Description of Results with Gemma3 4b model in short:

✅ Accuracy of analyzing validness correctly for bad example: 55.56% (5 out of 9 correct)

✅ Accuracy of analyzing validness correctly for valid example: 78.05% (32 out of 41 correct)

✅ Accuracy of analyzing novelty correctly for bad example: 55.56% (5 out of 9 correct)
✅ Accuracy of analyzing novelty correctly for valid example: 95.24% (40 out of 42 correct)

✅ Accuracy of analyzing clarity correctly for bad example: 44.44% (4 out of 9 correct)
✅ Accuracy of analyzing clarity correctly for valid example: 88.10% (37 out of 42 correct)

✅ Accuracy of analyzing feasibility correctly for bad example: 55.56% (5 out of 9 correct)

✅ Accuracy of analyzing feasibility correctly for valid example: 97.62% (41 out of 42 correct)

The model is performing **best with valid examples**:

- **Clarity:** 88.10% → Somehow excellent at judging clarity in good ideas
- **Novelty:** 95.24% → Excellent at recognizing original and innovative ideas
- **Feasibility:** 97.62% → Excellent at confirming feasibility in valid ideas
- **Validness:** 78.05% → Very good accuracy on overall validity of good ideas

The model is little Struggling With Bad Examples, but majority ratio is more than 50%:

- Clarity: 44.4% →Struggles to mark poorly explained ideas as fail in majority.
- Novelty: 55.6% →Little better but still performing weak at catching unoriginal ideas.
- Feasibility: 55.56% → Little better here; but still most of the infeasible ideas are not caught.
- Validness: 55.56% →Performs somehow good but still fails to reject at an excellent level.

In Validness model is taking 11 seconds, in novelty model is taking around 10 seconds to process single idea, for Clarity time taken is around 3 seconds and in Feasibility the time period to process single is around around 2 seconds.

Detail Results

✅ ✅ means idea has that failing metric and model correctly marked this.

✅ means ideas hasn't have that failing metric and model correctly considered that ideas as passed by assigning a score >= 3.

❌ means model incorrectly marked.

Validness For Bad Examples tested with Gemma 3 4b

(Model's resource allocation was 1gb of RAM

To process 9 ideas, a model took 6 and half seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 9

Score Distribution:

Score 1: 1 idea(s)

Score 2: 4 idea(s)

Score 3: 3 idea(s)

Score 4: 1 idea(s)

Score 5: 0 idea(s)

Validness Score by Ideas:

- Building a ChatGPT Clone with OpenAI API: 3.00
- Using Decision Trees for Binary Classification: 4.00
- Generating Earthquake Predictions with ChatGPT: 1.00
- Training Neural Networks to Predict Earthquakes Using social media comments: 2.00
- Using LLMs to Classify Plant Species: 3.00
- Optimizing Sorting Algorithms with LLMs: 2.00
- Developing Artificial General Intelligence (AGI): 2.00
- Direct Brain-AI Communication Using Neural Implants: 3.00
- Improving Artificial Intelligence: 2.00

---- > **Average Overall Validness Score = 2.44**

Validness Evaluation Summary Based on Failing Metrics:

❌ 'Building a ChatGPT Clone with OpenAI API' scored 3.00 — Should be unclear, but marked as clear (failing: validness)

✅ 'Using Decision Trees for Binary Classification' scored 4.00 — Clear and validness is not a failing metric

❌ 'Generating Earthquake Predictions with ChatGPT' scored 1.00 — Valid idea but incorrectly marked as unclear (validness not in failing metric)

✅ ✅ 'Training Neural Networks to Predict Earthquakes Using social media comments' scored 2.00 — Correctly marked as unclear (failing: validness)

✅ 'Using LLMs to Classify Plant Species' scored 3.00 — Clear and validness is not a failing metric

❌ 'Optimizing Sorting Algorithms with LLMs' scored 2.00 — Valid idea but incorrectly marked as unclear (validness not in failing metric)

❌ 'Developing Artificial General Intelligence (AGI)' scored 2.00 — Valid idea but incorrectly marked as unclear (validness not in failing metric)

✅ 'Direct Brain-AI Communication Using Neural Implants' scored 3.00 — Clear and validness is not a failing metric

✅ ✅ 'Improving Artificial Intelligence' scored 2.00 — Correctly marked as unclear (failing: validness)

Accuracy:

✅ **Accuracy of analyzing validness correctly: 55.56% (5 out of 9 correct)**

Validness For Valid Examples tested with Gemma 3 4b

To process 42 ideas, a model took 16 seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 41

Json Decode error at this point (Ideas 32):
JSONDecodeError for idea: Machine Learning approach for Enterprise Data with a focus on SAPLeonardo

Raw response was:

{

"Rationale": "The thesis idea addresses a relevant and increasingly important problem – the integration of ML into enterprise environments. However, it currently lacks sufficient depth and specific methodological detail to warrant a high overall validity score. The core motivation is sound; enterprises *do* struggle with data extraction and need better ways to leverage their assets. The comparison between external vs. integrated ML solutions (TensorFlow vs. SAP Leonardo) is a good startin ...

(I tried to fix it using demjson3 or rapidJson library, but it wasn't fixed properly. The reason I found on internet is, sometimes lower parameter models lack following strict json instruction at some point).

Score Distribution:

Score 1: 3 idea(s)

Score 2: 6 idea(s)

Score 3: 30 idea(s)

Score 4: 2 idea(s)

Score 5: 0 idea(s)

Validness Score by Ideas:

- Artificial intelligence-assisted data analysis with BayesDB: 3.00
- Data analysis and simulation approach to capacity planning: 3.00
- Faster linear algebra for data analysis and machine learning: 3.00
- Emotional response modeling in financial markets : Boston Stock Exchange data analysis: 3.00
- Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?: 3.00
- Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy: 3.00
- Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes: 3.00
- Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models: 3.00
- Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models: 3.00
- Using LLMs to aid developers with code comprehension in codebases: 3.00
- Telepathic Machine Learning: Training AI Models with Brain Waves: 2.00
- Infinite Data Compression Using a Single Byte: 1.00
- The Square Root of a Cat: Applying Algebraic Structures to Living Organisms: 1.00
- Training a Neural Network Using Only White Noise: 2.00
- Reverse Evolution: Teaching Dinosaurs to Use Smartphones: 2.00
- Predicting Earthquake Locations Using Sentient AI Pigeons: 2.00
- Quantum Blockchain for Faster-than-Light Financial Transactions: 2.00
- Sentiment Analysis on Dolphin Communication Using Large Language Models: 3.00
- Using AI to Detect Ghosts in Abandoned Buildings: 2.00
- Infinite Battery Life Using Perpetual Motion Machines: 1.00
- Lead Scoring with Machine Learning: 4.00
- Using Machine Learning Methods for Evaluating the Quality of Technical Documents: 3.00

- Application of machine learning algorithms for classification and regression problems for mobile game monetization: 3.00
- Applying Machine Learning in Equity Trading: 3.00
- Predicting Default Loans using Machine Learning: 3.00
- Dynamic Model Selection for Automated Machine Learning in Time Series: 3.00
- Application of Machine Learning in Economic Optimization: 3.00
- Sanity Checks for Explanations of Deep Neural Networks Predictions: 4.00
- Machine Learning in Application-Based Case Management: 3.00
- Machine Learning for All: a Methodology for Choosing a Federated Learning Approach: 3.00

---- > **Average Overall Validness Score = 2.76**

Validness Evaluation Summary Based on Failing Metrics:

✅ 'Emotional response modeling in financial markets : Boston Stock Exchange data analysis' scored 3.00 — Correctly marked as valid

✅ 'Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?' scored 3.00 — Correctly marked as valid

✅ 'Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy' scored 3.00 — Correctly marked as valid

✅ 'Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes' scored 3.00 — Correctly marked as valid

✅ 'Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Using LLMs to aid developers with code comprehension in codebases' scored 3.00 — Correctly marked as valid

❌ 'Telepathic Machine Learning: Training AI Models with Brain Waves' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Infinite Data Compression Using a Single Byte' scored 1.00 — Should have passing score minimum >= 3 but got low score = 1.00

❌ 'The Square Root of a Cat: Applying Algebraic Structures to Living Organisms' scored 1.00 — Should have passing score minimum >= 3 but got low score = 1.00

❌ 'Training a Neural Network Using Only White Noise' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Reverse Evolution: Teaching Dinosaurs to Use Smartphones' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Predicting Earthquake Locations Using Sentient AI Pigeons' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Quantum Blockchain for Faster-than-Light Financial Transactions' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Sentiment Analysis on Dolphin Communication Using Large Language Models' scored 3.00 — Correctly marked as valid

❌ 'Using AI to Detect Ghosts in Abandoned Buildings' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Infinite Battery Life Using Perpetual Motion Machines' scored 1.00 — Should have passing score minimum >= 3 but got low score = 1.00

✅ 'Lead Scoring with Machine Learning' scored 4.00 — Correctly marked as valid

✅ 'Using Machine Learning Methods for Evaluating the Quality of Technical Documents' scored 3.00 — Correctly marked as valid

✅ 'Application of machine learning algorithms for classification and regression problems for mobile game monetization' scored 3.00 — Correctly marked as valid

✅ 'Applying Machine Learning in Equity Trading' scored 3.00 — Correctly marked as valid

✅ 'Predicting Default Loans using Machine Learning' scored 3.00 — Correctly marked as valid

✅ 'Dynamic Model Selection for Automated Machine Learning in Time Series' scored 3.00 — Correctly marked as valid

✅ 'Application of Machine Learning in Economic Optimization' scored 3.00 — Correctly marked as valid

✅ 'Sanity Checks for Explanations of Deep Neural Networks Predictions' scored 4.00 — Correctly marked as valid

✅ 'Machine Learning in Application-Based Case Management' scored 3.00 — Correctly marked as valid

✅ 'Machine Learning for All: a Methodology for Choosing a Federated Learning Approach' scored 3.00 — Correctly marked as valid

Accuracy:

✅ **Accuracy of analyzing validness correctly: 78.05 % (32 out of 41 correct)**

Novelty For Bad Examples tested with QWQ 32b

To process 9 ideas, a model took 5 seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 9

Score Distribution:

Score 1: 2 idea(s)

Score 2: 2 idea(s)

Score 3: 4 idea(s)

Score 4: 1 idea(s)

Score 5: 0 idea(s)

Dimension Averages:

ProblemNovelty: 2.33

MethodologicalInnovation: 2.11

PotentialImpact: 2.56

CombinationUniqueness: 2.44

Novelty Score by Ideas:

- Building a ChatGPT Clone with OpenAI API: 2.00
- Using Decision Trees for Binary Classification: 1.00
- Generating Earthquake Predictions with ChatGPT: 2.00
- Training Neural Networks to Predict Earthquakes Using social media comments: 3.00
- Using LLMs to Classify Plant Species: 3.00
- Optimizing Sorting Algorithms with LLMs: 3.00
- Developing Artificial General Intelligence (AGI): 4.00
- Direct Brain-AI Communication Using Neural Implants: 3.00
- Improving Artificial Intelligence: 1.00

---- > **Average Overall Novelty Score = 2.44**

Novelty Evaluation Summary Based on Failing Metrics:

❌ 'Building a ChatGPT Clone with OpenAI API' scored 2.00 — Valid idea but incorrectly marked as unclear (novelty not in failing metric)

✅ ✅ 'Using Decision Trees for Binary Classification' scored 1.00 — Correctly marked as unclear (failing: novelty)

✅ ✅ 'Generating Earthquake Predictions with ChatGPT' scored 2.00 — Correctly marked as unclear (failing: novelty)

✅ 'Training Neural Networks to Predict Earthquakes Using social media comments' scored 3.00 — Clear and novelty is not a failing metric

✅ 'Using LLMs to Classify Plant Species' scored 3.00 — Clear and novelty is not a failing metric

✅ 'Optimizing Sorting Algorithms with LLMs' scored 3.00 — Clear and novelty is not a failing metric

✅ 'Developing Artificial General Intelligence (AGI)' scored 4.00 — Clear and novelty is not a failing metric

✅ 'Direct Brain-AI Communication Using Neural Implants' scored 3.00 — Clear and novelty is not a failing metric

❌ 'Improving Artificial Intelligence' scored 1.00 — Valid idea but incorrectly marked as unclear (novelty not in failing metric)

Accuracy:

✅ **Accuracy of analyzing Novelty correctly: 77.78% (7 out of 9 correct)**

Novelty For Valid Examples tested with Gemma 3 4b

To process 42 ideas, a model took 8 seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 42

Score Distribution:

Score 1: 2 idea(s)

Score 2: 2 idea(s)

Score 3: 36 idea(s)

Score 4: 2 idea(s)

Score 5: 0 idea(s)

Dimension Averages:

ProblemNovelty: 2.95

MethodologicalInnovation: 2.83

PotentialImpact: 2.81

CombinationUniqueness: 3.12

Novelty Score by Ideas:

• Generative Adversarial Networks for Multi-Instrument Music Synthesis: 3.00

• Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality: 3.00

• Self-supervised Domain Adaptation of Language Models for the Process Industry: 3.00

• Deep Learning Techniques Applied to Constituency Parsing of German: 3.00

• Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments: 3.00

• Graph Neural Networks for Electrical Grid State Estimation: 3.00

• Representation Learning on Electronic Health Records Using Graph Neural Networks: 3.00

• Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems: 3.00

• Solving Machine Learning Problems: 3.00

• Optimization Methods for Machine Learning underStructural Constraints: 3.00

• Probabilistic data analysis with probabilistic programming: 4.00

• Artificial intelligence-assisted data analysis with BayesDB: 3.00

• Data analysis and simulation approach to capacity planning: 3.00

• Faster linear algebra for data analysis and machine learning: 3.00

• Emotional response modeling in financial markets : Boston Stock Exchange data analysis: 3.00

• Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?: 3.00

• Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy: 3.00

• Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes: 3.00

• Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models: 3.00

- Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models: 3.00
- Using LLMs to aid developers with code comprehension in codebases: 3.00
- Telepathic Machine Learning: Training AI Models with Brain Waves: 3.00
- Infinite Data Compression Using a Single Byte: 1.00
- The Square Root of a Cat: Applying Algebraic Structures to Living Organisms: 2.00
- Training a Neural Network Using Only White Noise: 3.00
- Reverse Evolution: Teaching Dinosaurs to Use Smartphones: 4.00
- Predicting Earthquake Locations Using Sentient AI Pigeons: 2.00
- Quantum Blockchain for Faster-than-Light Financial Transactions: 3.00
- Sentiment Analysis on Dolphin Communication Using Large Language Models: 3.00
- Using AI to Detect Ghosts in Abandoned Buildings: 3.00
- Infinite Battery Life Using Perpetual Motion Machines: 1.00
- Machine Learning approach for Enterprise Data with a focus on SAPLeonardo: 3.00
- Lead Scoring with Machine Learning: 3.00
- Using Machine Learning Methods for Evaluating the Quality of Technical Documents: 3.00
- Application of machine learning algorithms for classification and regression problems for mobile game monetization: 3.00
- Applying Machine Learning in Equity Trading: 3.00
- Predicting Default Loans using Machine Learning: 3.00
- Dynamic Model Selection for Automated Machine Learning in Time Series: 3.00
- Application of Machine Learning in Economic Optimization: 3.00
- Sanity Checks for Explanations of Deep Neural Networks Predictions: 3.00
- Machine Learning in Application-Based Case Management: 3.00
- Machine Learning for All: a Methodology for Choosing a Federated Learning Approach: 3.00

---- > **Average Overall Novelty Score = 2.90**

Novelty Evaluation Summary Based on Failing Metrics:

✅ 'Generative Adversarial Networks for Multi-Instrument Music Synthesis' scored 3.00 — Correctly marked as valid

✅ 'Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality' scored 3.00 — Correctly marked as valid

✅ 'Self-supervised Domain Adaptation of Language Models for the Process Industry' scored 3.00 — Correctly marked as valid

✅ 'Deep Learning Techniques Applied to Constituency Parsing of German' scored 3.00 — Correctly marked as valid

✅ 'Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments' scored 3.00 — Correctly marked as valid

✅ 'Graph Neural Networks for Electrical Grid State Estimation' scored 3.00 — Correctly marked as valid

✅ 'Representation Learning on Electronic Health Records Using Graph Neural Networks' scored 3.00 — Correctly marked as valid

✅ 'Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems' scored 3.00 — Correctly marked as valid

✅ 'Solving Machine Learning Problems' scored 3.00 — Correctly marked as valid

✅ 'Optimization Methods for Machine Learning underStructural Constraints' scored 3.00 — Correctly marked as valid

✅ 'Probabilistic data analysis with probabilistic programming' scored 4.00 — Correctly marked as valid

✅ 'Artificial intelligence-assisted data analysis with BayesDB' scored 3.00 — Correctly marked as valid

✅ 'Data analysis and simulation approach to capacity planning' scored 3.00 — Correctly marked as valid

✅ 'Faster linear algebra for data analysis and machine learning' scored 3.00 — Correctly marked as valid

✅ 'Emotional response modeling in financial markets : Boston Stock Exchange data analysis' scored 3.00 — Correctly marked as valid

✅ 'Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?' scored 3.00 — Correctly marked as valid

✅ 'Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy' scored 3.00 — Correctly marked as valid

✅ 'Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes' scored 3.00 — Correctly marked as valid

✅ 'Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Using LLMs to aid developers with code comprehension in codebases' scored 3.00 — Correctly marked as valid

✅ 'Telepathic Machine Learning: Training AI Models with Brain Waves' scored 3.00 — Correctly marked as valid

❌ 'Infinite Data Compression Using a Single Byte' scored 1.00 — Should have passing score minimum >= 3 but got low score = 1.00

❌ 'The Square Root of a Cat: Applying Algebraic Structures to Living Organisms' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Training a Neural Network Using Only White Noise' scored 3.00 — Correctly marked as valid

✅ 'Reverse Evolution: Teaching Dinosaurs to Use Smartphones' scored 4.00 — Correctly marked as valid

❌ 'Predicting Earthquake Locations Using Sentient AI Pigeons' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Quantum Blockchain for Faster-than-Light Financial Transactions' scored 3.00 — Correctly marked as valid

✅ 'Sentiment Analysis on Dolphin Communication Using Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Using AI to Detect Ghosts in Abandoned Buildings' scored 3.00 — Correctly marked as valid

❌ 'Infinite Battery Life Using Perpetual Motion Machines' scored 1.00 — Should have passing score minimum >= 3 but got low score = 1.00

✅ 'Machine Learning approach for Enterprise Data with a focus on SAPLeonardo' scored 3.00 — Correctly marked as valid

✅ 'Lead Scoring with Machine Learning' scored 3.00 — Correctly marked as valid

✅ 'Using Machine Learning Methods for Evaluating the Quality of Technical Documents' scored 3.00 — Correctly marked as valid

✅ 'Application of machine learning algorithms for classification and regression problems for mobile game monetization' scored 3.00 — Correctly marked as valid

✅ 'Applying Machine Learning in Equity Trading' scored 3.00 — Correctly marked as valid

✅ 'Predicting Default Loans using Machine Learning' scored 3.00 — Correctly marked as valid

✅ 'Dynamic Model Selection for Automated Machine Learning in Time Series' scored 3.00 — Correctly marked as valid

✅ 'Application of Machine Learning in Economic Optimization' scored 3.00 — Correctly marked as valid

✅ 'Sanity Checks for Explanations of Deep Neural Networks Predictions' scored 3.00 — Correctly marked as valid

✅ 'Machine Learning in Application-Based Case Management' scored 3.00 — Correctly marked as valid

✅ 'Machine Learning for All: a Methodology for Choosing a Federated Learning Approach' scored 3.00 — Correctly marked as valid

Accuracy:

✅ **Accuracy of analyzing Novelty correctly: 90.48% (38 out of 42 correct)**

Clarity For Bad Examples tested with Gemma 3 4b

To process 9 ideas, a model took 3 seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 9

Score Distribution:

Score 1: 1 idea(s)

Score 2: 7 idea(s)

Score 3: 1 idea(s)

Score 4: 0 idea(s)

Score 5: 0 idea(s)

Clarity Score by Ideas:

- Building a ChatGPT Clone with OpenAI API: 2.00
- Using Decision Trees for Binary Classification: 2.00
- Generating Earthquake Predictions with ChatGPT: 2.00
- Training Neural Networks to Predict Earthquakes Using social media comments: 2.00
- Using LLMs to Classify Plant Species: 3.00
- Optimizing Sorting Algorithms with LLMs: 2.00
- Developing Artificial General Intelligence (AGI): 2.00
- Direct Brain-AI Communication Using Neural Implants: 2.00

- Improving Artificial Intelligence: 1.00

---- > **Average Overall Clarity Score = 2.00**

Clarity Evaluation Summary Based on Failing Metrics:

✅ ✅ 'Building a ChatGPT Clone with OpenAI API' scored 2.00 — Correctly marked as unclear (failing: clarity)

✅ ✅ 'Using Decision Trees for Binary Classification' scored 2.00 — Correctly marked as unclear (failing: clarity)

❌ 'Generating Earthquake Predictions with ChatGPT' scored 2.00 — Valid idea but incorrectly marked as unclear (clarity not in failing metric)

❌ 'Training Neural Networks to Predict Earthquakes Using social media comments' scored 2.00 — Valid idea but incorrectly marked as unclear (clarity not in failing metric)

❌ 'Using LLMs to Classify Plant Species' scored 3.00 — Should be unclear, but marked as clear (failing: clarity)

✅ ✅ 'Optimizing Sorting Algorithms with LLMs' scored 2.00 — Correctly marked as unclear (failing: clarity)

❌ 'Developing Artificial General Intelligence (AGI)' scored 2.00 — Valid idea but incorrectly marked as unclear (clarity not in failing metric)

❌ 'Direct Brain-AI Communication Using Neural Implants' scored 2.00 — Valid idea but incorrectly marked as unclear (clarity not in failing metric)

✅ ✅ 'Improving Artificial Intelligence' scored 1.00 — Correctly marked as unclear (failing: clarity)

Accuracy:

✅ **Accuracy of analyzing Clarity correctly: 44.44% (4 out of 9 correct)**

Clarity For Valid Examples tested with Gemma 3 4b

To process 42 ideas, a model took less than 1 second approx. to process each idea.

Analysis Results:

Total Valid Entries: 42

Score Distribution:

Score 1: 0 idea(s)

Score 2: 5 idea(s)

Score 3: 20 idea(s)

Score 4: 17 idea(s)

Clarity Score by Ideas:

- Generative Adversarial Networks for Multi-Instrument Music Synthesis: 4.00
- Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality: 3.00
- Self-supervised Domain Adaptation of Language Models for the Process Industry: 4.00
- Deep Learning Techniques Applied to Constituency Parsing of German: 3.00
- Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments: 3.00
- Graph Neural Networks for Electrical Grid State Estimation: 3.00
- Representation Learning on Electronic Health Records Using Graph Neural Networks: 3.00
- Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems: 3.00
- Solving Machine Learning Problems: 3.00
- Optimization Methods for Machine Learning underStructural Constraints: 4.00
- Probabilistic data analysis with probabilistic programming: 4.00
- Artificial intelligence-assisted data analysis with BayesDB: 4.00
- Data analysis and simulation approach to capacity planning: 4.00
- Faster linear algebra for data analysis and machine learning: 4.00
- Emotional response modeling in financial markets : Boston Stock Exchange data analysis: 4.00
- Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?: 3.00
- Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy: 3.00
- Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes: 3.00
- Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models: 3.00
- Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models: 3.00

- Using LLMs to aid developers with code comprehension in codebases: 3.00
- Telepathic Machine Learning: Training AI Models with Brain Waves: 2.00
- Infinite Data Compression Using a Single Byte: 2.00
- The Square Root of a Cat: Applying Algebraic Structures to Living Organisms: 2.00
- Training a Neural Network Using Only White Noise: 3.00
- Reverse Evolution: Teaching Dinosaurs to Use Smartphones: 3.00
- Predicting Earthquake Locations Using Sentient AI Pigeons: 3.00
- Quantum Blockchain for Faster-than-Light Financial Transactions: 2.00
- Sentiment Analysis on Dolphin Communication Using Large Language Models: 3.00
- Using AI to Detect Ghosts in Abandoned Buildings: 3.00
- Infinite Battery Life Using Perpetual Motion Machines: 2.00
- Machine Learning approach for Enterprise Data with a focus on SAPLeonardo: 3.00
- Lead Scoring with Machine Learning: 4.00
- Using Machine Learning Methods for Evaluating the Quality of Technical Documents: 4.00
- Application of machine learning algorithms for classification and regression problems for mobile game monetization: 4.00
- Applying Machine Learning in Equity Trading: 3.00
- Predicting Default Loans using Machine Learning: 4.00
- Dynamic Model Selection for Automated Machine Learning in Time Series: 4.00
- Application of Machine Learning in Economic Optimization: 4.00
- Sanity Checks for Explanations of Deep Neural Networks Predictions: 4.00
- Machine Learning in Application-Based Case Management: 4.00
- Machine Learning for All: a Methodology for Choosing a Federated Learning Approach: 4.00

---- > **Average Overall Novelty Score = 3.29**

Clarity Evaluation Summary Based on Failing Metrics:

✅ 'Generative Adversarial Networks for Multi-Instrument Music Synthesis' scored 4.00 — Correctly marked as valid

✅ 'Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality' scored 3.00 — Correctly marked as valid

✅ 'Self-supervised Domain Adaptation of Language Models for the Process Industry' scored 4.00 — Correctly marked as valid

✅ 'Deep Learning Techniques Applied to Constituency Parsing of German' scored 3.00 — Correctly marked as valid

✅ 'Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments' scored 3.00 — Correctly marked as valid

✅ 'Graph Neural Networks for Electrical Grid State Estimation' scored 3.00 — Correctly marked as valid

✅ 'Representation Learning on Electronic Health Records Using Graph Neural Networks' scored 3.00 — Correctly marked as valid

✅ 'Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems' scored 3.00 — Correctly marked as valid

✅ 'Solving Machine Learning Problems' scored 3.00 — Correctly marked as valid

✅ 'Optimization Methods for Machine Learning underStructural Constraints' scored 4.00 — Correctly marked as valid

✅ 'Probabilistic data analysis with probabilistic programming' scored 4.00 — Correctly marked as valid

✅ 'Artificial intelligence-assisted data analysis with BayesDB' scored 4.00 — Correctly marked as valid

✅ 'Data analysis and simulation approach to capacity planning' scored 4.00 — Correctly marked as valid

✅ 'Faster linear algebra for data analysis and machine learning' scored 4.00 — Correctly marked as valid

✅ 'Emotional response modeling in financial markets : Boston Stock Exchange data analysis' scored 4.00 — Correctly marked as valid

✅ 'Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?' scored 3.00 — Correctly marked as valid

✅ 'Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy' scored 3.00 — Correctly marked as valid

✅ 'Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes' scored 3.00 — Correctly marked as valid

✅ 'Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Using LLMs to aid developers with code comprehension in codebases' scored 3.00 — Correctly marked as valid

❌ 'Telepathic Machine Learning: Training AI Models with Brain Waves' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'Infinite Data Compression Using a Single Byte' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

❌ 'The Square Root of a Cat: Applying Algebraic Structures to Living Organisms' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Training a Neural Network Using Only White Noise' scored 3.00 — Correctly marked as valid

✅ 'Reverse Evolution: Teaching Dinosaurs to Use Smartphones' scored 3.00 — Correctly marked as valid

✅ 'Predicting Earthquake Locations Using Sentient AI Pigeons' scored 3.00 — Correctly marked as valid

❌ 'Quantum Blockchain for Faster-than-Light Financial Transactions' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Sentiment Analysis on Dolphin Communication Using Large Language Models' scored 3.00 — Correctly marked as valid

✅ 'Using AI to Detect Ghosts in Abandoned Buildings' scored 3.00 — Correctly marked as valid

❌ 'Infinite Battery Life Using Perpetual Motion Machines' scored 2.00 — Should have passing score minimum >= 3 but got low score = 2.00

✅ 'Machine Learning approach for Enterprise Data with a focus on SAPLeonardo' scored 3.00 — Correctly marked as valid

✅ 'Lead Scoring with Machine Learning' scored 4.00 — Correctly marked as valid

✅ 'Using Machine Learning Methods for Evaluating the Quality of Technical Documents' scored 4.00 — Correctly marked as valid

✅ 'Application of machine learning algorithms for classification and regression problems for mobile game monetization' scored 4.00 — Correctly marked as valid

✅ 'Applying Machine Learning in Equity Trading' scored 3.00 — Correctly marked as valid

✅ 'Predicting Default Loans using Machine Learning' scored 4.00 — Correctly marked as valid

✅ 'Dynamic Model Selection for Automated Machine Learning in Time Series' scored 4.00 — Correctly marked as valid

✅ 'Application of Machine Learning in Economic Optimization' scored 4.00 — Correctly marked as valid

✅ 'Sanity Checks for Explanations of Deep Neural Networks Predictions' scored 4.00 — Correctly marked as valid

✅ 'Machine Learning in Application-Based Case Management' scored 4.00 — Correctly marked as valid

✅ 'Machine Learning for All: a Methodology for Choosing a Federated Learning Approach' scored 4.00 — Correctly marked as valid

Accuracy:

✅ **Accuracy of analyzing Novelty correctly: 88.10% (37 out of 42 correct)**

Feasibility For Bad Examples tested with Gemma 3 4b

To process 9 ideas, a model took 2 seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 9

Score Distribution:

Score 1: 0 idea(s)

Score 2: 4 idea(s)

Score 3: 4 idea(s)

Score 4: 1 idea(s)

Score 5: 0 idea(s)

Feasibility Score by Ideas:

- Building a ChatGPT Clone with OpenAI API: 3.00
- Using Decision Trees for Binary Classification: 4.00
- Generating Earthquake Predictions with ChatGPT: 2.00
- Training Neural Networks to Predict Earthquakes Using social media comments: 3.00
- Using LLMs to Classify Plant Species: 3.00
- Optimizing Sorting Algorithms with LLMs: 3.00

- Developing Artificial General Intelligence (AGI): 2.00
- Direct Brain-AI Communication Using Neural Implants: 2.00
- Improving Artificial Intelligence: 2.00

---- > **Average Overall Feasibility Score = 2.67**

Feasibility Evaluation Summary Based on Failing Metrics:

✅ 'Building a ChatGPT Clone with OpenAI API' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Using Decision Trees for Binary Classification' scored 4.00 — Clear and feasibility is not a failing metric

❌ 'Generating Earthquake Predictions with ChatGPT' scored 2.00 — Valid idea but incorrectly marked as unclear (feasibility not in failing metric)

✅ 'Training Neural Networks to Predict Earthquakes Using social media comments' scored 3.00 — Clear and feasibility is not a failing metric

❌ 'Using LLMs to Classify Plant Species' scored 3.00 — Should be unclear, but marked as clear (failing: feasibility)

❌ 'Optimizing Sorting Algorithms with LLMs' scored 3.00 — Should be unclear, but marked as clear (failing: feasibility)

✅ ✅ 'Developing Artificial General Intelligence (AGI)' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Direct Brain-AI Communication Using Neural Implants' scored 2.00 — Correctly marked as unclear (failing: feasibility)

❌ 'Improving Artificial Intelligence' scored 2.00 — Valid idea but incorrectly marked as unclear (feasibility not in failing metric)

Accuracy:

✅ **Accuracy of analyzing Clarity correctly: 55.56% (5 out of 9 correct)**

Feasibility For Valid Examples tested with Gemma 3 4b

To process 42 ideas, a model took 1 and half seconds approx. to process each idea.

Analysis Results:

Total Valid Entries: 42

Score Distribution:

Score 1: 3 idea(s)

Score 2: 6 idea(s)

Score 3: 29 idea(s)

Score 4: 4 idea(s)

Score 5: 0 idea(s)

Feasibility Score by Ideas:

- Generative Adversarial Networks for Multi-Instrument Music Synthesis: 3.00
- Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality: 3.00
- Self-supervised Domain Adaptation of Language Models for the Process Industry: 3.00
- Deep Learning Techniques Applied to Constituency Parsing of German: 3.00
- Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments: 3.00
- Graph Neural Networks for Electrical Grid State Estimation: 3.00
- Representation Learning on Electronic Health Records Using Graph Neural Networks: 4.00
- Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems: 3.00
- Solving Machine Learning Problems: 3.00
- Optimization Methods for Machine Learning underStructural Constraints: 3.00
- Probabilistic data analysis with probabilistic programming: 3.00
- Artificial intelligence-assisted data analysis with BayesDB: 3.00
- Data analysis and simulation approach to capacity planning: 3.00
- Faster linear algebra for data analysis and machine learning: 3.00
- Emotional response modeling in financial markets : Boston Stock Exchange data analysis: 3.00
- Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?: 3.00
- Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy: 3.00
- Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes: 3.00

• Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models: 3.00

• Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models: 3.00

• Using LLMs to aid developers with code comprehension in codebases: 3.00

• Telepathic Machine Learning: Training AI Models with Brain Waves: 2.00

• Infinite Data Compression Using a Single Byte: 1.00

• The Square Root of a Cat: Applying Algebraic Structures to Living Organisms: 1.00

• Training a Neural Network Using Only White Noise: 2.00

• Reverse Evolution: Teaching Dinosaurs to Use Smartphones: 2.00

• Predicting Earthquake Locations Using Sentient AI Pigeons: 2.00

• Quantum Blockchain for Faster-than-Light Financial Transactions: 2.00

• Sentiment Analysis on Dolphin Communication Using Large Language Models: 3.00

• Using AI to Detect Ghosts in Abandoned Buildings: 2.00

• Infinite Battery Life Using Perpetual Motion Machines: 1.00

• Machine Learning approach for Enterprise Data with a focus on SAPLeonardo: 3.00

• Lead Scoring with Machine Learning: 4.00

• Using Machine Learning Methods for Evaluating the Quality of Technical Documents: 3.00

• Application of machine learning algorithms for classification and regression problems for mobile game monetization: 4.00

• Applying Machine Learning in Equity Trading: 3.00

• Predicting Default Loans using Machine Learning: 4.00

• Dynamic Model Selection for Automated Machine Learning in Time Series: 3.00

• Application of Machine Learning in Economic Optimization: 3.00

• Sanity Checks for Explanations of Deep Neural Networks Predictions: 3.00

• Machine Learning in Application-Based Case Management: 3.00

• Machine Learning for All: a Methodology for Choosing a Federated Learning Approach: 3.00

---- > **Average Overall Novelty Score = 2.81**

Feasibility Evaluation Summary Based on Failing Metrics:

✅ 'Generative Adversarial Networks for Multi-Instrument Music Synthesis' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Machine Learning Image Segmentation to Improve Object Recognition in Mixed Reality' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Self-supervised Domain Adaptation of Language Models for the Process Industry' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Deep Learning Techniques Applied to Constituency Parsing of German' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Applying Deep Reinforcement Learning in the Navigation of Mobile Robots in Static and Dynamic Environments' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Graph Neural Networks for Electrical Grid State Estimation' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Representation Learning on Electronic Health Records Using Graph Neural Networks' scored 4.00 — Clear and feasibility is not a failing metric

✅ 'Deep Reinforcement Learning for Decentralized Autonomous Decision-Making in Federated Satellite Systems' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Solving Machine Learning Problems' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Optimization Methods for Machine Learning underStructural Constraints' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Probabilistic data analysis with probabilistic programming' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Artificial intelligence-assisted data analysis with BayesDB' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Data analysis and simulation approach to capacity planning' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Faster linear algebra for data analysis and machine learning' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Emotional response modeling in financial markets : Boston Stock Exchange data analysis' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Evaluating Large Language Models for Automated Cyber Security Alarm Analysis Processes' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Automatic Evaluation of Companies' Alignment with EU Taxonomy Using Large Language Models' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Variational Auto-Encoder for Latent Uncertainty Encoding in Large Language Models' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Using LLMs to aid developers with code comprehension in codebases' scored 3.00 — Clear and feasibility is not a failing metric

✅ ✅ 'Telepathic Machine Learning: Training AI Models with Brain Waves' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Infinite Data Compression Using a Single Byte' scored 1.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'The Square Root of a Cat: Applying Algebraic Structures to Living Organisms' scored 1.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Training a Neural Network Using Only White Noise' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Reverse Evolution: Teaching Dinosaurs to Use Smartphones' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Predicting Earthquake Locations Using Sentient AI Pigeons' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Quantum Blockchain for Faster-than-Light Financial Transactions' scored 2.00 — Correctly marked as unclear (failing: feasibility)

❌ 'Sentiment Analysis on Dolphin Communication Using Large Language Models' scored 3.00 — Should be unclear, but marked as clear (failing: feasibility)

✅ ✅ 'Using AI to Detect Ghosts in Abandoned Buildings' scored 2.00 — Correctly marked as unclear (failing: feasibility)

✅ ✅ 'Infinite Battery Life Using Perpetual Motion Machines' scored 1.00 — Correctly marked as unclear (failing: feasibility)

✅ 'Machine Learning approach for Enterprise Data with a focus on SAPLeonardo' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Lead Scoring with Machine Learning' scored 4.00 — Clear and feasibility is not a failing metric

✅ 'Using Machine Learning Methods for Evaluating the Quality of Technical Documents' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Application of machine learning algorithms for classification and regression problems for mobile game monetization' scored 4.00 — Clear and feasibility is not a failing metric

✅ 'Applying Machine Learning in Equity Trading' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Predicting Default Loans using Machine Learning' scored 4.00 — Clear and feasibility is not a failing metric

✅ 'Dynamic Model Selection for Automated Machine Learning in Time Series' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Application of Machine Learning in Economic Optimization' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Sanity Checks for Explanations of Deep Neural Networks Predictions' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Machine Learning in Application-Based Case Management' scored 3.00 — Clear and feasibility is not a failing metric

✅ 'Machine Learning for All: a Methodology for Choosing a Federated Learning Approach' scored 3.00 — Clear and feasibility is not a failing metric

Accuracy:

✅ **Accuracy of analyzing Novelty correctly: 97.62% (41 out of 42 correct)**