their intersection and their two mutual set differences:

$$(X \cup Y) = \bigcup[(X \cap Y), (X - Y), (Y - X)] \tag{7}$$

In addition, Eq. (6) also converges to the Dice-Sorensen Index for $\theta_x = \theta_y = .5$:

$$\underset{Dice}{z}(X, Y) := \frac{2 \cdot |(X \cap Y)|}{|X| + |Y|} = \frac{2\sum_i(x_i \cdot y_i)}{\sum_i(x_i^2) + \sum_i(y_i^2)} \tag{8}$$

A functional definition of similarity

Equation (8) can be interpreted as the divergence that occurs as a consequence of decisions to opt for Y instead of X. This interpretation differs from a structural definition of similarity, and it is closer to a functional definition. A structural definition of similarity reduces objects to their essential elements and then compares them. A functional definition of similarity is concerned with differences in outcomes, in the context where the objects operate. Functional similarity can be defined as the fungibility of two options, which captures how much one can act as a substitute for the other, independently of their physical congruence (Hahn et al., 2003; Medin et al., 1993). The relevance and value of this alternative perspective on similarly has increased with the growing use of new Data Science methods

Fig. 1 Two urns filled with the same amount of red and blue balls, shuffled differently. Generated by *DeepAI*

within their traditional areas of applications concerning decision-making in Finance, Management, etc., and also increasingly in applications in the natural sciences (classification, artificial intelligence, etc.).

This difference in the definition of similarity can be understood through many examples. A simple and pertinent one is the following: two urns contain the same number of blue and red balls, are shuffled in two different ways, see Fig. 1. A structural approach will detect that the positions of the balls are mutually independent, and therefore conclude that the two urns are as different as possible. A functional approach will detect that the probability of drawing a blue ball is identical for each urn, so the two urns are functionally identical.

From this example, other formulas for measuring functional similarity can be derived. These require specifying objects in terms of probabilities of outcomes rather than vectors:

$$P(X) : \{p(x_{i=1}), p(x_{i=2}), \dots)\}$$

So the functional similarity between two objects would result in the intersection of the masses of the two distributions. Consider the following formula:

$$\underset{\text{eF1}}{z}(X, Y) := \frac{2 \cdot \sum_i [p(x_i) \cdot p(y_i)]}{\sum_i p(x_i)^2 + \sum_i p(y_i)^2} \tag{9}$$

Equation (9)works as an extension of real numbers of Eq. (8). Yet, for x defined on a binary scale, Eq. (9) converges to the F1 score, that is, the harmonic mean between precision and recall of a test of binary classification[1].

[1] In fact, assuming that X is the ground truth, and that Y is the variable for the positive outcomes (i.e. $y_i = 1$ means that i is test-positive), hence

$$\text{Precision}(Y) := \frac{\sum_i x_i \cdot y_i}{\sum_i y_i}$$

and

A benefit of adopting a functional definition is that one can assess the similarity of two distributions indirectly, deriving a complementary or inverse formula from the divergence between the two. Typical measures of divergence, like the χ^2 or the Kullback-Leiber not have a form in $(0:1)$ and are not defined for $p(x) = 0$. The Hellinger Divergence, instead

$$\underset{Hell.}{d}(X, Y) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{p(x_i)} - \sqrt{p(y_i)}\right)^2} \tag{10}$$

is constrained in $(0:1)$, so one can adopt its complement to estimate functional similarity:

$$\underset{Hell.}{z}(X, Y) := 1 - \underset{Hell.}{d}(X, Y) \tag{11}$$

To conclude, both perspectives can legitimately be used to justify formulas to capture similarity. A structuralist definition of similarity is closer to an objective one, because if two physical objects are materially identical, then it follows that they are perfectly fungible for any purpose. Nevertheless, there are some small benefits from adopting a functional definition of similarity. One, for example, is the general heuristic that consequences are often easier to sample correctly and treat mathematically. In many applications, tracing consequences is much less expensive than tracing objective features, possibly because there is an indefinite number of features that can be traced about objects, and missing one could bias the estimation of structural similarity. Instead, if a consequence is unobserved, likely, its consequences are not very relevant, because if it were, it would be known and observed. On the other hand, it may hold that the structural similarity $z(X, Y)$ is equal to $z(X, Y')$, but the 'different pieces' ('missing pieces') may be different, and differently relevant for X. For example, a car missing a seat (or with a replaced seat) can still work properly, while a car missing wheels (or with damaged wheels) cannot. The functional definition of similarity works as a useful *ad hoc* that is transparent in how the evaluated terms $p(x)$ are stylised. Finally, one can notice that the functional definition requires fewer assumptions than the structural. As Tversky noticed, a structural definition must assume symmetry in similarity:

Footnote 1 (continued)

$$\text{Recall}(Y) := \frac{\sum_i x_i \cdot y_i}{\sum_i x_i}$$

Then it follows

$$\text{F1}(X, Y) = \frac{2}{[\text{Precision}(Y)]^{-1} + [\text{Recall}(Y)]^{-1}} = \frac{2 \cdot \sum_i x_i \cdot y_i}{\sum_i x_i^2 + \sum_i y_i^2}$$

The F1 score is a typical indicator of 'fidelity' between a condition (X) and an experimental observation (Y). It may satisfy the definition of fungibility, in the sense that one can affirm that if it is sufficiently high, one is allowed to treat the positive cases (Y) as a ground truth (X) without an high hazard; in other words, with *little* consequences, independently by how exactly the testing procedure is capable to infer the condition. A difference between F1 and Eq. ((9) consists in the latter considering the possibility that ground truth cannot be defined on clear-cut binary outcomes (Goutte & Gaussier, 2005). As a consequence, the testing procedure results in more uncertain outcomes. Another difference is that in F1 i indexes the observations while in Eq. ((9) i indexes the outcomes of X, constrained by the definition of probabilities $\sum P(X) = 1$. Therefore, even for functional similarity, one can instead resort to a different operative definition for z, considering, for example, Fuzzy Jaccard (Eq. 4) or Tanimoto (Eq. 5).

$$z(X,Y) = \hat{z} \iff [z(Y,X) = \hat{z}]$$

leading actors to accept Eq. (1) as true. This is not needed for a functional definition. An object can be fungible for another but not *vice versa*.

The role of similarity in quantitative studies on interdisciplinarity

Scientific disciplines are iconic and well-recognisable branches of knowledge. A discipline is not just a set of linked ideas: disciplines also have specialised methods and a particular way of organising the professional life of experts (Becher, 1981; Becher, 1994; Börner et al., 2012; Bu et al., 2021; Hodgson & Donald, 2022; Jacobs & Frickel, 2009; Sugimoto & Weingart, 2015; Stichweh, 1992). Considering only the few general branches of science, scientific activities are sufficiently well defined by the conjunction of an object of enquiry plus a methodological tradition. This approach leads to simple distinctions, such as "Theoretical sciences" vs. "Applied", or "Natural" vs "Social" (Fanelli & Glanzel, 2013). At higher definition, more subdivisions are recognised as relevant, and the concepts of transmission and certification of knowledge are more emphasised.

Despite the growing interest in IDR from science-policy experts, measuring IDR is not straightforward. As Cantone (2024) demonstrated, there are many paradigms of measurement. Disciplinary similarity emerged as an influential factor for measuring disciplinary diversity. Across paradigms, the considered taxonomy of i disciplines is stylised as:

$$\mathcal{I} : \{i_i, i_2, \dots, i, j, \dots i_k\}$$

Henceforth, the formalism $p_i(x)$ symbolises the proportion of i in the considered metadata (e.g. the references, etc.) of the unit of analysis x.

A canonical formula of disciplinary diversity is then the Rao-Stirling index Stirling (2007):

$$\Delta_2(x) = \sum_{(i,j)} p_i(x) \cdot p_j(x) \cdot [z(i,j) - 1] \tag{12}$$

where in this case (i, j) is the placeholder formalism for all the couplets of elements of $\mathcal{I}$. $[z(i,j) - 1]$ is the component of 'disparity' in the equation (Shu et al., 2022; Zhang et al., 2016), and it also appeared in alternative parametric formulas for diversity (Mutz, 2022; Leydesdorff, 2018; Leydesdorff et al., 2019; Wang et al., 2017). The disparity factor can be easily equated as the complement of the similarity between i and j disciplines, that is $z(i, j)$ in Eq. (12).

Similarity between disciplines: measures and meaning

The component of Disparity has been introduced across the paradigms of measurement of IDR to correct the implicit assumption that the knowledge and activities embodied within disciplinary categories are uniformly close to one another. This assumption is rarely justifiable: as previously mentioned, at a general level, natural sciences are clustered together,

In other words, the role of coefficients of similarity is to correct the measurement of IDR (e.g *via* Eq.12) for the effective conceptual proximity between the couples of disciplines involved in the metadata of the unit of analysis. In light of this, one might question

what the notion of similarity between disciplines entails. For example, one could notice that, in the unfortunate occurrence of a leg injury, the diagnosis of the radiologist is always followed by the intervention of the orthopedist, and finally by the rehabilitative therapy of the physiatrist. So it could be stated that these sub-disciplines of Medicine are similar because they work together. This conception falls under the case of structural similarity, because by being frequently collaborative and synergetic one to each other, they realise a synergy towards a unique achievement, the total rehabilitation of the patient, that would not be possible without any of the specialists. On the contrary, similarity could originate from mutual fungibility, or even rivalry. Various examples can be found among the humanities and social sciences (e.g. "is behavioural theory a substitute for classical rational theory of economic actor?"). But other examples can be found in technology (e.g., statistical methods *vs.* machine learning). This approach considers the function of the discipline.

This debate would not be solved if not by looking in detail at the common methods of quantification of $z(i, j)$. The works of Leydesdorff (2005) Adnani et al. (2020), Huang et al. (2021), and Shu et al. (2022) document a variety of well-established approaches to estimate disciplinary similarity. Some of these approaches can be relatable to the formal theory presented in Section "Theoretical background on the measurement of similarity", others potentially expand it with new insights. The core feature of virtually all the established approaches is the assumption that the relevant metadata to measure disciplinary diversity is the network of citations among articles published in journals ascribed as highly relevant for the elicited taxonomy of disciplines. The steering principle is that the more two disciplines cite each other (network proximity), the more similar they are.

Let

$$c_{i\to j}$$

be the sum of all the citations from articles published in journals linked to discipline i towards articles linked to j (outward links). Pairwise, by inverting the direction of the arrow, the quantity

$$c_{i\leftarrow j}$$

represents the sum of citations received by j from i (inward links). Given a $\mathcal{I}$ taxonomy, from these formalisms can be derived vectors of disciplines citing others

$$\underset{\rightarrow}{\mathbf{i}} : \{c_{i\to j_1}, c_{i\to j_2}, \dots\}$$

and of the disciplines being cited by others

$$\underset{\leftarrow}{\mathbf{i}} : \{c_{j_1\to i}, c_{j_2\to i}, \dots\}$$

Together, these compose the matrix of citations C, that is, the adjacency matrix of the considered network of journals.

From the vectors of C, one can quantify the similarity among disciplines, for example, by applications of the methods outlined in Section "Theoretical background on the measurement of similarity". This study establishes another formula in the specific literature of IDR. The first is the Salton's cosine:

$$\hat{z}(i_1, i_2) := \cos(\mathbf{i_1}, \mathbf{i_2}) = \frac{\sum_c [c(i_1) \cdot c(i_2)]}{\sqrt{\sum_c [c(i_1)]^2} \cdot \sqrt{\sum_c [c(i_2)]^2}} \tag{13}$$

which can be understood as a translation of the Bravais-Person coefficient of linear correlation, which always preserves the sign of its inputs, i.e., for positive vectors it will always be a positive number (Egghe & Leydesdorff, 2009).

Peculiarly, all the considered formulas account for inward and outward citations disjointly, i.e., they can be applied to pairs of citing disciplines or cited disciplines. This distinction has been proven not to matter empirically for the specific estimation of similarity, since typically for large networks the two numbers tend to coincide. Nevertheless, it can be remarked that the similarity of the references (outward links) stands as a structural similarity in metadata. In this case, two disciplines are similar because they share a high intersection of shared ideas. On the contrary, the similarity in being cited (inward links) highlights the functional fungibility of the two disciplines. To conclude, there are also methods of quantification that consider jointly the sums of inward and outward citations, for example, the Ochiai method:

$$\underset{Ochiai}{z}(i,j) = \frac{c_{i\to j}\cdot c_{j\to i}}{\sqrt{(\sum \underset{\to}{\mathbf{i}} + \sum \underset{\leftarrow}{\mathbf{i}})\cdot(\sum \underset{\to}{\mathbf{j}} + \sum \underset{\leftarrow}{\mathbf{j}})}} \tag{14}$$

Materials

At the beginning of the experimental procedure, a textual prompt was written and repeatedly tested without saving its results until a final version was established. The prompt aims to output a code that runs in the language R and generates a matrix of scores of similarity for the input of the list of disciplines. Two taxonomies of disciplines have been elicited; these fit in the final version of the prompt. The fitted prompts are then imputed to the LLM through packages that connect the R Studio software with the APIs (Application Programming Interface) of the LLM. Finally, R-Studio receives the answer, executes the code, and saves the output matrix of similarities as a stand-alone object. The process has been repeated many times with the final version of the prompt to simulate a proper process of sampling and resampling of outcomes. Outcomes from these multiple trials are compared to similarities estimated through traditional methods presented in Section "Theoretical background on the measurement of similarity".

The considered taxonomies are derived by the "Leiden Five Fields" taxonomy[2] or L5F, and from the disciplinary groups of Clarivate's Journal Citation Report[3] (9S). The tested LLMs are: ChatGPT 4o ("CGPT4o), Claude 3.5 Sonnet ("C-Sonnet"), and Google Gemini 1.5 Pro ("Gemini").

Prompt

The prompt was submitted to the LLMs through the automated connection to their respective APIs, using wrapper software to pipe the results directly to *R Studio*. For C-Sonnet we used the `ClaudeR` package, for Gemini we used `gemini.R`, and for CGPT4o `tidychatmodels`. Redundancy in the instructions of the prompt is justified by the need to

[2] https://www.leidenranking.com/information/fields

[3] https://jcr.clarivate.com/jcr/browse-categories

automate the generation of the similarity matrix into a *R* object. The prompt was progressively improved through many rounds of trials to reduce formal and substantial errors in the format of the response, but only results from the final version of the prompt have been employed in the analysis of the study. There is text between two curly brackets in the prompt, which indicates how the taxonomies have been fitted in the prompt.

Prompt
You are an expert in quantitative studies in science
Your task is to provide an R script that, when executed, will generate a symmetrical square matrix object in R with $\{\{\mathcal{I}^2\}\}$ elements. Hence, the length () of the matrix that will be generated by the script must be equal to $\{\{\mathcal{I}^2\}\}$
The matrix should not be assigned to an object. It should run on a script, not on a markdown, which means no backticks should be found in your answer
The content of the matrix is estimates of the disciplinary similarity among the following categories:
{{CATEGORIES of $\mathcal{I}$}}
Requirements:
• The similarity values should be between 0 and 1
• The matrix should be symmetrical along the diagonal. It means that once evaluated, the code must be TRUE for isSymmetric()
• The similarity values on the diagonal should all be 1
• Approximate the real similarity values as closely as possible
• Keep three decimal places for all values
• The code must generate a matrix of $\{\{\mathcal{I}^2\}\}$ elements
• In the code for the matrix, the names for rows and columns must be the categories that I provided
• Be sure that the dimension name generating the matrix reports "dimnames = list(c(...), c(...))" where "..." are {{CATEGORIES of $\mathcal{I}$}}
Before providing your final answer, think through the problem step-by-step. Consider the following:
• The nature of research in each category
• The methodologies commonly used
• The overlap in subject matter
• The frequency of interdisciplinary collaborations
• The similarity in publication venues and citation patterns
After considering all pairs of categories, present your final similarity matrix in the following format:
It must be a code in R that generates a matrix object that represents your estimates
It must start with a matrix(c(1.000... and contain no "\n" within the code
Remember to name correctly dimnames: rows and columns correctly. The dimnames should be {{CATEGORIES of $\mathcal{I}$}}
You are not allowed to generate a code with no dimname. You are not allowed to name dimnames as "V1" or rows "1", or with similar numeric references
The matrix must be squared and symmetrical, made of $\{\{\mathcal{I}^2\}\}$ elements. Less than $\{\{\mathcal{I}^2\}\}$ elements is not allowed. More than $\{\{\mathcal{I}^2\}\}$ elements is not allowed
The output must consist exclusively of R code. Do not output anything that is not R code
Do not open your message with phrases like "Here is the..." or "Sure...". All the content of your answer must be exclusively R code

Table 1 Leiden five fields (L5F)

Field	Label	Alternative
Biomedical and health sciences	BHS	Biomedical sciences
Life and Earth sciences	LES	Life sciences
Mathematics and Computer Science	MCS	Mathematics
Physical sciences and engineering	PSE	Engineering
Social sciences and humanities	SSH	Social sciences
Statistics of the sampling algorithm		
Number of LLM tested		3
Number of temperatures tested		3
Unique couples of fields ($k(i, j)$)		10
Number of equivalent estimates for original combinations		24
Number of equivalent estimates for alternative combinations		6
Total number of runs		108
Total number of estimates drawn		3, 240

Sampling similarities in the Leiden Five Fields

The Leiden Five Fields taxonomy (Table 1) is inspired by Waltman and van Eck (2012) and Traag et al. (2019). It is ideal for a preliminary glimpse of the precision of and the agreement among LLMs since it is synthetic but well-representative of the main divisions among sciences, coupling disciplines into only five fields of research.

For each of the 3 LLMs, a similarity matrix was drawn 18 times to fit the prompt to the original five fields: 6 times at temperature = 0, 6 times at temperature = .5, and 6 times at temperature = 1. By enforcing $z(i,i) := 1$ (see Eq. 1) and $z(i,j) = z(j,i)$ there are only $\binom{5}{2} = 10$ unique combinations of fields with stochastic similarity.

For each LLM, the sampling algorithm has been run across 5 other iterations, with each iteration drawing 18 matrices (6 for each temperature). The difference for these additional 5 iterations is that in each iteration, the name of one field has been altered into an alternative form (see Table 1) to test the *Resilience* of the estimates to trivial alterations. These alternative categories remove one part of the name of the original field, as a substantial, not only nominal, change in the terms of referenced disciplinarity.

The final result is that the similarity $z(i, j)$ of each unique couple of L5F has been sampled 24 times for each combination per LLM and temperature, while combinations involving alternative categories have been sampled only 6 times per LLM and temperature. In total, 3, 240 estimates have been drawn for the L5F taxonomy.

Sampling similarities in the disciplines of Journal Citation Report

Clarivate's Journal Citation Report associates a large list of scientific journals to a taxonomy of more than 200 "subject categories", organised in 21 disciplinary groups. From these groups, we derived a taxonomy of only 9 subjects (9S). In the 9S taxonomy, some categories are identical to a group, some are merges of groups, and some groups are ignored, see Table 2.

A taxonomy of 9 subjects combines $\binom{9}{2} = 36$ unique couples; for this reason, instead of iterating six different variations of the taxonomy, the algorithm was run for only two types of iterations: once with the original categories of 9S, and once with the alternative names instead of their original (e.g. Life Science vs. Biology). If substantial differences in estimates are expected by altering L5F, we expect that alternatives for 9S should not deviate the estimates as much. In 9S, the disciplinarity is better defined, and the proposed modifications in the wording of the disciplines do not alter the semantic reference of the categories. Hence, this time the alteration counts as a test of the *Robustness* of the LLM; in particular, the deviations between original and alternative names should be lower for 9S than L5F.

In this case, for each LLM, the similarity matrixes have been drawn 99 times by fitting the prompt on the original 9S taxonomy: 33 times at temperature = 0, 33 times at temperature = .5, and 33 times at temperature = 1, for a total of 10, 692 estimations. In addition, the iteration with the alternative names has been run 7 times for each LLM and temperature, for a total of 2, 268 estimations on alternative combinations.

Table 2 Nine disciplinary subjects (9S)

Clarivate's group	Category	Label	Alternative name
Agricultural Sciences	Biology	BIO	Life Science
Arts & Humanities			
Biology & Biochemistry	Biology	BIO	Life Science
Chemistry	Chemistry	CHE	
Clinical Medicine	Clinical Medicine	MED	Health Science
Computer Science	Mathematics & Computer Science	MCS	Mathematics & Informatics
Economics & Business	Economics & Business	ECB	
Engineering	Engineering & Materials Science	EMS	
Environment/Ecology	Biology	BIO	Life Science
Geosciences	Geology	GEO	Earth Science
History & Archaeology			
Literature & Language			
Materials Science	Engineering & Materials Science	EMS	
Mathematics	Mathematics & Computer Science	MCS	Mathematics & Informatics
Multidisciplinary			
Philosophy & Religion			
Physics	Physics	PHY	
Plant & Animal Science	Biology	BIO	Life Science
Psychiatry/Psychology	Psychology & Social Sciences	PSS	Human Sciences
Social Sciences	Psychology & Social Sciences	PSS	Human Sciences
Visual & Performing Arts			

Statistics of the sampling algorithm	
Number of LLM tested	3
Number of temperatures tested	3
Unique couples of categories ($k(i, j)$)	36
Number of equivalent estimates for original combinations	33
Number of equivalent estimates for alternative combinations	7
Total number of runs	120
Total number of estimates drawn	12, 960

Citation-based methods of similarity

The result of sampling similarity from the LLM has been compared to estimates of traditional methods considered in Section "Theoretical background on the measurement of similarity". Since the established assumption is to quantify the similarity of disciplines from a network of citations, we accessed the Journal Citation Report database for the year 2023. This is a network of all citations among journals indexed with one or more of Clarivate's Group. The C adjacency matrix of this network is presented in Table 3. Journals with multiple associations across the 9 Categories are counted once for each category.

For each couple of categories $i \neq j$, 13 estimates have been considered from seven estimators (Table 4). Six estimators of these have been applied twice, once for the columns and once for the rows of C, while this distinction does not matter for the Ochiai. In three

Table 3 Cross-citations among 9S categories

Citer	Cited category								
Category	BIO	CHE	MED	ECB	EMS	GEO	MCS	PHY	PSS
BIO	12, 337, 674	1, 502, 395	3, 032, 856	936, 409	983, 026	257, 796	244, 477	53, 175	149, 909
CHE	1, 290, 528	5, 307, 020	252, 689	47, 334	1, 543, 809	197, 877	110, 568	267, 028	8, 375
MED	2, 514, 098	141, 590	10, 825, 769	91, 331	75, 182	3, 408	149, 676	35, 655	461, 103
ECB	534, 253	52, 141	123, 724	2, 299, 166	143, 066	36, 196	231, 546	6, 170	417, 533
EMS	804, 319	2, 025, 145	183, 718	199, 542	5, 286, 851	117, 437	717, 523	559, 538	30, 479
GEO	165, 901	273, 146	7, 999	36, 110	88, 397	567, 985	23, 852	22, 714	8, 168
MCS	180, 487	82, 938	156, 559	222, 188	643, 488	17, 268	2, 701, 321	109, 473	108, 173
PHY	51, 685	185, 815	37, 663	5, 231	378, 003	23, 622	105, 137	1, 698, 720	5, 973
PSS	131, 280	6, 126	548, 930	351, 896	15, 526	7, 236	109, 581	5, 777	1, 747, 167

Table 4 Summary of considered estimators of similarity on the citational network

Estimator	Scale	Equation
Cosine of vectors	Sum of citations	Eq. (13)
Generalised Jaccard (Tanimoto)	Sum of citations	Eq. (5)
Dice-Sorensen	Sum of citations	Eq. (8)
Fuzzy Jaccard	Relative proportions of citations	Eq. (4)
Extended F1	Relative proportions of citations	Eq. (9)
Hellinger	Relative proportions of citations	Eq. (11)
Ochiai	Sum of citations	Eq. (14)

cases, the estimator requires normalizing the vectors of C to their proportions instead of absolute methods (4).

Results

The temperature of LLM influences bias and precision of the similarity scores

The temperature of the LLM has a slight, but not trivial, effect ($\beta_{Temp} \neq 0$) on both the estimation of the generic $z(i, j)$ similarity score, see Table 5. The column $\bar{z}(i,j)$ is the average z among the $k = 10$ combinations of i and j. $n = 24$ is the number of estimates for each combination. For the L5F taxonomy, arguably the effect of increasing the

Table 5 Effect of temperature on the estimation of similarity

Taxon	Model	Temp	n	$k(i,j)$	$\bar{z}(i,j)$	$p(z \mid \beta_{\text{Temp}} = 0)$	$\bar{s}_{i,j}(z)$	p(Levene)
L5F	C-Sonnet	0	24	10	0.41	0.01	0.04	0.93
	C-Sonnet	0.5	24	10	0.42		0.05	
	C-Sonnet	1	24	10	0.44		0.06	
	Gemini	0	24	10	0.29	0.48	0.04	~ 1
	Gemini	0.5	24	10	0.29		0.05	
	Gemini	1	24	10	0.28		0.05	
	CGPT4o	0	24	10	0.55	0.76	0.06	0.17
	CGPT4o	0.5	24	10	0.54		0.10	
	CGPT4o	1	24	10	0.55		0.10	
9S	C-Sonnet	0	33	36	0.36	~ 0	0.02	0.06
	C-Sonnet	0.5	33	36	0.43		0.07	
	C-Sonnet	1	33	36	0.40		0.07	
	Gemini	0	33	36	0.20	~ 0	0.03	~ 0
	Gemini	0.5	33	36	0.23		0.07	
	Gemini	1	33	36	0.23		0.06	
	CGPT4o	0	33	36	0.56	~ 0	0.07	0.5
	CGPT4o	0.5	33	36	0.59		0.07	
	CGPT4o	1	33	36	0.61		0.09	

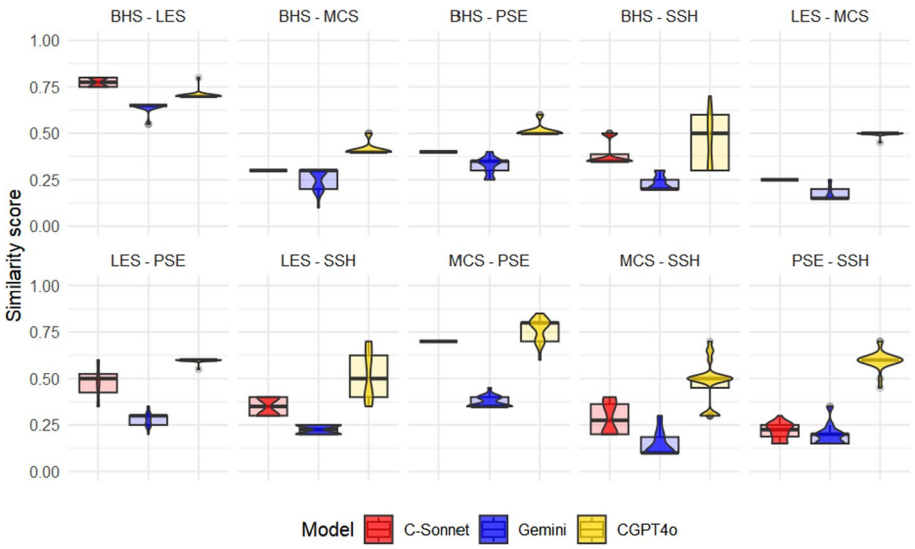

Fig. 2 Estimates of similarity are represented with a box-violin method. The bold line is the median, and the box is the interquartile range; however, a mirrored kernel curve ("violin") of the density of the estimates is over-imposed over the box. A flat line is an ideal result in terms of *Precision* of the LLM estimation

temperature from 0 to 1 is negligible, nevertheless, using a model that corrects for the fixed effect of the 10 groups of combinations, the β_{Temp} effect is always statistically significant ($p(z \mid \beta_{\text{Temp}} = 0)$ in Table 5) for t-Student tests across the 9S taxonomy[4].

As expected, temperature has also a consistent positive impact on the variance of the estimates; this can be noticed by observing the column $\bar{s}_{i,j}(z)$, which measures the average value across the standard deviation of the estimates within each combination of categories i and j. We tested this impact with Levene's test of homoskedasticity across temperature values on the pool of estimates per combinations of taxonomy and model, i.e., not grouped for the combinations of categories, differently from the previous test on the average. The results are statistically significant only for the combination of Gemini and 9S ($p \sim 0$). The high p-values on the tests to reject homoskedasticity paired with the consistent increase in standard deviation at higher temperatures, suggest that raising the temperature above 0 has only a small negative influence on the precision of the estimation.

We restricted the replicability analysis to only estimates generated with Temperature set at 0, as it is the most reliable, and reducing randomness will provide the highest precision of outcomes (lower variance). Sensitivity analysis is based on average differences between original and traditional nomenclatures, and is conducted without removing the estimates generated with higher temperature, to preserve a larger sample size.

[4] The formula in R to reproduce the t-test is the following: `fixest::feols(formula = z ~ Temp | Categories, data = .)`

Replicability

At *Temp.* = 0 C-Sonnet and Gemini have a sufficiently good *Precision*, except when social sciences are involved in the estimations, see Figs. 2 and 3.

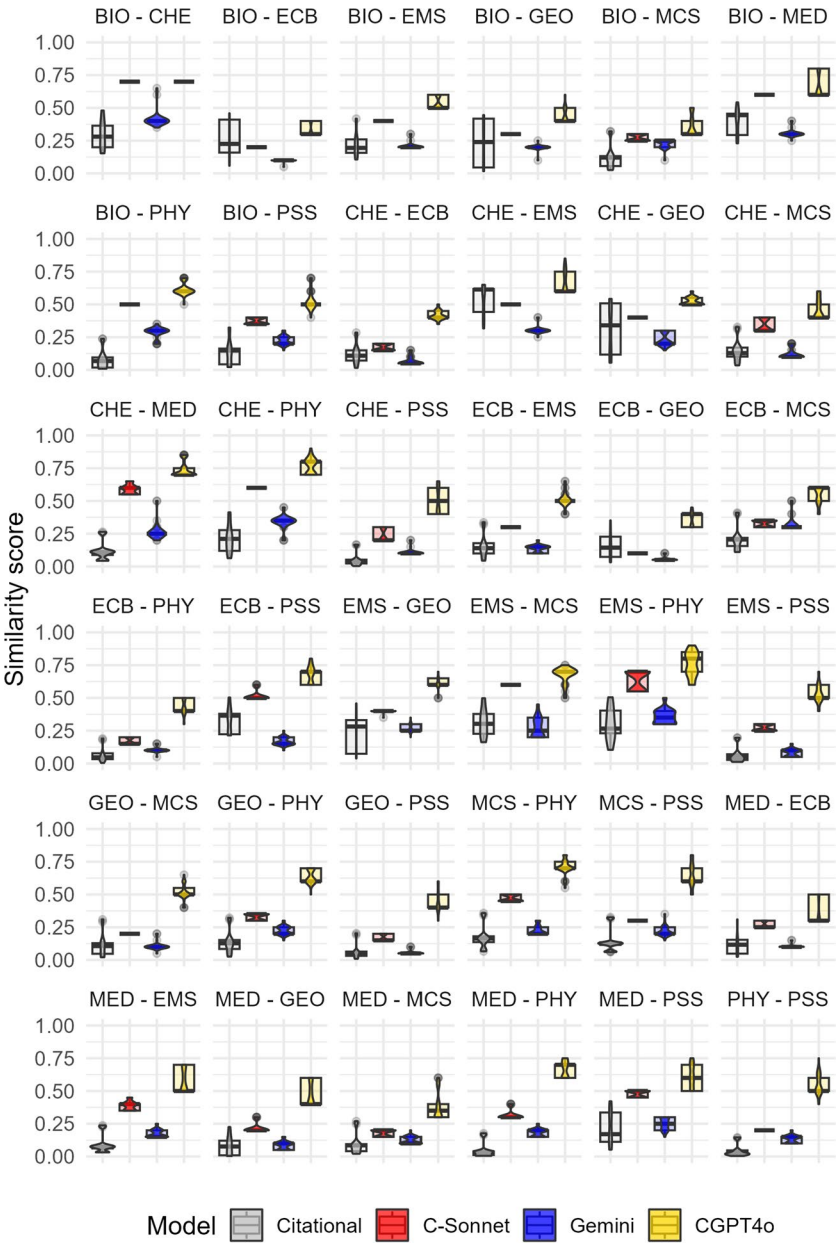

Fig. 3 The bold line is the median, the box is the interquartile range, and the mirrored kernel curve ("violin") is the density of the estimates. A flat line is an ideal result in terms of *Precision* of the LLM estimation

Nevertheless, different LLMs do not always reach an agreement. For L5F, this is particularly noticeable for all combinations of the five fields except when the field of Biomedical and Health Sciences (BHS) is involved. Adopting the 9S, it is possible to assess when an LLM agrees with citation-based methods. Gemini is consistently the closest LLM to these approaches. This fact alone does not necessarily assure that Gemini is objectively the best of the three models. Indeed, it provides evidence that Gemini internalised some form of "knowledge" about the scores generated by citation-based methods. 9S ChatGPT 4o is consistently the farthest from the scores generated by citation-based methods. This result should be considered along with the fact that in the 9S taxonomy, the scores are considerably lower than for L5F, and in some cases, ChatGPT 4o seems to end up closer to the centre of the scale .5, a behaviour consistent to what is seen in Fig. 2. This could be (weak) evidence that ChatGPT 4o is just hallucinating numbers to complete the task it has been given.

Resilience

A test for *Resilience* is conducted with the estimates for L5S. The average is used as an estimate for the similarity when the fields are reported exactly as they are *vs.* in the reduced alternative are compared in Table 1.

While for Biomedical and Health Sciences (BHS) and Life and Earth Sciences (LES) results are mixed, the results for Physical Sciences and Engineering (PSE) and Social Sciences and Humanities (SSH) are strong evidence of all three LLMs being sufficiently capable to operate with a correct *Sensitivity* towards meaningful distinction among disciplinary categories. When "Physics" is prompted instead of "Physical Sciences and Engineering", the similarities with all the other fields are coherently reduced, except for CGPT4o showing a small increase of similarity between Mathematics and Computer Science (see "Alt PSE" in Table 5). This behaviour is consistent with a form of LLM 'understanding' that, in general, Science, Technology, Engineering and Mathematics (STEM) disciplinary subjects are more interdisciplinary than a specialisation in Physics. This becomes even more evident when we notice that by removing "Humanities" from SSH, the similarity increases (see "Alt SSH" in Tab 1 with natural sciences. The LLMs recognise that the abstract idea of 'social sciences' is closer to natural sciences when it is severed from its connection with the Humanities. Given the prominence of books in the outputs of the humanities and their generally weaker coverage in citation datasets, the positive value of using LLMs to capture similarity is notable.

Robustness

Of the 60 combinations of models and alternatives for the L5F, the average absolute deviation (the average length of the bars in Fig. 4) is equal to .054; for 9S this equals 0.067 (see Table 6). These statistics indicate that the tested LLMs are very sensitive and not very robust to trivial changes in the nomenclature. Trivial alternative names at a lower granularity induce deviations in the similarity that are not significantly inferior to substantial changes. Looking at Fig. 5, it may seem that this effect is clustered around problematic couples of disciplines (e.g. Medicine and Engineering, see Alt MED and EMS in Fig. 5), but in reality even the median absolute deviation for L5F (.39) is not

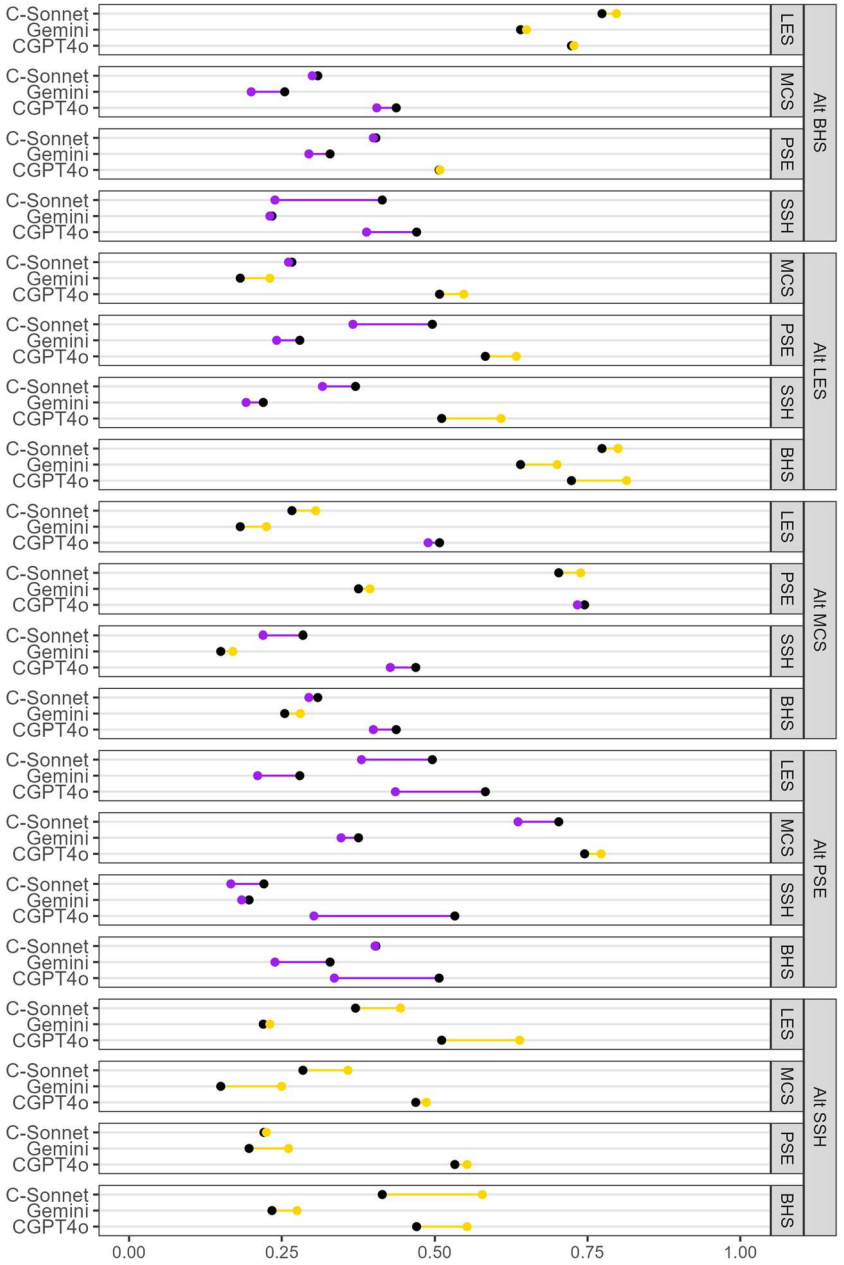

Fig. 4 Check of resilience: substantial alternatives. The black dot is the average estimate for the unaltered nomenclature; when the bar is yellow, the alternative has a higher average similarity; when the bar is purple, lower

Table 6 Mean absolute difference between alternative nomenclatures

Sample	In L5F	In 9S
All	.054	.067
C-Sonnet	.057	.073
Gemini	.040	.064
CGPT4o	.066	.064

higher than the median for 9S (.5). Curiously, after disaggregating this diagnostic statistic across the three models, again the outlier is CGPT4o, even if the decrease is too small to be deemed statistically significant ($p = .46$).

Explainability

We asked ChatGPT, Gemini, and Claude directly how they define "similarity" and "discipline similarity". Answers are provided in Table 7.

They all regarded similarity as "the degree to which two (or more) entities share characteristics, properties, or meaning", a definition that aligns with how humans conceptualise similarity. C-Sonnet provides a pure structural definition of similarity. CGPT4o mentions the possibility of similarity having to do with a functional relationship, recognising that similarity depends on the context of measurement. Gemini stands out for providing a form of methodology for measurement, primarily as structural, also mentioning that the "meaning" of the feature is context-dependent.

C-Sonnet defines disciplinary similarity as the structural similarity ("overlap") of two "academic or professional" fields, based on "interactions". This definition corresponds, although vaguely, to the cognitive definition of similarity in IDR. Curiously, C-Sonnet explicitly recognises the multi-dimensionality of the task of measuring disciplinary similarity (Cantone, 2024). CGPT4o is vaguer and does not provide a very useful evaluation, repeating that "convergence" must stand for scopes, principles, methods, and social structures. Gemini provides examples but seems to harmonise the previous two perspectives. All three LLMs seem to favour a structural definition of disciplinary similarity based on the overall attributes. At this stage, LLMs may derive the attributes from insights closely paralleling humans' intuitions, nevertheless, none of these insights seem to explicitly grasp the finesse of scientometric methods to establish the disciplinary similarity.

Discussion

This study contributes to address the question: "Should researchers employ large language models to estimate disciplinary similarity?". The findings suggest an affirmative answer. Apart from the potential bias present in certain models and their generally limited robustness, the study did not find relevant obstacles in the use of LLMs for estimating similarity matrices. In light of the encouraging outcomes observed here, future developments in this area may enhance all core aspects for the effective estimation of disciplinary similarity through LLMs.

The most obvious benefit of adopting this AI-driven approach is the lower barriers to access. For example, referencing the specific operations of the study, the retrieval of the

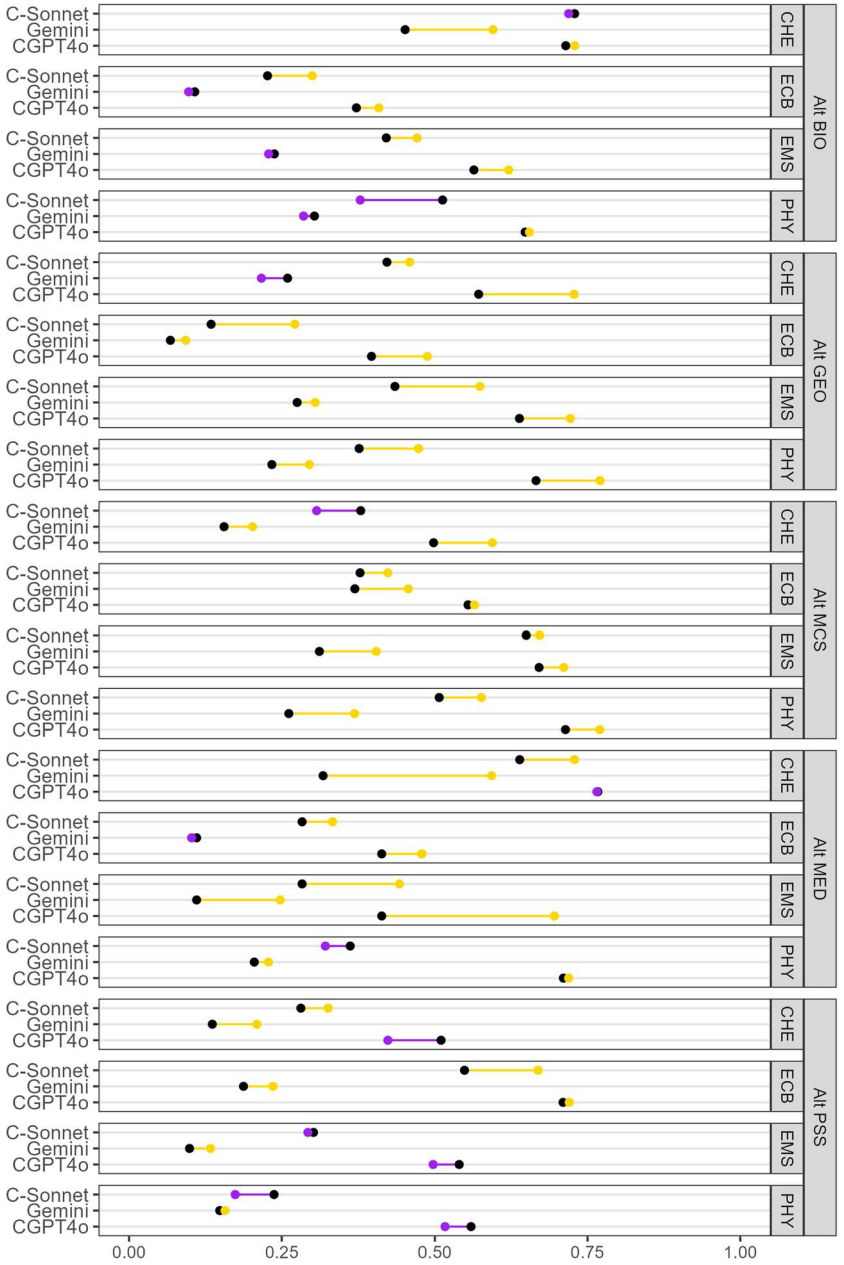

Fig. 5 Check of robustness: nominal alternatives. The black dot is the average estimate for the unaltered nomenclature; when the bar is yellow, the alternative has a higher average similarity; when the bar is purple, lower

network of citations in the Journal Citation Report required a specific right of access, which is typically granted to academic institutions, but not to independent researchers. Then, given the dimension of the network (around 11 gigabytes of data for the table of citations),

Table 7 Definition of similarity and disciplinary similarity by LLMs

Model	Similarity
C-Sonnet	The degree to which two or more entities share common characteristics or properties. The inverse of the difference or distance between entities in a defined space of comparison
CGPT4o	The extent to which two entities (objects, concepts, situations, etc.) exhibit comparable properties, behaviours, or relationships within a specific context. These properties could be physical, functional, relational, or abstract, depending on what is being compared
Gemini	The degree of shared meaning between two entities. This means: • Extracting Key Features: I identify the core elements that define each entity. These could be words, concepts, relationships, patterns, or any other meaningful characteristic • Comparing Features: I assess the overlap and divergence between the extracted features. The more shared features and the stronger their connection, the higher the similarity • Context Matters: The meaning of features can change depending on context. I consider the specific situation and purpose when evaluating similarity
Model	**Disciplinary similarity**
C-Sonnet	The degree of overlap and connection between two academic or professional fields based on their fundamental characteristics and interactions. This similarity can be quantified through a multi-dimensional analysis of key components
CGPT4o	The extent to which two disciplines converge in their objectives, underlying principles, research methods, knowledge structures, and societal contributions. In essence, it reflects how much overlap exists in the content and approaches of two fields of study or practice
Gemini	The degree of interconnectedness between two fields of study. This encompasses: • Shared Foundational Concepts: Disciplines built upon similar core principles and theories are more likely to be similar. For example, physics and engineering share foundational concepts in mathematics and mechanics • Overlapping Methodologies: Fields utilizing comparable research methods, analytical techniques, or problem-solving approaches exhibit higher similarity. For instance, sociology and anthropology often employ qualitative research methods • Interdisciplinary Connections: Disciplines frequently borrowing ideas, collaborating on research, or addressing shared problems indicate stronger similarity. An example is the overlap between neuroscience and psychology in studying human behaviour

data must be pre-processed with dedicated languages for data querying, then exported into a software for data analysis. Roughly speaking, with the adoption of high-performing laptops, the estimation of similarity with citational methods required four hours of computations between pre-processing and export in *SQL*, and processing in *R*. On the contrary, the generation of a similarity matrix from an LLM through the API requires very few minutes, and the whole experiment on LLMs consisting of a total of 228 queries costs less than 15$.

The results of this study imply that in the long run, the technology of LLMs will allow the following achievements:

1. A singular query to an LLM for a matrix of similarities can be an alternative to traditional methods based on citations. Aside from the aforementioned gains in the reduction of computational labour, there is also a conceptual benefit: typically, authors do not retrieve a full network of citations to estimate the disciplinary diversity. Instead, they use the network emerging from their sample of articles for the estimations. In other words, the same sample is used for both internal parameterisation of a model and inference through the model. This is not an advisable analytical practice because it may induce a systemic error in the inference. LLMs will always provide information out-of-the-sample

where statistical error is independent from the systemic error of the sampling process of the data. Nevertheless, this substitution relies on specific assumptions that the LLM is precise and robust. This study did not identify such a perfect candidate among the surveyed LLMs, but with rapid development, it could emerge in future studies.

2. A different approach would be to average a singular value across estimates sampled from different LLMs, for each pair of disciplines. In this case, the most important decision regards how to establish the weights of the average to correctly assess the researcher's confidence in the outputs of different LLMs. A simple solution could be to set weights proportional to the precision of the LLM. An even better weighting scheme would account for both the precision and the robustness of LLM, for the specific pair of disciplines.
3. Finally, the information of citational networks and LLM can be combined. Aside from just averaging scores, a more advanced method would involve the concept of Bayesian updates of the scores. This would require estimating the parameters of a Beta distribution from the scores generated by LLM, multiplying these for the Likelihood function of the estimates of similarity through traditional methods, and finally normalizing them to retain the variance in the scale between 0 and 1. From the updated distribution, it is possible to identify the maximum (i.e., the mode of the posterior distribution), which is the most reliable value for the similarity score.

Aside from the promising application of this proposal, this study holds important limitations that should be acknowledged and possibly challenged in future studies. A first limitation regards the number of surveyed LLMs. While ChatGPT, Claude, and Gemini are among the most acknowledged, alternatives emerged in AI's Grok and the European Mistral's Le Chat. Chinese brands also produced high-quality LLMs as DeepSeek or Alibaba's Qwen. In addition, all of these have fast cycles of development so the experiment presented in this study should be replicated, including more and newer versions of the models.

Another limitation of the study regards the dimension of the similarity matrix. The study tested taxonomies with five to nine disciplines, but taxonomies of ten to twenty-five disciplines are common, too. Of course, the size of the matrix of similarities is the square of the taxonomy, and even enforcing symmetry as the prompt does, a query for all the combinations of 25 disciplines is requesting no less than 600 estimates for each trial. This could stress the capacity of the LLMs to format correctly consistently over trials, and be a technical bottleneck in the process of automation of sampling. Solutions to smooth the process of resampling in future experiments could include pre-formatting a list of combinations of disciplines, instead of querying a whole matrix, at the cost of higher expenses in API tokens, i.e., higher computational costs on the side of LLM side. It is hypothesised that a test on a higher number of disciplines could much better assess the effective *Resilience* of the LLM.

To conclude, beyond the evident limitations discussed, one may raise a deeper question: what exactly is the internal mechanism within LLMs that connects a request for a similarity matrix to a numerical output? We propose a metaphor to illustrate our two main conjectures. Imagine a humanoid robot seated at a restaurant table. A waiter brings out the chef's finest dishes, which the robot ingests, making them disappear into its mechanical mouth. The robot is then asked to provide an objective evaluation of the culinary experience. One possibility is that the robot has developed its own sense of taste, perhaps through some form of synthetic sensory apparatus. If the robot's

judgments closely align with human evaluations, it could suggest a degree of objectivity in the concept of "flavor". Conversely, if the robot produces unusual judgments, what we might call "hallucinations", this could mean that its apparatus detects aspects hidden from human perception instead. Its assessments might not predict human taste reliably, but the information provided could offer novel perspectives. And since the idea of a universally objective taste remains debatable, it would be unfair to label the robot's responses as simply inaccurate. In this study, we metaphorically aimed to assess how well the robot can replicate human taste (*Agreement*), and how consistent it is with its own past evaluations, rather than offering arbitrary responses (*Precision*).

Alternatively, the robot might not have developed any real sensory ability. Upon identifying the dish, it could simply draw on its knowledge base of expert opinions and deliver an average judgment. When applied to similarity estimation, this kind of "cheating" would actually be ideal–since it would mean that the LLM is simply saving us time by retrieving highly specific, relevant information from existing sources. Qualitative analyses, maybe in the form of dialogic surveys should be welcomed, to prove this conjecture. Evidence for this second hypothesis is in the excellence of Gemini in being in *Agreement* with traditional methods. Gemini might be accessing an extensive body of scientific knowledge for training through Google Scholar. Possibly, within this body of knowledge, there are papers published in scientometric journals, reporting estimates for similarity scores. Then exactly as the robot that just summarised the opinions of past experts, by drawing on this knowledge, the LLM could try align itself with past literature. Gemini proved itself to be the advanced LLM for the task of similarity estimation by rarely failing at formatting output in the phase of prototypation of the prompt, and by being capable to explains its methods in detail too. None of these results should be considered definitive due to the inherently dynamic nature of the comparative performances among brands of LLMs.

Acknowledgements This study is an original idea of Giulio Giacomo Cantone, who performed all the quantitative analyses of the study and curated the mathematical background of the article. The prompt is an adaptation of Giulio Giacomo Cantone and Er-Te Zheng, from an original prototype of Er-Te Zheng. Er-Te Zheng conducted qualitative analysis and contributed overall at all stages of the theorisation, literature review, and discussion of the article. Venera Tomaselli provided the conceptualisation of the statistical procedures. She curated the design of the experiment and the writing of the manuscript. Paul Nightingale contributed by reviewing literature on interdisciplinarity and large language models, and by managing permissions in collecting data on citations. We kindly thank Dr. Josie Coburn for her essential guidance in retrieving data for citational methods for the 9S taxonomy. We also thank Prof. Luca Martino of the University of Catania for insightful conversations on future developments of this application. Er-Te Zheng thanks the School of Information, Journalism and Communication of University of Sheffield for funding the GTA Scholarship. Cantone and Nightingale are grateful for the financial support of the Schmit Foundation and the ESRC's Metascience programme.

Declarations

Conflict of interest The authors of this manuscript have no conflict of interest.

References

Adnani, H., Cherraj, M., & Bouabid, H. (2020). Similarity indexes for scientometric research: A comparative analysis. *Malaysian Journal of Library and Information Science, 25*(3), 31–48. https://doi.org/10.22452/mjlis.vol25no3.3

Avila-Robinson, A., Mejia, C., & Sengoku, S. (2021). Are bibliometric measures consistent with scientists' perceptions? The case of interdisciplinarity in research. *Scientometrics, 126*(9), 7477–7502. https://doi.org/10.1007/s11192-021-04048-0

Becher, T. (1981). Towards a definition of disciplinary cultures. *Studies in Higher Education, 6*(2), 109–122. https://doi.org/10.1080/03075078112331379362

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education, 19*(2), 151–161. https://doi.org/10.1080/03075079412331382007

Bornmann, L., & Lepori, B. (2024). The use of ChatGPT to find similar institutions for institutional benchmarking. *Scientometrics*. https://doi.org/10.1007/s11192-024-05039-7

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature, 534*(7609), 684–687. https://doi.org/10.1038/nature18315

Bu, Y., Li, M., Gu, W., et al. (2021). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology, 72*(5), 523–539. https://doi.org/10.1002/asi.24433

Börner, K., Klavans, R., Patek, M., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464. https://doi.org/10.1371/journal.pone.0039464

Cantone, G. G. (2024). How to measure interdisciplinary research? A systemic design for the model of measurement. *Scientometrics*. https://doi.org/10.1007/s11192-024-05085-1

D'Este, P., Llopis, O., Rentocchini, F., et al. (2019). The relationship between interdisciplinarity and distinct modes of university-industry interaction. *Research Policy, 48*(9), 103–799. https://doi.org/10.1016/j.respol.2019.05.008

Dillion, D., Tandon, N., Gu, Y., et al. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology, 60*(5), 1027–1036. https://doi.org/10.1002/asi.21009

Committee on Facilitating Interdisciplinary Research. (2005). *Facilitating interdisciplinary research*. National Academies Press.

Fanelli, D., & Glanzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLoS ONE, 8*(6), e66938. https://doi.org/10.1371/journal.pone.0066938

Farquhar, S., Kossen, J., Kuhn, L., et al. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Fontana, M., Iori, M., Montobbio, F., et al. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy, 49*(7), 104063. https://doi.org/10.1016/j.respol.2020.104063

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *Advances in information retrieval* (pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Haeussler, C., & Sauermann, H. (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy, 49*(6), Article 103987. https://doi.org/10.1016/j.respol.2020.103987

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition, 87*(1), 1–32. https://doi.org/10.1016/S0010-0277(02)00184-1

Hodgson, G. M., & Donald, T. (2022). Campbell on the institutions of scientific knowledge and the limits to interdisciplinarity. *Journal of Institutional Economics, 18*(6), 969–980. https://doi.org/10.1017/S1744137422000121

Huang, Y., Glänzel, W., Thijs, B., Porter, A. L., & Zhang, L. (2021). The comparison of various similarity measurement approaches on interdisciplinary indicators. Working Papers of ECOOM—Centre for Research and Development Monitoring 670612, KU Leuven, Faculty of Economics and Business (FEB), ECOOM—Centre for Research and Development Monitoring

Jacobs, J. A., & Frickel, S. (2009). Interdisciplinarity: A Critical Assessment. *Annual Review of Sociology, 35*(1), 43–65. https://doi.org/10.1146/annurev-soc-070308-115954

Jones, N. (2024). AI now beats humans at basic tasks—New benchmarks are needed, says major report. *Nature, 628*(8009), 700–701. https://doi.org/10.1038/d41586-024-01087-4

Larivière, V., & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126–131. https://doi.org/10.1002/asi.21226

Leahey, E. (2016). From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual Review of Sociology, 42*(1), 81–100. https://doi.org/10.1146/annurev-soc-081715-074219

Leahey, E., & Barringer, S. N. (2020). Universities' commitment to interdisciplinary research: To what end? *Research Policy, 49*(2), Article 103910. https://doi.org/10.1016/j.respol.2019.103910

Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology, 56*(7), 769–772. https://doi.org/10.1002/asi.20130

Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics, 116*(3), 2113–2121. https://doi.org/10.1007/s11192-018-2810-y

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics, 13*(1), 255–269. https://doi.org/10.1016/j.joi.2018.12.006

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition, 24*(2), 235–249. https://doi.org/10.3758/BF03200884

Marres, N., & de Rijcke, S. (2020). From indicators to indicating interdisciplinarity: A participatory mapping methodology for research communities in-the-making. *Quantitative Science Studies, 1*(3), 1041–1055. https://doi.org/10.1162/qss_a_00062

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Mei, Q., Xie, Y., Yuan, W., et al. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences, 121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Mugabushaka, A. M., Kyriakou, A., & Papazoglou, T. (2016). Bibliometric indicators of interdisciplinarity: the potential of the Leinster-Cobbold diversity indices to study disciplinary diversity. *Scientometrics, 107*(2), 593–607. https://doi.org/10.1007/s11192-016-1865-x

Mutz, R. (2022). Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? An information-theoretical and statistical estimation approach. *Scientometrics, 127*(12), 7397–7414. https://doi.org/10.1007/s11192-022-04336-3

Nejjar, M., Zacharias, L., Stiehle, F., et al. (2024). LLMs for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process, 37*(1), e2723. https://doi.org/10.1002/smr.2723

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Petković, M., Škrlj, B., Kocev, D., et al. (2021). Fuzzy Jaccard Index: A robust comparison of ordered lists. *Applied Soft Computing, 113*(107), 849. https://doi.org/10.1016/j.asoc.2021.107849

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719–745. https://doi.org/10.1007/s11192-008-2197-2

Rafols, I. (2019). S &T indicators in the wild: Contextualization and participation for responsible metrics. *Research Evaluation, 28*(1), 7–22. https://doi.org/10.1093/reseval/rvy030

Rafols, I., Leydesdorff, L., O'Hare, A., et al. (2012). How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy, 41*(7), 1262–1282. https://doi.org/10.1016/j.respol.2012.03.015

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24–43. https://doi.org/10.1016/0040-5809(82)90004-1

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy, 40*(3), 463–472. https://doi.org/10.1016/j.respol.2010.11.001

Rousseau, R., Zhang, L., & Hu, X., et al. (2019). Knowledge integration: Its meaning and measurement. In W. Glanzel, H. F. Moed, & U. Schmoch (Eds.), *Springer handbook of science and technology indicators* (pp. 69–94). Springer.

Shanahan, M. (2024). Talking about large Language models. *Communications in ACM, 67*(2), 68–79. https://doi.org/10.1145/3624724

Shu, F., Dinneen, J. D., & Chen, S. (2022). Measuring the disparity among scientific disciplines using Library of Congress Subject Headings. *Scientometrics, 127*(6), 3613–3628. https://doi.org/10.1007/s11192-022-04387-6

Stichweh, R. (1992). The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science. *Science in Context, 5*(1), 3–15. https://doi.org/10.1017/S0269889700001071

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*. https://doi.org/10.1098/rsif.2007.0213

Stirling, A. (2023). Against misleading technocratic precision in research evaluation and wider policy—A response to Franzoni and Stephan, uncertainty and risk-taking in science. *Research Policy, 52*(3), Article 104709. https://doi.org/10.1016/j.respol.2022.104709

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation, 71*(4), 775–794. https://doi.org/10.1108/JD-06-2014-0082

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*. https://doi.org/10.2478/jdis-2024-0013

Thijs, B., Huang, Y., & Glänzel, W. (2021). Comparing different implementations of similarity for disparity measures in studies on interdisciplinarity. Working Papers of Department of Management, Strategy and Innovation, Leuven 670614, KU Leuven, Faculty of Economics and Business (FEB), Department of Management, Strategy and Innovation, Leuven

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies, 1*(1), 239–263. https://doi.org/10.1162/qss_a_00011

Wang, X., Wang, Z., Huang, Y., et al. (2017). Measuring interdisciplinarity of a research system: Detecting distinction between publication categories and citation categories. *Scientometrics, 111*(3), 2023–2039. https://doi.org/10.1007/s11192-017-2348-4

Willett, P. (2014). The calculation of molecular structural similarity: Principles and practice. *Molecular Informatics, 33*(6–7), 403–413. https://doi.org/10.1002/minf.201400024

Zhang, L., Rousseau, R., & Glanzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology, 67*(5), 1257–1265. https://doi.org/10.1002/asi.23487

Zheng, E. T., Fu, H. Z., Thelwall, M., & Fang, Z. (2024). *Can tweets predict article retractions? A comparison between human and LLM labelling* https://doi.org/10.48550/arXiv.2403.16851

Zwanenburg, S., Nakhoda, M., & Whigham, P. (2022). Toward greater consistency and validity in measuring interdisciplinarity: A systematic and conceptual evaluation. *Scientometrics, 127*(12), 7769–7788. https://doi.org/10.1007/s11192-022-04310-z