# The Philosophy Engine

Automating Philosophical Discovery
Through Literature, Argument, and Counterargument

Matthew Bonci  Lehan Zhang  David Zollikofer

*This is the named edition of the methodology report submitted to the AI Philosophy Competition. Essay titles, theses and texts are withheld until the competition's results are announced.*

**Explore the interactive site** lehanzhang.com/research/ai-philosophy
**Watch the video** lehanzhang.com/research/ai-philosophy/video

Every stage of our pipeline is anchored in the published literature. A script organised 32,685 articles from 24 philosophy journals into 185 literature clusters by their shared references and ranked them. AI models then mapped selected debates from their full texts and proposed candidate theses, either from inside a single debate (within-cluster generation, our baseline) or by combining two debates that rarely cite each other (cross-cluster generation, a variation designed to elicit less conventional ideas). Each candidate thesis was screened, developed into a memo, written up and revised under independent AI referees and an AI editor, and finally compared in paired games with published papers from the same debate, including hidden papers its author never saw. Several papers passed the AI editor; AI referees then reviewed them, the AI referee reviews determined the selection, and the team confirmed it, choosing one paper per debate. All three come from the within-cluster method. Humans designed the process and fixed its rules; AI models proposed the ideas, wrote every sentence, and reviewed and judged the papers.

## Contents

**1 The entry at a glance** 3

**2 The literature as foundation** 3

**3 Within-cluster generation: the baseline** 5

**4 Cross-cluster generation: a variation** 7

**5 From candidate thesis to paper** 9

**6 Judging against published work** 11

**7 The three submitted papers** 12

**8 Who did what** 13

**9 Resources** 13

**10 Models, harnesses and reproducibility** **13**

**A Selected prompts** **15**

**B The judge's scoring guide** **16**

# 1 The entry at a glance

Our pipeline is built around one question: where in the current literature is there room for a new and defensible contribution, and can an AI system develop such a contribution into a paper that holds its own against published work in the same debate? The three submitted papers are its answer.

The three papers are labelled A, C and K throughout. Their titles, debates and abstracts are withheld until the competition's results are announced. The letters are the papers' fixed identifiers throughout the submission. Table 5 in Section 7 summarises each paper.

Figure 1 shows the whole pipeline. It has three layers. The *literature layer* turns the journal literature into debates that can be worked on: clusters of articles that cite the same work, ranked for how promising they are, with the full texts of their members (Section 2). The *generation layer* proposes candidate theses. Its baseline, within-cluster generation, maps a single debate and looks for its unfinished business: open problems, unanswered objections and principles in tension (Section 3). Its variation, cross-cluster generation, combines items from two debates that rarely cite each other and subjects each combination to seven gates (Section 4). The *paper layer* takes every candidate thesis, whatever its origin, through screening, a memo, a paper and up to two rounds of revision under independent AI referees and an editor, and then compares the accepted paper with published papers (Sections 5 and 6). Three design commitments run through all three layers.

- *Grounded in the literature.* Debates are found in the citation record, theses are proposed from maps of the texts, quotations are checked word for word, and novelty is tested against the library and public indexes.
- *Adversarial by construction.* Referees, editors and judges are separate AI calls that see only the document in front of them, never the author's reasoning; a changed thesis is treated as a new idea and screened again.
- *Measured against published work.* Each accepted paper is compared, in both reading orders, with published papers from its own debate and with hidden papers that were withheld from its author.

# 2 The literature as foundation

Philosophical contributions are made to particular conversations. A pipeline that generated arguments without knowing which conversations exist, what they have settled and what they leave open would produce ideas that are either familiar or unmoored. The first layer of the pipeline therefore reconstructs the literature as a set of debates, and every later stage draws on it.

## 2.1 From citations to debates

To find the conversations, a script grouped articles by the work they build on. Two articles that cite many of the same sources are likely to be addressing the same questions, so shared references are a direct, argument-neutral trace of a debate.

The data are 32,685 works from 24 philosophy journals, of which 27,346 are research articles once reviews, front matter and other non-articles are removed. Two articles are linked when they share at least two references and their reference lists overlap with a cosine similarity of at least 0.05 (shared references divided by the geometric mean of the two lists' lengths); 19,254 articles have such a link. A standard community-detection method (weighted Louvain clustering) then divided the linked articles into groups, with its resolution parameter chosen to maximise the

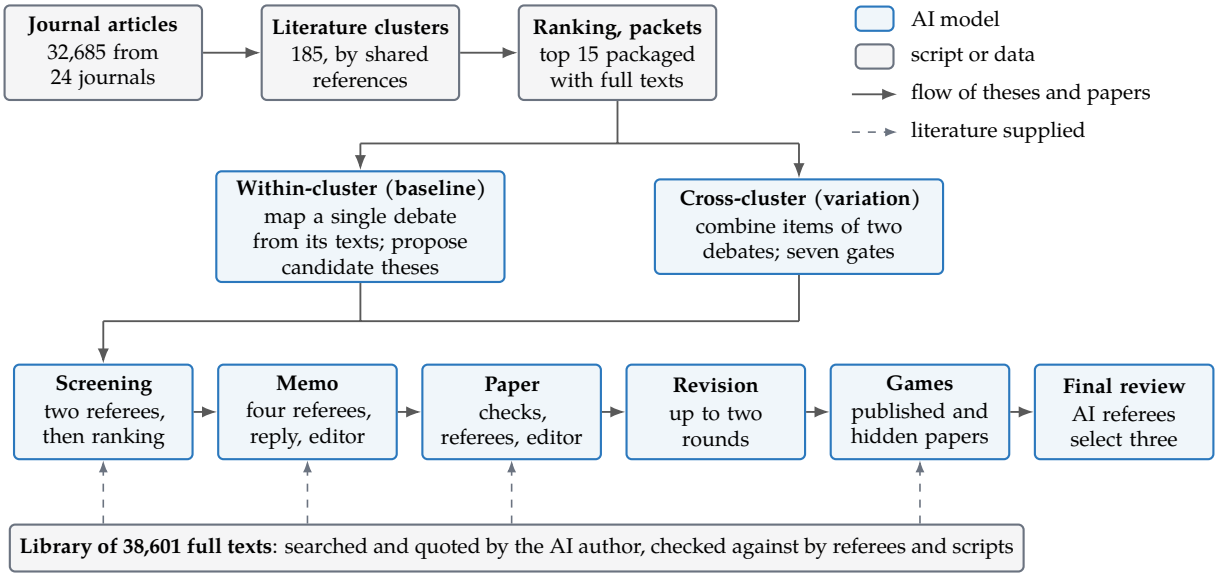



**Figure 1.** The pipeline. Top row: the literature layer, run by scripts. Middle row: the two ways of generating candidate theses; within-cluster generation is the baseline and produced all three submitted papers. Bottom row: the paper layer, which every candidate thesis enters at screening. Blue boxes are steps performed by AI models, grey boxes by scripts or data. Solid arrows carry candidate theses and papers from step to step; dashed arrows mark where the library of full texts is supplied to a step.

number of articles in groups of 20 to 300 papers. Dropping the 242 groups with fewer than 20 papers left 185 literature clusters. In all, 27,346 research articles gave 19,499 core articles with at least five recorded references, 1,281,667 links between them, 427 groups and 185 clusters of at least 20 core articles. Figure 2 illustrates the method. Articles without enough references were attached by the text similarity of their titles and abstracts. The clusters are stable: re-running the method five more times from different random starts reproduces the median cluster with an overlap of 87 per cent. A cluster usually corresponds to a single debate, though it can merge two debates that share a canon.

To decide where to work, a script ranked the 185 clusters on nine equally weighted measures in three groups: activity and relevance (recent output, the share of articles in leading generalist journals, authorship by members of the judging committee, growth), researchability (access to core texts, coverage of references, stability of the cluster, availability of short papers) and connections to other clusters. The 15 best-ranked clusters were packaged for the generators (Table 1). The debates of the submitted papers were then chosen without any human reading of a paper: for the runs that produced Papers A and C, a group of four AI assessors compared 15 candidate debates for room for a contribution, likely reception, feasibility with our library and risk, and proposed, among others, the debates of Papers A and C; for Paper K, the within-cluster generator's eligibility rule selected its cluster (Section 3).

## 2.2 What the literature does in the pipeline

Once the debates are fixed, the literature serves six distinct functions, summarised in Table 2.

*Packets.* Each of the 15 packets holds the members of one cluster with their full texts (669 texts across the packets), the membership list and the citation graph among the members, a core of highly cited papers, recent papers and contrasting readings, and a list of central papers whose texts were missing. The generators work from these packets, not from the open web.

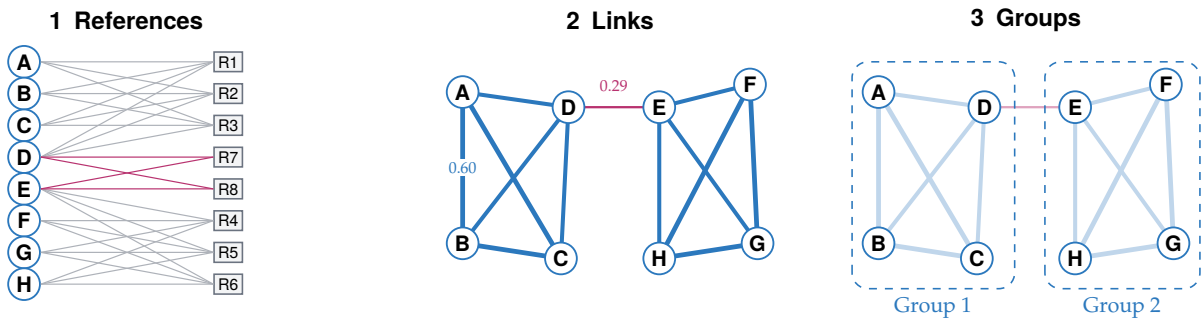



**Figure 2.** How shared references define debates, illustrated with eight made-up papers, A to H; they are not papers from the study. (1) Papers A–D cite works R1–R3, papers E–H cite R4–R6, and D and E also cite R7 and R8 (magenta). Each paper has two further references of its own, not drawn. (2) Two papers are linked when they share at least two references; line width shows their overlap, the number of shared references divided by the square root of the product of the two reference-list lengths (A–B: $3/\sqrt{5 \cdot 5} = 0.60$; D–E: $2/\sqrt{7 \cdot 7} = 0.29$). The study kept links with an overlap of at least 0.05. (3) Densely linked papers form groups; a link across groups does not merge them. The real study found its groups with weighted Louvain community detection; the outlines here only illustrate the idea.

*The library.* The licensed library holds 38,601 full texts of journal articles. The AI author searches it with a full-text search tool (BM25 ranking), reads what it finds and quotes from it. Scripts then check that the quoted passages of every paper occur verbatim in the library texts and that every cited work is held, and a paper that fails goes back to its author. Referees, too, receive the opening excerpts of the works a document cites and the closest search hits for its thesis, so their judgements of novelty and accuracy rest on the texts themselves.

*Maps of debates.* Before proposing any thesis, the generators map the debate: who holds which position, which principles pull against each other, which problems the papers themselves flag as open, and which objections have gone unanswered. Each entry is tied to a passage in a source text. The map is what makes a candidate thesis answerable to a debate rather than to a topic: a thesis must say which item of the map it addresses and who would have to change their view.

*Prior-art checks.* A thesis that has already been defended adds nothing, however well it is argued. Novelty is therefore tested repeatedly against the literature: in the cross-cluster gates by a search of the library in the vocabularies of both debates and of a public index across all fields (Crossref), followed by a web search; in the paper pipeline by a prior-art referee at screening and a prior-art investigator at the memo stage, both supplied with library excerpts.

*Benchmarks.* The papers against which each submitted paper is measured come from its own debate. For Papers C and K, comparison papers and hidden test papers were chosen by a fixed rule from the debate's most-cited recent papers in leading journals; for Paper A, whose run came first, an AI assistant chose them from published papers of the same debate, spread across venues and decades. The hidden papers were removed from the author's search index, and in later runs also blocked from retrieval. For Paper K, whose cluster's texts had all been given to the generator, the rule's exception applied: the hidden papers are the first three recent papers from outside the cluster that cite at least three of its papers.

## 3 Within-cluster generation: the baseline

Most philosophical contributions grow out of a debate's own unfinished business. The within-cluster method models this directly. It stays inside one literature cluster, reconstructs the state

**Table 1.** The 15 packaged clusters in rank order, with their band on each of the nine ranking measures. Bands are relative deciles among the 185 clusters, from 1 (lowest) to 10 (highest); darker cells are higher. The score is the plain mean of the nine bands, and a person set the cutoff at 15. Rows above the rule are the six clusters the standalone within-cluster generator started from (Section 3); ranks 6 and 7 have the same score, and the ranking script's order is kept. The clusters of the submitted papers carry the paper's letter. Measures: recent output; share of articles in leading generalist journals; authorship by members of the judging committee; growth; access to core texts; reference coverage; stability under re-clustering; availability of short papers; connections to other clusters.

| | | Activity and relevance | | | | Researchability | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Rank** | **Cluster** | Recent output | Top journals | Committee | Growth | Core access | Ref. coverage | Stability | Short papers | Connections | **Score** |
| 1 | Cluster 1 (C) | 9 | 9 | 10 | 9 | 10 | 7 | 10 | 9 | 7 | 8.89 |
| 2 | Cluster 2 | 4 | 10 | 10 | 10 | 4 | 10 | 10 | 4 | 10 | 8.00 |
| 3 | Cluster 3 | 7 | 9 | 10 | 8 | 7 | 9 | 8 | 4 | 9 | 7.89 |
| 4 | Cluster 4 | 9 | 8 | 4 | 10 | 10 | 2 | 9 | 10 | 8 | 7.78 |
| 5 | Cluster 5 | 7 | 9 | 8 | 6 | 9 | 9 | 8 | 4 | 6 | 7.33 |
| 6 | Cluster 6 (K) | 9 | 8 | 4 | 9 | 10 | 9 | 3 | 9 | 3 | 7.11 |
| 7 | Cluster 7 (A) | 8 | 10 | 10 | 9 | 5 | 8 | 6 | 4 | 4 | 7.11 |
| 8 | Cluster 8 | 10 | 8 | 4 | 8 | 9 | 10 | 6 | 4 | 3 | 6.89 |
| 9 | Cluster 9 | 10 | 8 | 10 | 6 | 10 | 10 | 2 | 4 | 2 | 6.89 |
| 10 | Cluster 10 | 9 | 8 | 8 | 10 | 2 | 4 | 9 | 4 | 8 | 6.89 |
| 11 | Cluster 11 | 7 | 10 | 4 | 9 | 7 | 8 | 9 | 4 | 4 | 6.89 |
| 12 | Cluster 12 | 5 | 10 | 10 | 4 | 7 | 8 | 7 | 4 | 7 | 6.89 |
| 13 | Cluster 13 | 7 | 10 | 8 | 6 | 7 | 5 | 9 | 4 | 5 | 6.78 |
| 14 | Cluster 14 | 6 | 10 | 4 | 9 | 4 | 3 | 6 | 9 | 10 | 6.78 |
| 15 | Cluster 15 | 4 | 10 | 8 | 6 | 7 | 6 | 7 | 4 | 9 | 6.78 |

of its debate, and proposes theses that address what the debate itself leaves open. It never combines material across clusters. The method was implemented twice, and both implementations produced submitted papers.

### 3.1 Mapping inside the paper pipeline (Papers A and C)

In the runs on a single debate, the paper pipeline writes its own candidate theses. A new AI author session reads the debate's core papers in full and about 20 further works from the library, and writes a map of the debate: positions and who holds them, principles in tension, open problems the papers flag (with located quotations) and unanswered objections. The map contains no theses. The same session then writes 16 candidate theses from the map, and a separate call to the model writes 8 more from the map and the core texts. Each candidate thesis (the pipeline calls it an idea card) states a thesis, its core argument in steps, the map item it addresses, who must change what, the closest existing work and the precise difference, three consequences for distinct questions, and the strongest objection. Paper C grew from a thesis written by the mapping session, Paper A from one written by the second call.

### 3.2 The standalone within-cluster generator (Paper K)

A second implementation was built as a separate generator, with stricter requirements on how each thesis is tied to the literature. It starts from the six best-ranked clusters and keeps those with at least five available papers of at most 6,000 words, a rule fixed in advance; three clusters qualified. For each cluster, the generator receives the packet, reads the five most-cited short papers

**Table 2.** Six functions of the literature in the pipeline. Each row names what the pipeline takes from the published literature and the stage that uses it.

| Function | What the literature supplies | Used by |
|---|---|---|
| Finding debates | Citation links among 32,685 articles, grouped into 185 clusters and ranked | Choice of debates |
| Packets | For each of 15 clusters: its members' full texts, the citation graph among them, and its most-cited, recent and contrasting papers | Both generators |
| The library | 38,601 full texts, searched by the AI author, quoted in the papers, and used to check that every quotation occurs word for word | Author, referees, scripted checks |
| Maps of debates | Positions, principles in tension, open problems and unanswered objections, each tied to a located quotation | Candidate theses, screening |
| Prior-art checks | Searches of the library and of a public bibliographic index; excerpts of the works a document cites and of the closest search hits | Cross-cluster gates, referees |
| Benchmarks | Published comparison papers and hidden test papers from the same debate | Judging |

in full to calibrate what a short contribution can do, consults a reading set of at least ten papers, and maps the debate in four fixed sections: positions, principles in tension, open problems and unanswered objections, each item linked to source passages. It then proposes ten distinct candidate theses (prompt excerpt in Appendix A). Each must engage at least ten papers of its own cluster, each through a quotation of at least 40 characters that a script locates in the stored text, and must contain a scope plan for a 6,000-word paper: comparisons with the short papers, the argument's boundaries, its dependencies, its main risk of overreach and a word budget for each part. A script checked the packet hashes, the structure, the references to the map, the declared full readings, every quotation and every word budget. All three clusters returned ten valid theses. The ten for the cluster that produced Paper K entered the paper pipeline at screening; one of them became Paper K. Figure 3 shows how they are tied to the debate: together they engage 27 of its papers, each thesis ten, and many papers serve several theses.

## 4 Cross-cluster generation: a variation

New philosophical ideas sometimes come from applying the tools of a debate to the problems of another. The cross-cluster method tests whether this can be done systematically. It combines items from two debates that rarely cite each other, for instance a method from the first and an open problem from the second, and then tries hard to kill each combination before any of it is written up. It is a variation on the baseline: its candidate theses enter the same paper pipeline. Two versions ran; the account below follows the second, which drew on the 15 packets. Its runs produced candidates that the AI editor accepted, none of which was among the three selected.

*An item bank with verified quotations.* One AI call per packet read excerpts of five full texts and every member's abstract and wrote down what the debate's literature contains: positions, tensions, open problems, unanswered objections, shared assumptions, methods and cases. Each item has a short statement, a field-neutral version with schematic letters in place of the field's terms, structural tags, and one or two verbatim quotations. A script searched for every quotation in its source and dropped the item if it was not found: 4 of 367 items were dropped, leaving 363.

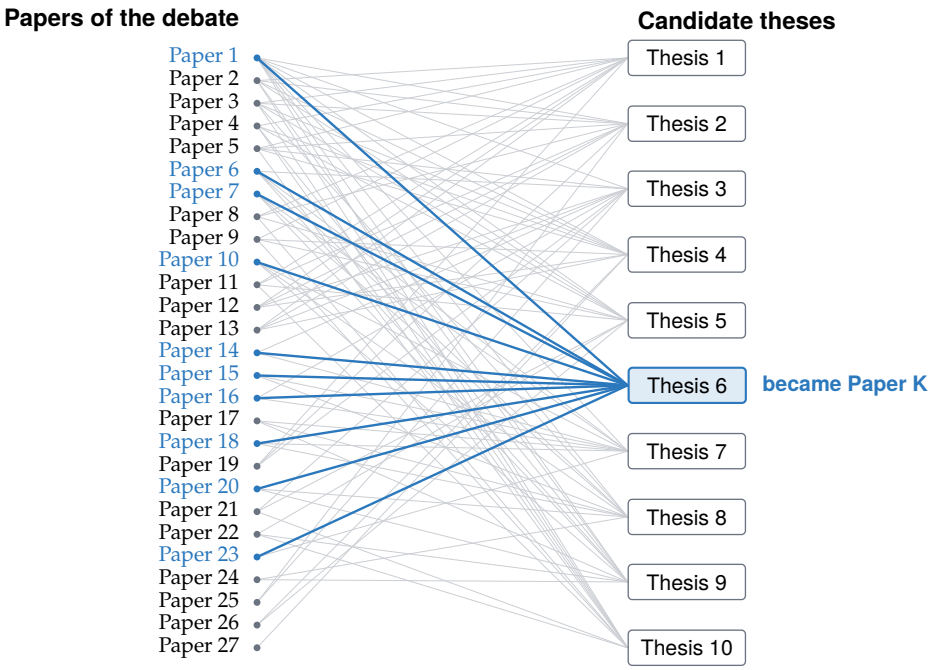



**Figure 3.** The ten candidate theses for the cluster that produced Paper K and the papers of the debate they engage (papers are numbered; their names are withheld). Each line is a closest-work comparison backed by a quotation that a script found word for word in the source text; each thesis engages ten papers of the debate. Blue: the ten papers engaged by Thesis 6, which became Paper K. Papers are ordered by the number of theses that engage them.

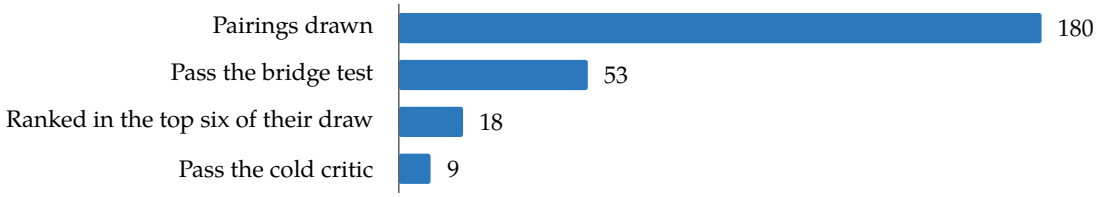



**Figure 4.** Combinations of the three main cross-cluster draws that survived the gates. The bridge test, the ranking and the cold critic did all the filtering; the co-citation, prior-art and web-search gates removed none of these combinations.

*Random draws.* Combinations were drawn at random, in a draw fixed in advance, from pairs of debates weighted by how closely their literatures are coupled: the most closely coupled pairs (11 of 105) were never drawn, and pairs in the middle range of coupling were favoured. Each combination joins one item from each debate under a recipe: apply the method of the first debate to an open problem of the second; turn an unanswered objection of the first against a position in the second; or set two tensions of the same shape side by side. Both items must share a structural tag. Three draws of 60 gave 180 combinations.

*Seven gates.* Every combination then had to pass seven gates in order (Table 3). The design principle is that early gates kill only what is hopeless, while later gates, applied to fewer and more developed candidates, ask harder questions. The bridge test is the route's central idea: written with the same schematic letters, both items must remain true of their own fields, which separates a shared structure from a shared word (Appendix A). Figure 4 shows how many combinations survived.

*Frozen rules and the hand-off.* Every script, prompt and configuration file of the generator, together with the models and the selection rules, was written down before the first draw, and no

**Table 3.** The seven gates of cross-cluster generation, in order. A combination that fails a gate goes no further. The last two columns give the number of combinations entering and leaving each gate in the three main draws together and in the fourth draw, which stayed inside the pair of debates chosen by the best survivor.

| Gate | Question asked | Answered by | Main draws | Pair draw |
|---|---|---|---|---|
| 1 Co-citation | Do three or more works already cite both source papers? If so, the literature already links them. | Script on the citation graph | $180 \to 180$ | $28 \to 28$ |
| 2 Bridge test | Stated with the same schematic letters, do both items still hold of their fields? Only hopeless pairings fail, never merely risky ones. | AI model | $180 \to 53$ | $28 \to 7$ |
| 3 Prior art | Is the combination already published? A partly anticipated one passes with a note of what remains new. | Script search of the library and a public index; AI referee | $53 \to 53$ | $7 \to 6$ |
| 4 Ranking | Which survivors are most promising? The top six of each draw go on; the rest are set aside, not refuted. | AI ranker, three shuffled orders | $53 \to 18$ | $6 \to 6$ |
| 5 Thesis writing | Each survivor is written up as a candidate thesis: claim, argument in steps, closest work and the difference. | AI writer | $18 \to 18$ | $6 \to 6$ |
| 6 Web search | Is the thesis already in print anywhere? | AI model with web search | $18 \to 18$ | $6 \to 6$ |
| 7 Cold critic | Is the link between the two debates substantive, or a shared word or loose analogy? | AI critic that sees only the thesis | $18 \to 9$ | $6 \to 2$ |

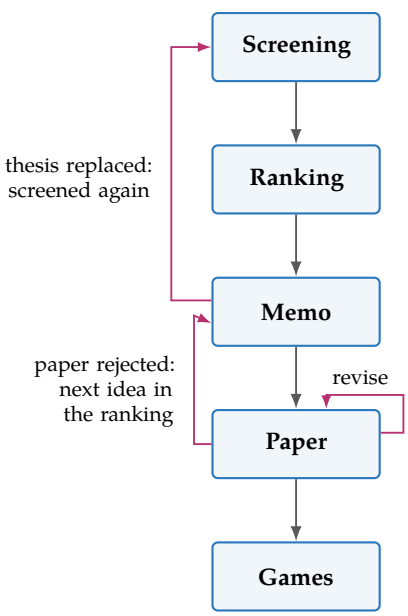
human read a candidate thesis from the real draws before it was handed on. By a rule fixed in advance, the survivor with the best normalised rank chose the pair of debates, here a pair of the 15 clusters. A fourth draw of 28 combinations stayed inside that pair. The paper pipeline received four candidate theses from it: the best survivor, one further thesis from the same pair, and the two survivors of the fourth draw, together with a map of the two debates (48 items) built by a script from the item bank.

## 5 From candidate thesis to paper

The generators propose; the paper pipeline decides. Its purpose is to subject each candidate thesis to the kind of scrutiny a paper receives from referees, as early and as cheaply as possible, and to develop only the theses that survive. Every candidate thesis enters at the same point, whatever its origin. Figure 5 shows the stages.

*Screening and ranking.* A prior-art referee classifies each candidate thesis as anticipated, partly anticipated or new, and a sceptical reader judges whether it is promising, doubtful or to be rejected. Both read the debate map and the literature excerpts. Survivors are then ranked side by side. Comparative ranking is needed because absolute scores do not discriminate: in earlier runs each referee gave a promise score of 2 out of 5 in 148 of 174 reports. From Paper C's run onwards, an AI editor orders all survivors at once, three times in shuffled order, and the mean rank decides which go on; Paper A's run, the earliest of the three, predates this ranker.

*The memo.* Each selected thesis gets its own author session, which writes a memo of two to three pages before any paper is drafted. Four referees report independently: a field specialist, a prior-




A prior-art referee and a sceptical reader read each candidate thesis with the debate map and literature excerpts. A thesis is dropped if anticipated, if an issue is fatal, or if the reader rejects it.

An AI editor orders the survivors side by side, three times in shuffled order. The best go on; the others wait in reserve.

The author writes a two-to-three-page memo. Four referees report independently, the author answers every issue, and an editor who sees only the memo, the reports and the reply decides whether the idea survives.

The author writes the paper. Scripts check length, citations and quotations; an AI check reviews attributions and anonymity; four fresh referees report; an editor blind to the paper's history accepts, asks for a revision or rejects.

The accepted version is compared with published papers from its debate and with hidden test papers (Section 6).

**Figure 5.** The paper pipeline. Each stage is carried out by separate AI calls; referees, editors and checks see only the document before them. Magenta arrows show the three ways back: a thesis that changes during the memo stage is screened again as a new idea, a revision request starts a new round of referees (at most two), and a rejected paper is replaced by the next memo in the ranking.

art investigator who also searches neighbouring fields, an argument critic, and a sceptical reader who asks whether the consequences are shown or merely asserted. Every issue is classified as fatal, answerable from the paper's own resources, or requiring engagement with a named work. The author replies to each issue. A separate check then asks whether the thesis is the same, narrowed or replaced; a replaced thesis is a new idea and goes back to screening. An editor who sees only the memo, the reports and the reply decides whether the idea survives. Later runs added a panel modelled on the competition's judges that contributes further objections; none of the submitted papers' runs used it.

*The paper.* The memo's author session writes the paper. Scripts check that it compiles, that every citation resolves, that it stays within 5,700 words before the bibliography, that every cited work is held, that quoted passages are verbatim, and that it contains no talk about its own production. An AI check reviews quotations, attributions and anonymity. In Paper K's run, an outside reader with nothing but the paper reconstructed its argument and listed the places where it lost the thread. Four fresh referees then report, and an editor who never sees the paper's history accepts it, asks for a revision, or rejects it. The editor may not accept while a fatal issue stands, while an answerable issue is open, or while a work bearing on the paper's novelty or validity is unengaged; a script enforces these rules on the editor's answer (Appendix A). Up to two rounds of revision are allowed. During revision, games against three of the six comparison papers (Section 6) decide between versions: a revision is kept only if it is accepted or wins more games without losing significance.

*Changes between runs.* The pipeline was improved while the runs that produced the candidate papers were under way. In order, it gained the comparative selection ranker, a stricter scoring guide for the judge (Section 6), a limit on the works an editor may demand, a guard against revisions that lose argument, and the word cap; then the memo panel; then stricter safeguards for the hidden test papers and checks on the scope of claims and definitions; and last, an author brief that puts the reader first and the outside reader. A run used the pipeline as it stood when the run began. Papers A and C therefore passed through fewer of these steps than Paper K.

**Table 4.** What each level of the judge's scoring guide means.

| Level | Level description |
|---|---|
| 5 | exemplary on this criterion: a careful specialist referee would ask for no change on it |
| 4 | strong on this criterion: only minor, local fixes are needed |
| 3 | competent on this criterion, with at least one substantial but repairable problem |
| 2 | serious problems on this criterion: fixing them would change the thesis or the main argument |
| 1 | fails the criterion |

*Origin of the theses.* The thesis of every submitted paper first appears in a candidate thesis written by an AI model: by the paper pipeline's mapping session (C) or its second call (A), or by the standalone within-cluster generator (K). The paper authors received only the pipeline's templated task prompts and the referees' reports. No human supplied a thesis, argument, objection, reply, sentence or source.

## 6 Judging against published work

Referees and editors decide whether a paper is acceptable; they do not tell us how it compares with what is already published. The judging stage answers that question directly, by setting each accepted paper against published papers from its own debate.

*Games.* A game is a paired comparison. An AI judge reads two papers without author names, decides which is better on each of the competition's six criteria (clarity, quality of argumentation, significance, originality, engagement with the literature, accuracy and scholarship) and overall, and then reads them again in the opposite order; a win scores one point and a tie half a point. Each accepted paper plays six comparison papers from its debate, twelve games in all, and then two or three hidden test papers that had been withheld from its author, in four or six games. During revision, games against three of the six comparison papers decided whether a revision was kept, and the author was shown the judge's account of what those papers do better; only the hidden papers played no part in development, so they are the cleaner test. The game judge for all three submitted papers was gpt-6.1-sol, a model from a different developer than the one that wrote and refereed the papers.

*Scoring.* The judge scores each of the competition's six criteria on five levels, each defined by the work a careful specialist referee would still ask for (Table 4; the full scoring guide is in Appendix B). All three submitted papers were scored the same way.

*Results.* Figure 6 shows the games of the three submitted papers. Paper A won 11 of twelve games against the comparison papers and all four against its two hidden papers; its run had released a third hidden paper to the author when referees asked for it, so that paper was left out of the test, which was played after the run. Paper C won 10 of twelve against the comparison papers and 5.5 of six against its three hidden papers; Paper K won 10 of twelve and all six. No submitted paper lost a game against the papers its author never saw.

*How demanding the judge is.* The judge's levels are relative to its own standards, not the verdicts of journals. In earlier runs, the judge scored nine published papers 382 times (Figure 7). Papers from Noûs and The Philosophical Review never reached level 5 on argument or scholarship. On average, essays from earlier runs of the pipeline scored above the published papers on clarity, argument and scholarship, and below the Noûs and Philosophical Review papers on significance,

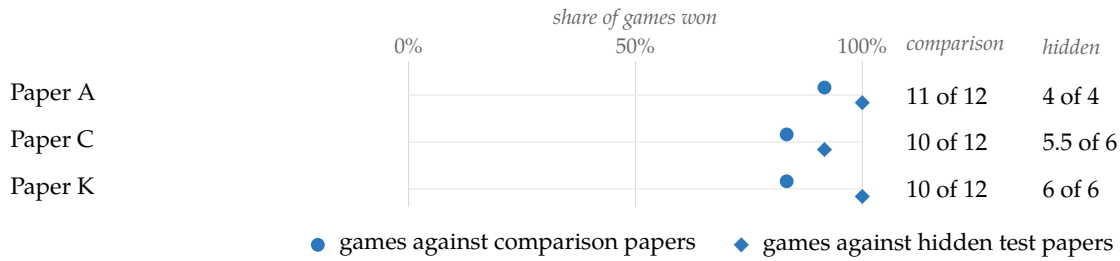



**Figure 6.** Share of games won by the three submitted papers against published papers from their debates: six comparison papers, each read in both orders (dots), and hidden test papers withheld from the author (diamonds; two papers for A, three for C and K). The columns on the right give the counts. All three papers were scored the same way.

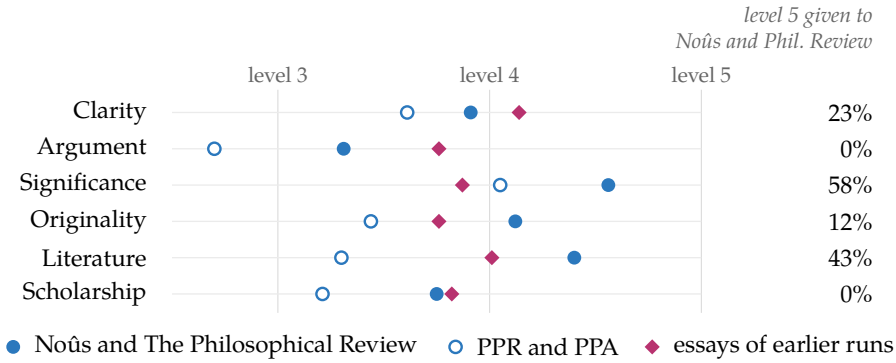



**Figure 7.** Mean level per criterion that the game judge gave in earlier runs to published papers from Noûs and The Philosophical Review (5 papers, 243 scores), to published papers from Philosophy and Phenomenological Research and Philosophy and Public Affairs (4 papers, 139 scores), and to essays from earlier runs of the pipeline. The right-hand column gives the share of level-5 scores among the Noûs and Philosophical Review papers.

originality and engagement with the literature. In a further test, six essays were edited to plant defects in seven ways and judged under both rubrics in both orders: edits that damaged clarity, introduced talk about the essay's production, or weakened engagement with the literature or scholarship were detected in 92 to 100 per cent of cases, while a hedged thesis and a borrowed step were detected under neither rubric. In the pipeline, those two defects are the business of the referees and of the editor's acceptance standard (Appendix A).

## 7 The three submitted papers

The pipeline ran on several debates, and across these runs its AI editor accepted several candidate papers. Over the same runs, AI models proposed 162 candidate theses, of which 17 were developed into full papers. AI referees then reviewed the accepted candidates with the instruction shown in Appendix A: "you are a top philosopher refereeing this paper for a nous submission. give an honest review and assessment of its philosophical contribution, novelty and execution." This review was run with six different models (last row of Table 7), and the authors took the three papers that were reviewed best across all six; they are identified in Table 5. Once the three papers were chosen, each received a final full referee report from an AI assistant, written as if for a top journal. The AI then revised each paper in response.

**Table 5.** The three submitted papers. Each comes from a different debate; all three come from within-cluster generation. Titles, debates and summaries are withheld until the competition's results are announced.

| | Route |
|---|---|
| **A** | Within-cluster: mapping inside the paper pipeline (run with high reasoning effort) |
| **C** | Within-cluster: mapping inside the paper pipeline (run with very high reasoning effort) |
| **K** | Within-cluster: standalone generator |

## 8 Who did what

Humans designed the process; AI models did the philosophical work. The team assembled the licensed library and the article database, wrote the clustering and ranking scripts, chose the ranking measures and the cut-offs, built the three components (the cross-cluster generator, the within-cluster generator and the paper pipeline), fixed and froze their rules, and configured, started and restarted the runs; in one run they resumed work on papers the run had stopped. They also wrote the instruction for the final review.

AI models proposed every candidate thesis, wrote every sentence of every paper, refereed, edited and judged the papers, and reviewed the candidates; their reviews determined the selection, and the team confirmed it, choosing one paper per debate. No human proposed an idea, wrote or edited a sentence of any paper, judged a draft, or supplied a work that a referee asked for.

## 9 Resources

Table 6 summarises the scale of the work and its cost. The cost covers every AI call recorded across all three team members' work and all trials, including runs and experiments that produced no submitted paper. Calls made through subscriptions are valued at the providers' list prices, so the figures are estimates rather than invoices.

**Table 6.** Scale and cost of the work. Counts of candidate theses and full papers refer to the runs that produced the candidate papers; calls and cost cover all recorded work.

| | |
|---|---:|
| Journal articles analysed | 32,685 |
| Literature clusters | 185 |
| Full texts in the library | 38,601 |
| Candidate theses proposed | 162 |
| Full papers written | 17 |
| Papers submitted | 3 |
| AI calls | about 4,560 |
| Estimated cost at list prices | about US$2,860 |
| Estimated cost per submitted paper | about US$950 |

## 10 Models, harnesses and reproducibility

Table 7 lists the models used in each role and the software through which they ran. Within the paper pipeline, referees, editors, checks and judges are fresh calls that see only the document

and the supplied literature excerpts; the author works in a continuing session per idea, with file tools and the library search tool, and without web search.

**Table 7.** Models and harnesses by role. "Very high" denotes the reasoning-effort setting xhigh.

| **Role** | **Model (reasoning effort)** | **Harness** |
|---|---|---|
| Choice of debates for A and C | Claude, as four AI assessors | Claude subagents |
| Within-cluster mapping and candidate theses in the paper pipeline (A, C) | Claude Opus 5.5 (high for A, very high for C) | The paper pipeline's agent runtime |
| Standalone within-cluster generator (K) | gpt-6-astra (high) | Codex command line, read-only sandbox |
| Cross-cluster generator | Claude Opus 5.5 (high); bridge test Claude Sonnet 5.5 (medium) | Claude Code command line (version 2.1.314) |
| Screening, memo and paper referees, editor, checks and author | Claude Opus 5.5 (high for A, very high for C and K) | The paper pipeline: agent sessions for the author, fresh calls for the others |
| Game judge (published and hidden papers) | gpt-6.1-sol | The paper pipeline |
| Final review of the candidates | Claude Opus 5.5, Claude Fable 5.1, gpt-6-luna, gpt-6-astra, gpt-6.1-sol, DeepSeek Flash 4.1 | |

*Version control and records.* The paper pipeline's code, prompts and rubrics are kept in a Git repository, and each run stores its full configuration with its records. The cross-cluster generator's code, prompts and configuration were recorded before its first draw, and the within-cluster generator's packets were checked before its theses were handed to the paper pipeline. Every model call is logged with its cost, and each run keeps every version of each paper and every report, so the development of each submitted paper can be traced step by step. The licensed library and the logs that quote it are available to the organisers on request.

# A Selected prompts

The pipeline runs on many prompts. The four below show its most distinctive decisions: how the cross-cluster route tests a combination, how the within-cluster generator ties a thesis to the literature, how the editor is held to a standard, and how the submitted papers were selected.

### Cross-cluster generator · the bridge test

*The field-neutral bridge test is the most distinctive idea of the cross-cluster route: it asks whether two items from different debates share a structure, not merely a word. Instructions to the AI model, excerpt.*

For each combination decide only whether it is hopeless. Do not kill a combination because its thesis would probably turn out false; high-risk combinations should pass.
Kill (keep = false) only for one of these reasons:
– equivocation: the items share a word but not a structure. Test: state both items with the SAME schematic letters; if either statement is then false of its field, it is equivocation;
– preconditions_unmet (O1): the target lacks a structure the method requires [...];
– no_deliverable: nothing statable would come out (no theorem, countermodel, verdict, or changed argument); [...]

END OF PROMPT TEXT

### Within-cluster generator · what each candidate thesis must show

*These requirements tie every candidate thesis to the literature of its debate and to the length of a competition paper. From the generator's specification, excerpt.*

Each card had to engage at least ten distinct, text-available papers from its own cluster, with source evidence and a stated relationship to the proposed claim. [...] Inside Core argument, a required subsection, "Scope: a 6,000-word paper," specified relevant comparisons to the selected short papers, argument boundaries, essential dependencies, a principal scope risk and a first check. [...] For each counted paper, an evidence quotation had to occur exactly in the stored text.

END OF PROMPT TEXT

### Paper pipeline · the editor's acceptance standard

*This rule stops the AI editor from accepting a paper while a fatal or answerable issue stands; a script enforces it on the editor's answer. Excerpt.*

The acceptance standard: a paper is accepted only if it meets the standard as it stands. A fatal issue that stands after the author's response, or a skeptical reader's verdict of reject that the document as it now stands does not answer (no consequence beyond the debate shown, or a thesis hedged until safe), means revise or fatal_reject, never accept. [...] Set answerable_resolved to true only if the document as it stands resolves every issue not binned needs_work [...].

END OF PROMPT TEXT

### Final review · the instruction to the AI referees

*The complete instruction under which AI referees reviewed the candidate papers before the three were selected.*

you are a top philosopher refereeing this paper for a nous submission. give an honest review and assessment of its philosophical contribution, novelty and execution.

END OF PROMPT TEXT

# B The judge's scoring guide

The scoring guide, used by the game judge for all three submitted papers, scores each of the competition's six criteria on five levels. Each criterion states its question, what it does not assess, and its levels; the rubric also defines caps that limit a level when a stated condition holds. The criteria and their questions come from the competition's published guidance; the levels, exclusions and caps are the project's own.

**Clarity.** *Is the thesis identifiable? Are important terms defined? Is the structure of the reasoning clear?*

Does not assess: whether the steps are valid or the premises defended (argument); how the contribution is positioned against other work (literature).

**5** The introduction states the question, the thesis and a roadmap; the thesis can be restated after one reading; every technical or coined term is defined at first use; each section's role in the argument is evident; prose is plain.
**4** Thesis, terms and structure are clear; one or two terms, transitions or passages need tightening.
**3** The thesis is identifiable but has to be assembled from several places, or some central term, or which step is meant to follow from which, must be guessed; labels, digressions or dense prose obscure parts of the structure.
**2** The thesis is ambiguous between materially different claims, or the argument turns on terms that are never defined; the main argument cannot be reconstructed without substantial guesswork.
**1** No thesis can be identified, or much of the text is unintelligible.

**Quality of argumentation.** *Do the conclusions follow from the premises? Are the premises adequately defended? Does the work answer serious objections?*

Does not assess: whether the reader can find the thesis, terms and steps (clarity); how important or new the conclusion is (significance, originality).

**5** The conclusion follows from explicitly stated premises; each crucial premise has support independent of the conclusion; the strongest objections, including those against the author's own positive view, are stated in their strongest form and answered with further argument; no gap remains that a referee would require fixing.
**4** The main inference is valid and the crucial premises are defended; one secondary premise, step or objection needs more work.
**3** The main inference works, but a crucial premise is asserted rather than defended, or a serious objection is missing or answered weakly.
**2** A crucial premise is undefended or doubtful, the conclusion needs an unstated assumption, or replies to objections beg the question.
**1** The conclusion does not follow, or the text asserts rather than argues.

**Significance.** *Would the thesis or argument make a worthwhile contribution to philosophical knowledge? Is the conclusion non-trivial?*

Does not assess: whether the conclusion is new (originality) or whether the main argument establishes it (argument): judge how much the conclusion would matter if it were established.

**5** If the conclusion is accepted, it changes how several important questions beyond the immediate debate should be approached, and the text shows how; it is of general interest to academic philosophers.
**4** A substantial contribution to an active debate, with consequences beyond the work it engages that are stated and argued for.
**3** A real but local contribution: it settles or sharpens a point within one narrow exchange, and wider consequences are asserted rather than shown.
**2** The contribution is minor, the conclusion is close to obvious, or the thesis has been weakened until it is safe.
**1** The conclusion is trivial or the question unimportant.

**Originality.** *Are the work's contributions original?*

Does not assess: how well the text positions itself against the closest work (literature) and how much the contribution matters (significance): judge what is in fact new, not how its novelty is announced.

**5** The thesis or argument would surprise a specialist: it is not anticipated by the closest existing work.
**4** A clearly new argument, distinction or application that is not in the closest existing work, though related work anticipates part of it.
**3** A genuine new element, but much of the contribution is anticipated by existing work; or a new argument for a familiar conclusion.
**2** Mostly restatement or recombination of known positions; little is new.
**1** The main contribution is known prior work.

**Engagement with the literature.** *Does the work understand and engage with the work most relevant to its contribution?*

Does not assess: whether quotations, citations, attributions, readings of texts and the reference list are accurate (scholarship); whether the contribution is in fact new (originality).

**5** The work most relevant to the contribution, including the strongest opposing views, is engaged: the text says what each cited work holds and how its own argument bears on it, and locates its contribution precisely against the closest work.
**4** The main relevant work is engaged; one relevant strand is missing or the engagement is thin in places.
**3** The main sources are present, but the contribution's relation to them is vague, or several citations serve only as pointers.
**2** Most of the relevant work is missing, or cited works are listed rather than engaged.
**1** No meaningful engagement with the work most relevant to the contribution.

**Accuracy and scholarship.** *Are sources represented fairly? Are quotations and citations accurate? Does the work acknowledge relevant prior work?*

Does not assess: which relevant work is covered, how deeply it is engaged and how the contribution is positioned against it (literature); talk about the essay's own production (clarity). Here acknowledging prior work means crediting ideas and results the item takes from others.

**5** Every attribution and quotation that can be checked is accurate and located (page or section where the argument relies on a passage); readings of texts are supported by the text; every in-text citation has a complete reference-list entry; ideas and results taken from others are credited.
**4** Attributions are accurate; minor slips in citations or reference-list entries.
**3** Attributions are broadly right, but some are loose, unlocated or taken second-hand from another work's report, or an idea or result taken from others is not credited.
**2** A source is misrepresented on a point that matters, or several cited works are missing from the reference list.
**1** A citation or quotation is fabricated, or a source the argument depends on is misrepresented.

---

This report was generated with AI under substantial human direction and editing. The authors checked it carefully, but cannot guarantee that it is free of errors.