Topic Modeling of the AI Ethics, Generative AI, and Large Language Model Literature with Latent Dirichlet Allocation: Methodology and Preliminary Results

Jean Vides
Systems Engineering Program, Faculty of Engineering
Universidad del Magdalena
jcvides@unimagdalena.edu.co
November 2026

***Abstract*—**The rapid diffusion of generative artificial intelligence (AI) and large language models (LLMs) has produced a fast-growing, interdisciplinary body of scholarship on their ethical, social, regulatory, and safety implications. This paper presents the methodology and preliminary results of a Latent Dirichlet Allocation (LDA) topic model applied to a Scopus corpus of 20,000 records published between 2018 and 2026, of which 19,505 documents were retained after screening. Titles and abstracts were cleaned with a deterministic pipeline combining lemmatization, bigram detection, and four layers of stopword filtering. The number of topics was selected by comparing C_v and NPMI coherence, topic diversity, dominance, and perplexity across $K = 4$–40. A 27-topic model with an asymmetric document–topic prior and $\eta = 0.5$ ($C_v = 0.576$; NPMI = 0.082; top-10 diversity = 0.774) was fitted with Gensim. Robustness was examined through random-seed, neighbouring-K, and hyperparameter sensitivity analyses, and the topic space was explored with HJ-Biplots and LDAvis. The topics separate into a technical layer (evaluation, fine-tuning, retrieval, multimodality, agents) and a normative layer (governance, bias, privacy, security, alignment, regulation). Publication volume is overwhelmingly concentrated in 2025–2026 and in the United States and China. Seed stability is moderate (mean Jaccard = 0.319; mean ARI = 0.241), which supports exploratory use of the model while low-volume topics should be treated as provisional.

***Keywords*—**topic modeling; latent Dirichlet allocation; generative AI; large language models; AI ethics; topic coherence; HJ-Biplot; LDAvis

1. INTRODUCTION

Scholarship on the ethics of artificial intelligence has evolved quickly. Early efforts concentrated on mapping and consolidating principles, such as the global landscape of AI ethics guidelines (Jobin et al., 2019) and the ethical framework proposed by Floridi et al. (2018). Critical analyses of algorithmic systems (Mittelstadt et al., 2016) were later followed by work on the specific risks of large language models (Bender et al., 2021; Weidinger et al., 2021). The public release of widely accessible generative systems has accelerated this output: in the corpus analysed here, 82.8% of the documents were published in 2025 and 2026.

At this scale, manual reading is no longer a viable way to characterize the field. Topic models treat text as data and offer a reproducible way to summarize large collections (Grimmer & Stewart, 2013), and Latent Dirichlet Allocation (LDA; Blei et al., 2003) remains the reference probabilistic model for this purpose. Its application nevertheless involves consequential decisions: how texts are cleaned, how many topics are fitted, whether the solution is stable, and whether the resulting topics are interpretable to humans (Chang et al., 2009; Greene et al., 2014). These decisions are often under-reported.

This paper documents an end-to-end LDA pipeline applied to Scopus records on generative AI, large language models, and their governance

and ethical dimensions. It is guided by three questions: (RQ1) what thematic structure characterizes the literature; (RQ2) how that structure is distributed over time and across countries; and (RQ3) how robust the solution is to random initialization, the number of topics, and the Dirichlet priors. The contributions are:

- a transparent, PRISMA-style account of corpus construction, screening, and de-duplication;
- a fully specified preprocessing and model-selection procedure based on multiple coherence and structure metrics;
- a three-part sensitivity analysis (seeds, neighbouring K, and priors); and
- a 27-topic thematic map interpreted jointly through HJ-Biplots and LDAvis.

Section 2 gives background; Section 3 describes the data and methods; Section 4 reports results; Section 5 discusses findings, methodological considerations, and limitations; and Section 6 concludes.

2. BACKGROUND

2.1 Latent Dirichlet allocation

LDA is a hierarchical generative model for collections of discrete documents in which each document is a mixture of a fixed number of latent topics and each topic is a probability distribution over the vocabulary (Blei et al., 2003). Four properties make it appropriate for this corpus. First, titles and abstracts are short, heterogeneous texts for which a probabilistic mixture is more stable than hard clustering. Second, the aim is to describe the thematic structure of the field and its evolution rather than to classify documents into predefined categories. Third, the document–topic matrix (Θ) supports downstream analysis with HJ-Biplots, LDAvis, and trend models. Fourth, AI-ethics literature is interdisciplinary: a single paper may combine governance, bias, privacy, and technical evaluation, which favors a mixed-membership representation.

2.2 Evaluating topic models

Because LDA is unsupervised, the number of topics and the quality of a solution must be assessed indirectly. Held-out likelihood alone can be at odds with human interpretability (Chang et al., 2009), which motivated automatic coherence measures validated against human ratings (Lau et al., 2014; Röder et al., 2015). Among them, C_v combines semantic similarity, a sliding window, and mutual confirmation (Röder et al., 2015), while normalized pointwise mutual information (NPMI; Bouma, 2009) penalizes spurious associations and is bounded in $[-1, 1]$. Because fitted models also vary with initialization and with K, stability across seeds and neighbouring values of K is an additional criterion (De Waal & Barnard, 2008; Koltcov et al., 2014; Greene et al., 2014).

2.3 Visual analytics for topic spaces

The HJ-Biplot (Galindo-Villardón, 1986), an extension of Gabriel's (1971) biplot, represents the rows and columns of a matrix simultaneously in a low-dimensional space, so that years, countries, or documents can be read against topics. LDAvis (Sievert & Shirley, 2014) complements it by projecting topics according to the distance between their word distributions and ranking terms by relevance. The first offers a structural reading of *where* entities lie in the thematic space; the second a lexical reading of *why* a topic is interpreted as it is.

3. DATA AND METHODS

3.1 Corpus collection

Records were collected from **Scopus only**; no Web of Science data were integrated. The query targets the intersection of generative AI and language models with governance, ethics, regulation, and safety:

("generative AI" OR "large language model" OR "language model" OR "foundation model*" OR GPT OR chatbot OR "text generation") AND (governance OR ethics OR regulation OR "responsible AI" OR "algorithmic accountability" OR "AI policy") AND PUBYEAR AFT 2017 AND PUBYEAR BEF 2028 AND (LIMIT-TO (LANGUAGE , "English")) AND PUBYEAR > 2017 AND PUBYEAR < 2027*

The query was run on 11 October 2026. The number of records was capped at 20,000 because of the computational budget and processing time available for model fitting, the K search, and the

sensitivity analyses; the modeling corpus is therefore a capped subset of the full Scopus result set. The export contains 20,000 records with title, abstract, keywords, year, source, affiliations, affiliation country, document type, and DOI. The observed window is 2018–2026; 2026 is incomplete because the export was made during that year.

3.2 Screening and de-duplication

Screening followed the PRISMA 2020 guideline (Page et al., 2021) as summarized in Table 1 and Fig. 1. All 20,000 records had both title and abstract. Records with fewer than 20 words of useful text (495) were excluded. Duplicates were detected by DOI (20) and by title (59) but were **retained** and are reported as a limitation. The final modeling corpus comprises 19,505 documents, spanning 86 normalized countries and 5,697 distinct sources.

Table 1. Corpus screening summary.

Metric	Value
Records identified (Scopus)	20,000
With title and abstract	20,000
Without useful text	0
DOI duplicates (retained)	20
Title duplicates (retained)	59
Excluded by length (< 20 words)	495
Usable documents	**19,505**
Documents > 350 words	739
Countries (normalized)	86
Distinct sources	5,697

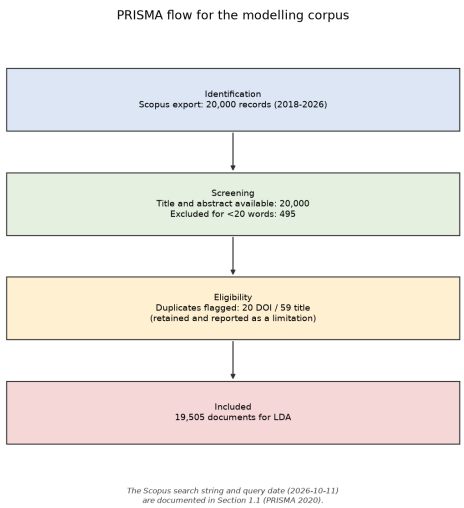


Fig. 1. PRISMA flow diagram of the Scopus corpus modeled.

3.3 Text preprocessing

The **title** and **abstract** were concatenated. Author and index keywords are available but were not appended, to avoid duplicating terms and to keep a homogeneous representation across documents. The following steps were applied deterministically:

1. Unicode normalization and lowercasing.
2. Removal of editorial noise: URLs, e-mail addresses, numeric citations, copyright and license notices, and “published by” sentences.
3. Whitespace tokenization of the cleaned text.
4. WordNet lemmatization (nouns first, then verbs); no stemming, to preserve interpretability.
5. Four-layer stopword removal: English functional stopwords (NLTK); editorial terms (e.g., publisher names, “journal”, “abstract”); methodological terms (e.g., “study”, “results”, “method”); and generic AI-ethics domain terms, while preserving valid acronyms.
6. Length filtering: tokens shorter than three characters were discarded, except valid acronyms such as *ai*, *llm*, *gpt*, *nlp*, and *gdpr*.
7. Bigram detection with *gensim.models.phrases* (minimum count 59, threshold 8), yielding terms such as *generative_ai* and *machine_learning*.

8. Vocabulary filtering: terms appearing in fewer than 39 documents or in more than 50% of documents were removed from the training dictionary.

Preprocessing used Python 3.13, NLTK, Gensim 4.4, pandas, and NumPy. Table 2 summarizes its output.

Table 2. Preprocessing results.

Metric	Value
Modeled documents	19,505
Total tokens	2,361,836
Tokens per document (mean)	121.09
Training vocabulary	4,533
Evaluation vocabulary	23,887
Documents with empty BoW	0
Total stopwords	288

3.4 Model specification

For a document d, LDA assumes the following generative process:

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad (1)$$

$$z_{d,n} \sim \text{Categorical}(\theta_d), \quad (2)$$

$$w_{d,n} \sim \text{Categorical}(\varphi_{z(d,n)}), \quad (3)$$

$$\varphi_k \sim \text{Dirichlet}(\eta), \quad (4)$$

where θ_d is the topic-proportion vector of document d, $z_{d,n}$ is the latent topic of its n-th word, and φ_k is the word distribution of topic k. The joint likelihood integrates out the latent variables and is approximated by variational inference or Gibbs sampling; Gensim's *LdaModel* uses online variational inference. Table 3 lists the final configuration.

Table 3. Final LDA configuration.

Parameter	Value
Number of topics (K)	27
α (document–topic prior)	asymmetric
η (topic–word prior)	0.5
Passes	50
Iterations per document	50
Random seed	42
Documents / vocabulary	19,505 / 4,533
Implementation	Gensim 4.4.0 *LdaModel*

3.5 Selecting the number of topics

Two coherence metrics were computed: C_v and NPMI (Section 2.2). Four structural diagnostics were also recorded: **topic diversity** (proportion of unique words among the top-10 words of all topics), **perplexity**, **dominance** (share of documents assigned to the most frequent topic), and the **normalized entropy** of the topic-size distribution. K = 4–40 (37 values) was evaluated with a symmetric α and 30 passes. The selected K was required to (i) lie on the coherence plateau, (ii) remain parsimonious and interpretable, (iii) show high diversity, (iv) keep dominance below an excessive-dominance threshold of 40%, and (v) be stable relative to its neighbours.

3.6 Hyperparameter search

For the selected K, a grid of 16 combinations (four values of α by four values of η) was evaluated, and the configuration with the highest C_v was retained. The final model was then trained with 50 passes.

3.7 Sensitivity analysis

Three sources of variation were examined. (a) **Random seed:** five seeds (42–46) were trained for each K in the neighbourhood 25–29 with 10 passes. (b) **Neighbouring *K*:** each neighbourhood model was compared with the reference model (K = 27, seed 42). (c) **Priors:** the 16-combination α/η grid. Runs were compared using the **Jaccard** overlap between the top-10 words of aligned topics, where alignment was solved with the Hungarian algorithm over the cosine similarity of topic–word distributions, and the **Adjusted Rand Index (ARI)** between dominant-topic assignments.

3.8 HJ-Biplot and LDAvis

For a centered matrix $\mathbf{X}$, the HJ-Biplot is obtained from the singular value decomposition

$$\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathrm{T}}, \quad (5)$$

with row coordinates $\mathbf{A} = \mathbf{UD}$ and column coordinates $\mathbf{B} = \mathbf{VD}$, so that the scalar product $\mathbf{a}_i \cdot \mathbf{b}_j$ approximates x_{ij} and rows and columns share one interpretive space. The representation quality of row i on the first two axes is the squared cosine

$$\cos^2{}_i = (a_{i1}^2 + a_{i2}^2) / \Sigma_k\, a_{ik}^2, \quad (6)$$

and analogously for columns; global quality is the mean $\cos^2$. Three biplots were built: **years** × **topics** (years with at least 20 documents; cell

value is the mean topic probability), **countries × topics** (the 12 most productive countries), and **documents × topics** (a random sample of 1,500 documents colored by dominant topic).

In LDAvis, topics are projected onto an intertopic distance map with bubble size proportional to prevalence, and terms are ranked by relevance,

$$\text{rel}(w \mid t) = \lambda \log p(w \mid t) + (1 - \lambda) \log [p(w \mid t)/p(w)], \quad (7)$$

where $\lambda \in [0, 1]$ trades conditional frequency against exclusivity. Relevance values are reported on the logarithmic scale; $\lambda = 0.6$ was used for exploration (Sievert & Shirley, 2014), and rankings for $\lambda = 0$ and $\lambda = 1$ were also exported.

3.9 Implementation and reproducibility

All code is organized as seven scripted phases (preprocessing, audit, K diagnostic, K decision, final model, sensitivity, and reporting) and run with a global seed of 42 in an environment reproduced with *uv sync*. Results are exported as CSV files, intermediate checkpoints as pickles, and the full configuration, library versions, metrics, and artifact inventory as a single JSON file.

4. RESULTS

4.1 Corpus characterization

The modeled corpus is very recent (Table 4). Only four documents date from 2018–2021, and 2022 contributes 31. Output then rises steeply, from 556 documents in 2023 to 2,763 in 2024, 6,883 in 2025, and 9,268 in 2026, so that 82.8% of the corpus was published in the last two years. The 2026 figure is a partial-year count.

Table 4. Documents per year. *2026 is incomplete.

Year	Documents	%
2018	2	0.01
2019	1	0.01
2021	1	0.01
2022	31	0.16
2023	556	2.85
2024	2,763	14.17
2025	6,883	35.29
2026*	9,268	47.52

The United States (23.71%) and China (20.32%) together account for 44.0% of the corpus, followed by India, Germany, and the United Kingdom (Table 5). Articles (42.65%) and conference papers (41.07%) dominate the document types (Table 6), with reviews at 9.26%. The single most frequent source is the proceedings of the Annual Meeting of the Association for Computational Linguistics (542 documents; 2.78%), followed by *Lecture Notes in Computer Science*, CEUR Workshop Proceedings, and the ACM CHI proceedings (Table 7). The most productive institutions are the University of California system (354 documents), Stanford University (326), Tsinghua University (322), Carnegie Mellon University (255), and Zhejiang University (249).

Table 5. Top ten countries.

Country	Documents	%
United States	4,538	23.71
China	3,888	20.32
India	1,022	5.34
Germany	999	5.22
United Kingdom	891	4.66
Italy	617	3.22
Canada	500	2.61
Australia	487	2.54
South Korea	449	2.35
Spain	316	1.65

Table 6. Main document types.

Document type	Documents	%
Article	8,319	42.65
Conference paper	8,011	41.07
Review	1,806	9.26
Book chapter	838	4.30
Note	164	0.84
Book	127	0.65

Table 7. Main sources.

Source	Docs	%
Proc. Annual Meeting of the Assoc. for Computational Linguistics	542	2.78
Lecture Notes in Computer Science	365	1.87
CEUR Workshop Proceedings	281	1.44
Conf. on Human Factors in Computing Systems	278	1.43
IEEE Access	260	1.33
Advances in Neural Information Processing Systems	232	1.19

Source	Docs	%
Communications in Computer and Information Science	218	1.12
Proceedings of Machine Learning Research	179	0.92

4.2 Model selection and final model

Coherence rises quickly up to about K = 11–14 and then flattens (Fig. 2, Table 8; the full grid is in Appendix A). The global maximum occurs at K = 40 (C_v = 0.5696), but its advantage over K = 27 (C_v = 0.5640; NPMI = 0.0749) is only 0.0056 and lies within the plateau. K = 27 also shows high top-10 diversity (0.7889), a dominance of 10.6% (well below the 40% threshold), no empty topics, and stable neighbours (Section 4.4). Perplexity varied little across the grid (−7.41 to −7.37). K = 27 was therefore selected.

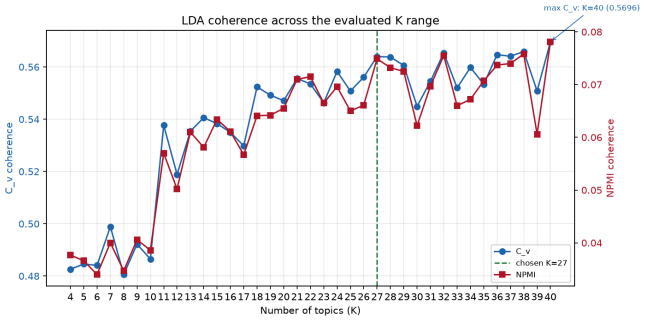


Fig. 2. Coherence by number of topics.

Table 8. Selected rows of the K search (symmetric α, 30 passes).

K	C_v	NPMI	Diversity	Dominance
4	0.4826	0.0377	0.8750	0.2740
8	0.4806	0.0347	0.7125	0.1863
12	0.5189	0.0502	0.7750	0.1518
16	0.5350	0.0611	0.7688	0.1148
20	0.5471	0.0655	0.7450	0.1346
24	0.5583	0.0696	0.7792	0.1160
27	**0.5640**	**0.0749**	**0.7889**	**0.1056**
30	0.5449	0.0622	0.7867	0.0854
34	0.5598	0.0672	0.7647	0.0783
38	0.5659	0.0758	0.7421	0.0748
40	0.5696	0.0781	0.7825	0.0928

The α/η grid (Table 9) yielded the best coherence for α = asymmetric and η = 0.5 (C_v = 0.5861), and the C_v range across all 16 combinations was only 0.0288. The final model (Table 10) attains C_v = 0.5761 and NPMI = +0.0818 with no empty topics. Its C_v differs from the grid value for K = 27 (0.5640) because the grid used a symmetric α and 30 passes; comparisons are valid only within a configuration.

Table 9. Eight best α/η combinations.

α	η	C_v	NPMI	Dom.
asymmetric	0.5	0.5861	0.0861	0.0890
0.01	0.5	0.5779	0.0812	0.0771
symmetric	0.5	0.5748	0.0808	0.0789
0.1	0.5	0.5713	0.0793	0.0815
asymmetric	0.01	0.5657	0.0776	0.1273
asymmetric	0.05	0.5656	0.0765	0.1268
0.01	0.05	0.5649	0.0748	0.1023
symmetric	symmetric	0.5640	0.0749	0.1056

Table 10. Final model metrics.

Metric	Value	Metric	Value
C_v	0.5761	Dominance	0.0844
NPMI	+0.0818	Norm. entropy	0.9475
Diversity (top-10)	0.7741	Empty topics	0
Perplexity	−7.2693		

Table 11. The 27 LDA topics: proposed label, number of documents with the topic as dominant, top ten terms by frequency, and the country contributing most documents to the topic. Labels were assigned by the authors.

ID	Proposed label	Docs	Top terms	Leading country
1	Human problem-solving and AI-assisted work	1,168	work, make, system, ai, problem, one, human, task, find, question	United States
2	AI governance and framework design	1,064	ai, framework, governance, system, design, across, structure, rather, evidence, support	United States
3	Generative AI ethics, technology, and societal challenges	1,647	ai, generative_ai, technology, artificial_intelligence, ethical, genai, challenge, development, tool, generative_artificial	United States
4	Privacy, explainability, and financial trust	181	privacy, explanation, financial, audit, framework, transparency, explainability, explainable, trust, sensitive	United States
5	ChatGPT response quality and performance evaluation	996	response, chatgpt, question, accuracy, score, performance, compare, test, quality, gpt-4	United States

ID	Proposed label	Docs	Top terms	Leading country
6	Perception, trust, and AI adoption surveys	437	participant, ai, survey, perceive, perception, finding, group, influence, among, impact	United States
7	Agentic systems, architecture, and enterprise workflows	686	agent, system, architecture, framework, agentic, multi-agent, agentic_ai, ai, rag, workflow	United States
8	Bias, fairness, gender, and stereotype in prompts	644	bias, fairness, gender, prompt, social, stereotype, across, demographic, output, group	United States
9	Clinical, medical, and healthcare AI	881	clinical, medical, patient, healthcare, health, care, support, medicine, include, clinician	United States
10	Prediction, foundation models, and biological sequence learning	503	prediction, foundation_model, machine_learn, ml, biological, feature, learn, predict, sequence, representation	China
11	Safety, cybersecurity, and adversarial robustness	512	safety, risk, security, attack, threat, vulnerability, cybersecurity, adversarial, detection, defense	United States
12	Multimodal and vision-language models	182	multimodal, image, visual, video, modality, vision, mllms, generation, foundation_model, vlms	China
13	Evaluation, benchmarks, and performance measurement	784	evaluation, benchmark, across, score, test, metric, prompt, assessment, performance, framework	United States
14	Training, fine-tuning, and machine unlearning	1,098	train, fine-tune, performance, task, unlearn, inference, optimization, introduce, framework, learn	China
15	Intelligent management, markets, and energy systems	700	management, system, framework, market, energy, intelligent, integrate, power, process, technology	China
16	Public policy, digital services, and urban platforms	111	policy, service, public, information, urban, platform, web, digital, government, access	United States
17	User interaction, chatbots, and conversational support	768	user, design, interaction, system, support, chatbots, chatbot, mental_health, conversational, dialogue	United States
18	Human cognition, emotion, and social simulation	551	human, social, cognitive, emotional, emotion, agent, theory, narrative, behavior, interaction	United States
19	Text detection, datasets, and classification	1,081	text, detection, dataset, classification, performance, prompt, fine-tune, generate, task, datasets	United States
20	Retrieval, knowledge graphs, and semantic generation	1,277	framework, reason, knowledge, semantic, retrieval, rag, structure, system, generation, query	China
21	Networks, robotics, simulation, and digital twins	223	plan, system, network, environment, physical, robot, simulation, control, scenario, traffic	China
22	Alignment, values, and moral decision-making	244	human, alignment, value, moral, preference, ethical, align, decision, decision-making, reason	United States
23	Application challenges, reviews, and field advances	1,498	application, challenge, review, development, include, field, advance, capability, potential, survey	China
24	Language, culture, translation, and multilingual corpora	126	language, cultural, translation, english, linguistic, multilingual, write, chinese, text, context	China
25	ChatGPT in academic and systematic literature reviews	499	chatgpt, review, include, publish, tool, identify, literature, scientific, systematic_review, qualitative	United States
26	Law, regulation, compliance, and software requirements	732	legal, process, code, requirement, automate, software, regulatory, regulation, compliance, case	United States
27	AI in education and student learning	912	student, learn, education, educational, feedback, tool, skill, assessment, chatgpt, teach	United States

4.3 Topic structure

Table 11 shows that the topics cover technical, applied, and normative themes. The largest topics by dominant-document count are T3 (generative AI ethics, 1,647 documents; 8.4% of the corpus), T23 (application challenges and reviews, 1,498), T20 (retrieval and knowledge graphs, 1,277), T1 (human problem-solving, 1,168), and T14 (training and unlearning, 1,098); the smallest are T16 (public policy, 111), T24 (language and culture, 126), T4 (privacy and explainability, 181), T12 (multimodal models, 182), and T21 (networks and robotics, 223). The United States leads 19 of the 27 topics by document volume and China leads the remaining eight (T10, T12, T14, T15, T20, T21, T23, and T24).

For interpretation, the topics were grouped into five thematic blocks: **risk, security, and policy** (T4, T11, T16, T22, T26); **governance and ethics** (T2, T3, T6, T8, T24); **applications and evaluation** (T9, T13, T14, T15, T23, T27); **interaction and agents** (T7, T17, T18, T20, T21); and a **technical core** (T1, T5, T10, T12, T19, T25). Summing dominant-topic counts per block (Table 12) shows that applications and evaluation is the largest block (30.1%), followed by the technical core (22.7%) and governance and ethics (20.1%).

Table 12. Thematic blocks (derived by summing dominant-topic counts).

Block	Topics	Docs	%
Risk, security, and policy	T4, T11, T16, T22, T26	1,780	9.1
Governance and ethics	T2, T3, T6, T8, T24	3,918	20.1
Applications and evaluation	T9, T13, T14, T15, T23, T27	5,873	30.1
Interaction and agents	T7, T17, T18, T20, T21	3,505	18.0
Technical core	T1, T5, T10, T12, T19, T25	4,429	22.7

4.4 Sensitivity and robustness

Seeds. For the selected $K = 27$, the mean Jaccard overlap between aligned topics across seed pairs is 0.3192 (SD = 0.0242; minimum 0.2756) and the mean ARI is 0.2413 (SD = 0.0280; minimum 0.1985) (Table 13). Neighbouring values of K behave similarly, so $K = 27$ is neither notably more nor less stable than its neighbours.

Table 13. Seed stability by K, mean (SD) over seed pairs.

K	Jaccard	ARI	C_v
25	0.346 (0.034)	0.246 (0.027)	0.546 (0.010)
26	0.310 (0.039)	0.246 (0.033)	0.546 (0.012)
27	**0.319 (0.024)**	**0.241 (0.028)**	**0.551 (0.010)**
28	0.294 (0.036)	0.223 (0.028)	0.543 (0.007)
29	0.305 (0.026)	0.259 (0.033)	0.550 (0.010)

Neighbouring *K*. Compared with the reference model, neighbouring models recover only part of the topic set at a Jaccard threshold of 0.5: 11 of 27 reference topics for $K = 25$, 8 for $K = 26$, 10 for $K = 28$, and 11 for $K = 29$ (Table 14). Topics therefore split and merge as K changes, although several topics persist across models (a best-match Jaccard of 1.0 is reached for $K = 25$, 26, and 28).

Table 14. Best-match Jaccard against the reference model ($K = 27$, seed 42).

K	Mean	Median	Ref. topics covered (≥ 0.5)
25	0.4412	0.4286	11 / 27
26	0.4048	0.3810	8 / 27
27	1.0000	1.0000	27 / 27
28	0.3844	0.4286	10 / 27
29	0.3715	0.3333	11 / 27

Priors. The coherence differences across the 16 α/η combinations are small (range 0.0288), and the selected configuration is among the best, suggesting that the structure is reasonably robust to the prior as long as α is asymmetric or close to symmetric. **Overall**, seed stability is moderate. The highest-volume topics (T3, T23, T20, and T1) are recovered most consistently, whereas the smallest topic (T16) is the most sensitive.

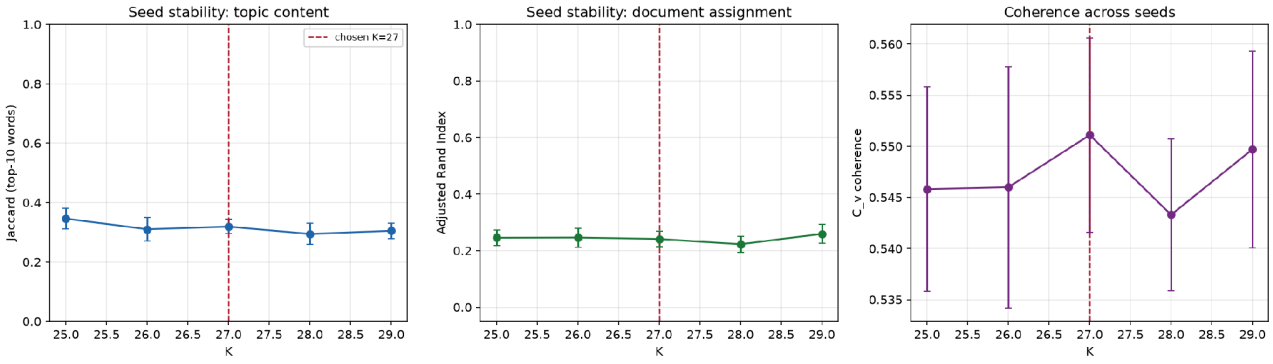


Fig. 3. Stability of the LDA solution across random seeds for $K = 25$–29.

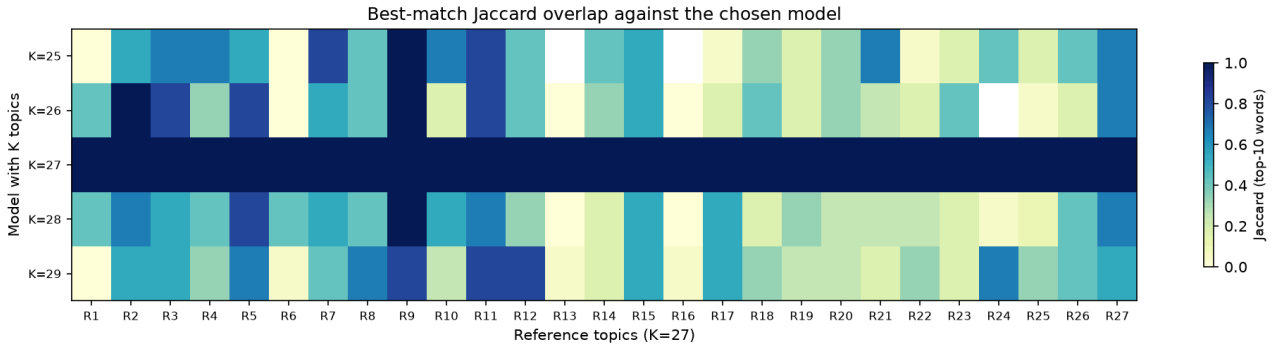


Fig. 4. Overlap between neighbouring-K models and the reference model ($K = 27$).

4.5 HJ-Biplot results

Table 15 reports the quality of the three biplots. The years × topics biplot is well represented (the first two axes retain 95.6% of the variance; mean $\cos^2 \approx 0.90$ for rows and 0.88 for columns), the countries × topics biplot is moderately represented (82.8% of the variance; mean $\cos^2$ of 0.56 and 0.46), and the documents × topics biplot is poorly represented (about 18.6%; mean $\cos^2$ of 0.18 and 0.11) and should be read qualitatively.

Table 15. HJ-Biplot quality.

Dataset	Rows × cols	Axes 1 / 2 (%)	$\cos^2$ rows	$\cos^2$ cols
Years × topics	5 × 27	77.5 / 18.1	0.900	0.883
Countries × topics	12 × 27	59.7 / 23.1	0.562	0.461
Documents (sample) × topics	1,500 × 27	10.3 / 8.2	0.176	0.108

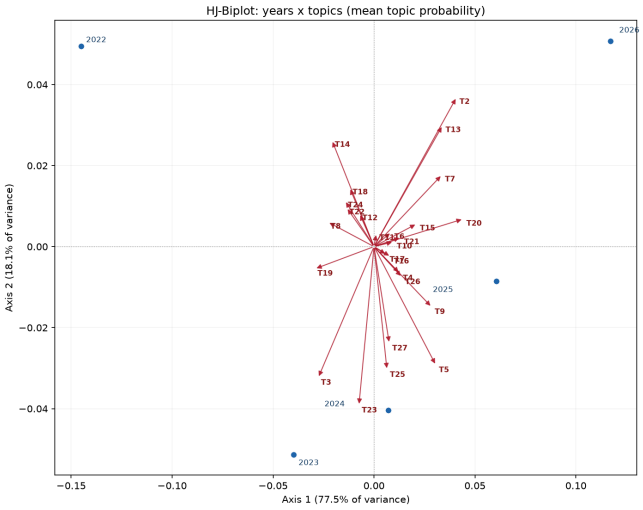


Fig. 5. HJ-Biplot of years × topics.

In the years biplot, the proximity of a year to a topic indicates that the topic gains relative weight in that year: 2022–2023 align with foundational and early-application topics, whereas 2024–2026 shift toward generative AI, agents, retrieval-augmented generation, and model evaluation.

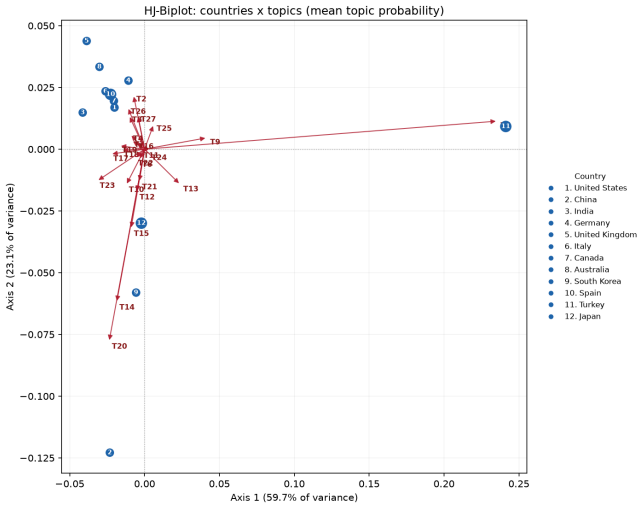


Fig. 6. HJ-Biplot of countries × topics.

In the countries biplot, the United States and China occupy central positions by volume, and other countries spread across the thematic space according to their mean topic profile. Positions of countries with low $\cos^2$ should be treated as unreliable.

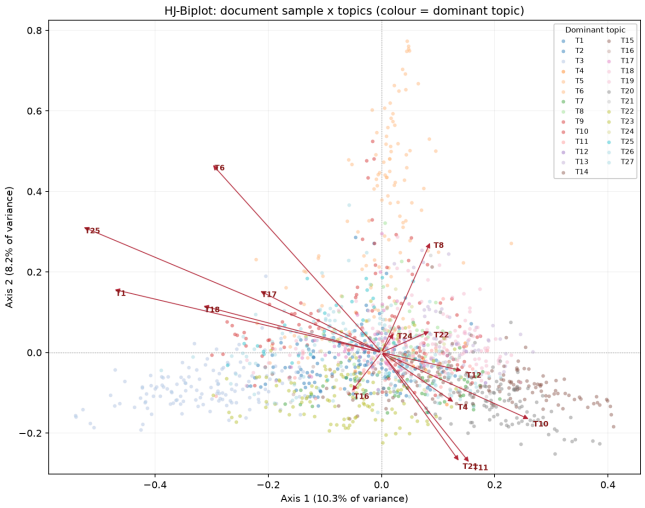


Fig. 7. HJ-Biplot of a 1,500-document sample, colored by dominant topic.

The document cloud suggests a separation between technical-evaluation documents (T5, T13, T19) and ethics and governance documents (T3, T8, T11), with an overlap zone around interaction and agents (T7, T17). Given the low representation quality, this reading is indicative only.

4.6 LDAvis results

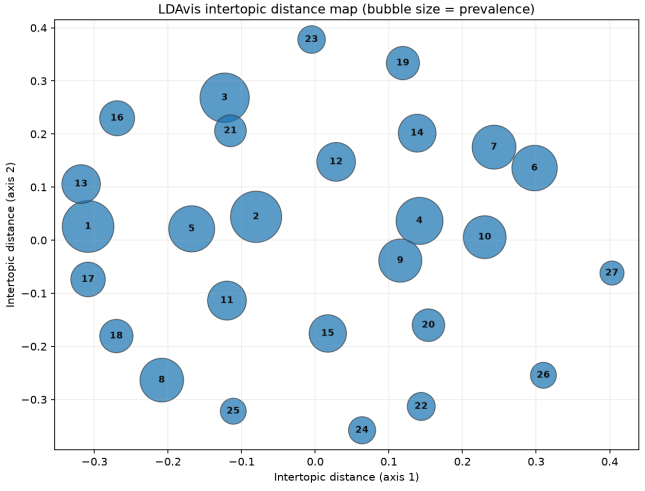


Fig. 8. LDAvis intertopic distance map.

On the intertopic map, the generative-AI (T3), evaluation (T13), and training (T14) topics lie close together because they share technical vocabulary, whereas ethics (T8) and governance (T2) topics separate from the technical core. Small topics (T16, T24) appear isolated, consistent with their low volume, and T3 has the largest bubble.

With $\lambda = 0.6$, the most relevant terms combine frequency and exclusivity. T11 is characterized by *safety*, *security*, *threat*, and *attack*; T9 by *clinical*, *medical*, and *patient*; T20 by *semantic*, *reason*, *knowledge*, and *retrieval*; and T8 by *bias* and *fairness*. T24 is the only topic whose leading term (*language*) has a positive relevance score. By contrast, the leading terms of T1 (*make*, *work*, *problem*) and T23 (*application*, *challenge*, *review*) are generic, indicating low specificity.

5. DISCUSSION

5.1 A technical layer and a normative layer

The 27-topic solution suggests that the literature is a composite of two layers rather than a homogeneous "AI ethics" field. A technical layer—benchmarking (T13), training and unlearning (T14), retrieval and knowledge graphs (T20), multimodality (T12), and agentic architectures (T7)—shares a large, overlapping vocabulary, as the LDAvis map shows. A normative layer—governance (T2), bias and fairness (T8), safety and security (T11), alignment and moral decision-making (T22), and law and compliance (T26)—occupies more clearly separated regions of the topic space. The convergence of the biplot and LDAvis readings strengthens this interpretation.

The block shares in Table 12 qualify the field's self-image: even though the corpus was retrieved with governance- and ethics-related terms, applications and evaluation (30.1%) and the technical core (22.7%) together outweigh governance and ethics (20.1%) and risk, security, and policy (9.1%). This may reflect the breadth of the model-related terms in the search string, but this explanation was not tested. Domain applications form sizeable, distinct topics, notably education (T27; 4.7% of documents) and health (T9; 4.5%), as do ChatGPT-centred evaluation (T5; 5.1%) and literature-review (T25; 2.6%) topics.

5.2 Temporal and geographic patterns

The corpus is dominated by 2025–2026, so any temporal reading is in practice a comparison of the most recent years; the 2018–2022 period contributes only 35 documents, and the years

biplot rests on five years. The apparent shift toward agents, retrieval, and evaluation topics is consistent with the rapid evolution of the technology but should be confirmed with a dedicated trend model. Geographically, the United States and China contribute 44.0% of the corpus. China leads eight topics, including several model-oriented ones (T10, T12, T14, T20, T21), but leadership by volume largely mirrors overall output and is not evidence of specialization; normalizing by each country's total output would be needed to assess that.

5.3 Methodological considerations

The coherence curve plateaus, so the choice of K = 27 rests on parsimony and interpretability as much as on coherence: the best-scoring K (40) offers only a marginal gain. Topic identity is only moderately stable. Seed-pair Jaccard around 0.32 and ARI around 0.24 mean that individual topics, especially small ones, may split or merge across runs, so topic-level claims should be restricted to the more stable, high-volume topics. The sensitivity runs used 10 passes against 50 in the final model, and the stability figures may therefore understate the stability of the final solution; this was not tested. Finally, coherence is not a substitute for human judgment (Chang et al., 2009), and the topic labels in Table 11 were assigned by the authors and have not been validated by independent coders.

5.4 Limitations

- **Single source and language.** Only Scopus records in English were used, which omits other indexes and languages.
- **Incomplete final year.** 2026 is a partial year (query date: 11 October 2026).
- **Capped corpus.** The corpus was limited to 20,000 records for reasons of computational budget and time, so the findings describe this subset and not necessarily the full literature matching the query; they may change if the full result set is modeled.
- **Duplicates.** Twenty DOI duplicates and 59 title duplicates were retained.
- **Text representation.** Only titles and abstracts were modeled; keywords were excluded.
- **Hard assignment.** Document counts per topic use the dominant topic only and ignore mixed membership.
- **Visualization quality.** The documents × topics biplot has low representation quality.

6. CONCLUSIONS AND FUTURE WORK

This paper presented a documented LDA pipeline applied to 19,505 Scopus documents on generative AI, large language models, and their ethical and governance dimensions. A 27-topic model selected on the coherence plateau (C_v = 0.576; NPMI = 0.082) captures a duality between a technical layer and a normative layer of the literature. Publication volume is strongly concentrated in 2025–2026 and in the United States and China. Seed stability is moderate, so the model is suitable for exploratory analysis, with low-volume topics treated as provisional.

Future work includes (i) human validation of topic labels; (ii) a dedicated temporal trend analysis once 2026 is complete; (iii) a comparison with embedding-based topic modeling, as implemented in the companion BERTopic pipeline of the repository; (iv) country-normalized thematic profiles; and (v) scaling the analysis to the full Scopus result set as computational resources allow, together with additional bibliographic sources, full de-duplication, and the inclusion of keyword fields.

DATA AND CODE AVAILABILITY

Code, configuration, and outputs are available at https://github.com/JeanVydes/bertlda. The full model configuration is stored in *outputs/results/config/model_configuration_full.json*, and an interactive LDAvis viewer is provided in *outputs/visualizations/LDAvis.html*.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference* (pp. 31–40).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22* (pp. 288–296).

De Waal, A., & Barnard, E. (2008). Evaluating topic models with stability. In *Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa* (pp. 79–84).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*(3), 453–467. https://doi.org/10.1093/biomet/58.3.453

Galindo-Villardón, M. P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Questiió, 10*(1), 13–23.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)* (LNCS 8724, pp. 498–513). Springer. https://doi.org/10.1007/978-3-662-44848-9_32

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent Dirichlet allocation: Stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 161–165). https://doi.org/10.1145/2615569.2615680

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539). https://doi.org/10.3115/v1/E14-1056

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). https://doi.org/10.1145/2684822.2685324

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). https://doi.org/10.3115/v1/W14-3110

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from language models* (arXiv:2112.04359). arXiv. https://arxiv.org/abs/2112.04359

APPENDIX A. FULL K SEARCH

Table A1. Complete K search (symmetric α, 30 passes). No K produced empty topics. The selected $K = 27$ is in bold.

K	C_v	NPMI	Diversity	Perplexity	Dominance	Entropy
4	0.4826	0.0377	0.8750	-7.4130	0.2740	0.9953
5	0.4846	0.0366	0.8200	-7.3861	0.2516	0.9891
6	0.4841	0.0340	0.7500	-7.3658	0.2298	0.9882
7	0.4989	0.0400	0.7429	-7.3513	0.2060	0.9864
8	0.4806	0.0347	0.7125	-7.3453	0.1863	0.9894
9	0.4922	0.0406	0.7444	-7.3327	0.1491	0.9917
10	0.4865	0.0386	0.7000	-7.3298	0.1390	0.9884
11	0.5377	0.0570	0.8091	-7.3059	0.1288	0.9860
12	0.5189	0.0502	0.7750	-7.3105	0.1518	0.9818

K	C_v	NPMI	Diversity	Perplexity	Dominance	Entropy
13	0.5354	0.0610	0.7769	-7.3168	0.1238	0.9689
14	0.5406	0.0581	0.8000	-7.3209	0.1774	0.9542
15	0.5383	0.0634	0.7467	-7.3229	0.1963	0.9564
16	0.5350	0.0611	0.7688	-7.3216	0.1148	0.9638
17	0.5298	0.0567	0.7588	-7.3233	0.1171	0.9573
18	0.5524	0.0641	0.7778	-7.3192	0.1360	0.9524
19	0.5492	0.0642	0.8000	-7.3205	0.1566	0.9499
20	0.5471	0.0655	0.7450	-7.3280	0.1346	0.9500
21	0.5556	0.0710	0.7952	-7.3287	0.1251	0.9556
22	0.5535	0.0715	0.7818	-7.3270	0.1247	0.9444
23	0.5463	0.0665	0.7826	-7.3371	0.1059	0.9505
24	0.5583	0.0696	0.7792	-7.3304	0.1160	0.9568
25	0.5508	0.0650	0.7800	-7.3335	0.1264	0.9047
26	0.5561	0.0661	0.7654	-7.3405	0.1183	0.9474
27	**0.5640**	**0.0749**	**0.7889**	**-7.3363**	**0.1056**	**0.9398**
28	0.5638	0.0732	0.7679	-7.3404	0.0852	0.9367
29	0.5606	0.0725	0.7655	-7.3312	0.0864	0.9199
30	0.5449	0.0622	0.7867	-7.3426	0.0854	0.9192
31	0.5545	0.0697	0.8065	-7.3472	0.0843	0.9157
32	0.5654	0.0755	0.7781	-7.3458	0.0757	0.9279
33	0.5520	0.0660	0.7545	-7.3501	0.0744	0.9278
34	0.5598	0.0672	0.7647	-7.3557	0.0783	0.9224
35	0.5534	0.0707	0.7800	-7.3563	0.1022	0.9203
36	0.5647	0.0737	0.7833	-7.3615	0.1113	0.9030
37	0.5641	0.0740	0.7892	-7.3628	0.0750	0.9268
38	0.5659	0.0758	0.7421	-7.3675	0.0748	0.9174
39	0.5508	0.0606	0.7615	-7.3717	0.0898	0.8916
40	0.5696	0.0781	0.7825	-7.3704	0.0928	0.9118