Producing Explainable AI with Large Language Models

Matthew Renze

AI Ethics

Question:

Can we use Large Language Models (LLMs)
to explain AI predictions in natural language?

Background

Background

Explainable AI (XAI)

Background

Explainable AI (XAI)

Natural language explanation (NLE)

Background

Explainable AI (XAI)

Natural language explanation (NLE)

Large Language Model (LLM)

Background

Explainable AI (XAI)

Natural language explanation (NLE)

Large Language Model (LLM)

Generative Pretrained Transformer (GPT)

Background

Explainable AI (XAI)

Natural language explanation (NLE)

Large Language Model (LLM)

Generative Pretrained Transformer (GPT)

COMPAS Synthetic Database

Background

Explainable AI (XAI)

Natural language explanation (NLE)

Large Language Model (LLM)

Generative Pretrained Transformer (GPT)

COMPAS Synthetic Database

SHapley Additive exPlanations (SHAP)

Scaled SHAP Values

c_charge_degree_importance: 0.000199

criminal_attitude_importance: 0.000004

family_criminality_importance: 0.000002

…

criminal_associates_importance: -0.062260

substance_abuse_importance: -0.099101

noncompliance_importance: -0.109318

Methods

Experiment

Experiment

Sample 100 case records

Experiment

Sample 100 case records

Generate explanations

Experiment

Sample 100 case records

Generate explanations

Correct explanations

Experiment

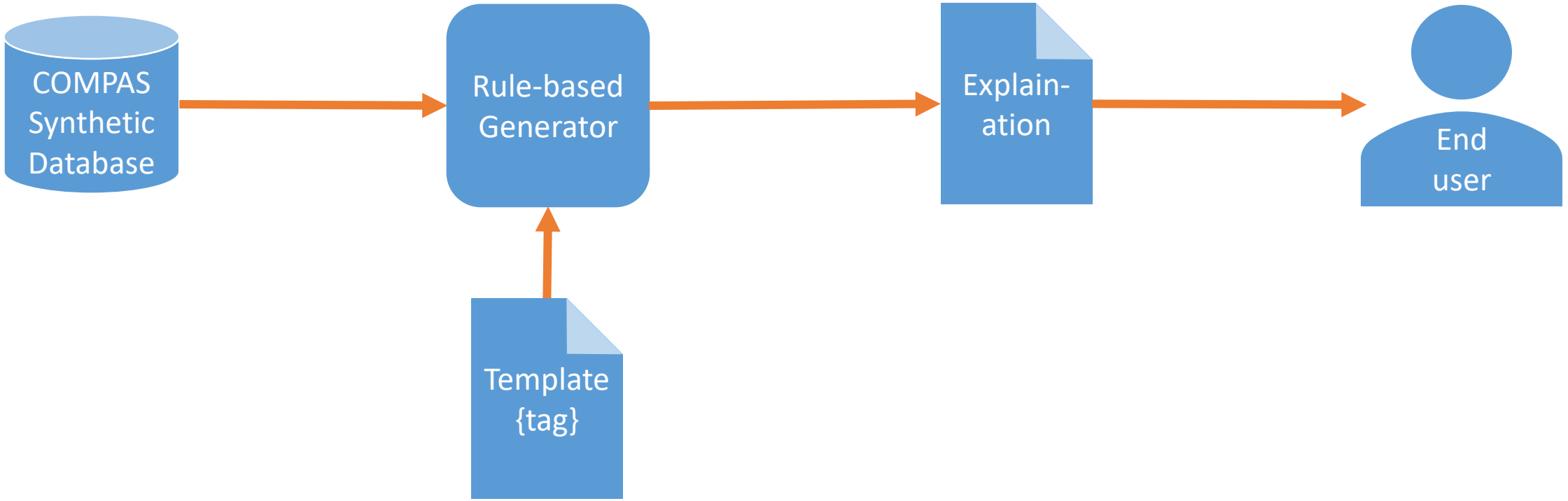
Sample 100 case records

Generate explanations

Correct explanations

Evaluate performance

Rule-based NLEs



{first-name},

Thank you for sharing your case record for our review.

Regarding your case record #{case-id}, you had a COMPAS risk score of {decile_score} on a scale of 1-10. This means you were scored as a {decile-score-level} risk for reoffending.

You received this score primarily because you had a {feature-level} risk score for {feature-name} ({feature-score}) {and | ,} {...}. {This was | These were} the {high-importance-feature-count} most important factors in receiving a {decile-score-level} overall risk score.

You also had a {feature-level} risk score for {feature-name} ({feature-score}) {and | ,} {...}. However, {this factor | these factors} only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including {feature-name} ({feature-score}) {and | ,} {...}, etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

{first-name},

Thank you for sharing your case record for our review.

Regarding your case record #{case-id}, you had a COMPAS risk score of {decile_score} on a scale of 1-10. This means you were scored as a {decile-score-level} risk for reoffending.

You received this score primarily because you had a {feature-level} risk score for {feature-name} ({feature-score}) {and | ,} {...}. {This was | These were} the {high-importance-feature-count} most important factors in receiving a {decile-score-level} overall risk score.

You also had a {feature-level} risk score for {feature-name} ({feature-score}) {and | ,} {...}. However, {this factor | these factors} only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including {feature-name} ({feature-score}) {and | ,} {...}, etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for noncompliance (6) and a high risk score for substance abuse (8). These were the two most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a high risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

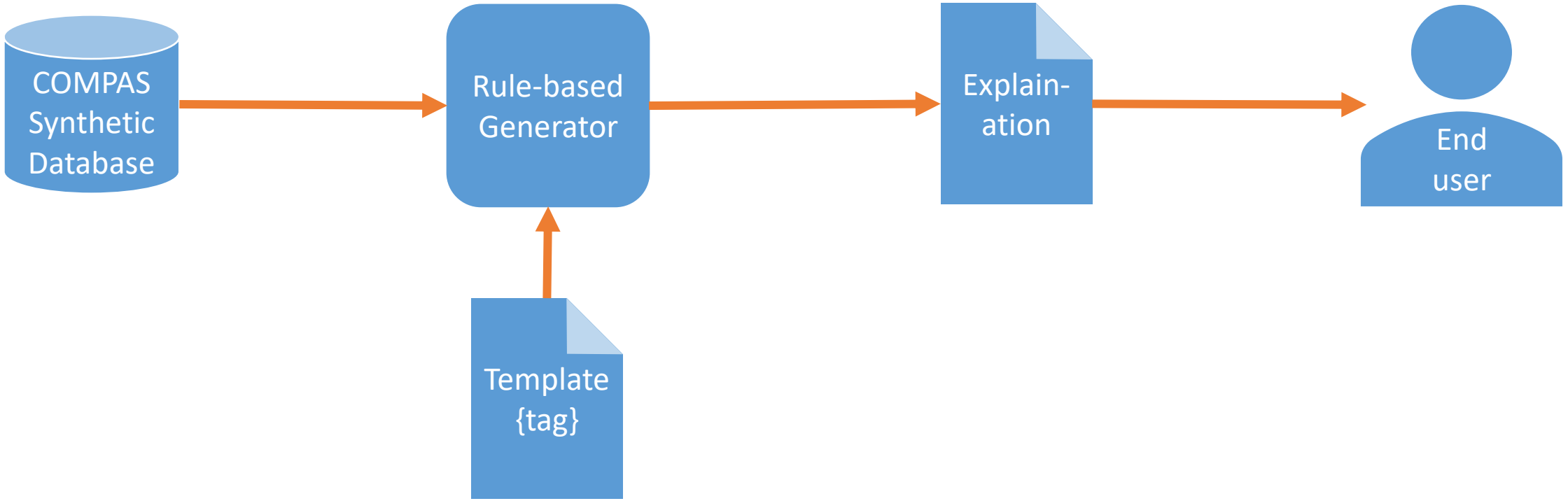
The remaining risk factors including financial problems (3), vocational risk factors (4), and social environment risk (5), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

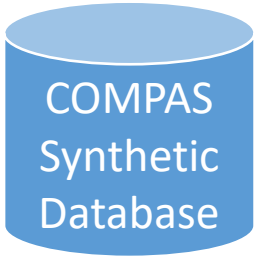
Sincerely,
Explainable AI

Rule-based NLEs

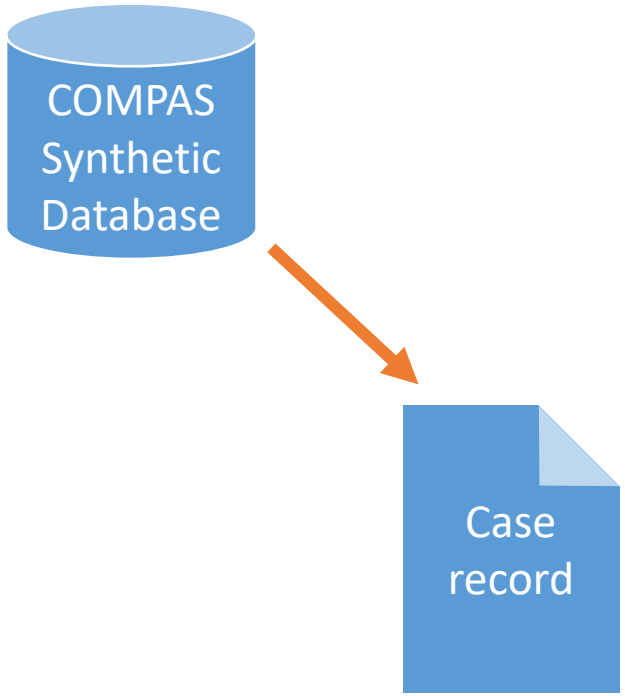


LLM-based NLEs

LLM-based NLEs



LLM-based NLEs



```
# Case Info
id: 142
name: richard veach
decile_score: 8
decile_score_level: high

# Features
priors_count: 9
c_charge_degree: F
juv_fel_count: 0
family_criminality: 5
criminal_attitude: 8
criminal_associates: 7
financial_problems: 3
substance_abuse: 8
noncompliance: 6
social_environment: 5
vocational: 4
```

```
# Case Info
id: 142
name: richard veach
decile_score: 8
decile_score_level: high

# Features
priors_count: 9
c_charge_degree: F
juv_fel_count: 0
family_criminality: 5
criminal_attitude: 8
criminal_associates: 7
financial_problems: 3
substance_abuse: 8
noncompliance: 6
social_environment: 5
vocational: 4
```

```
# Feature Levels
priors_count_level: high
c_charge_degree_level: high
juv_fel_count_level: low
family_criminality_level: medium
criminal_attitude_level: high
criminal_associates_level: high
financial_problems_level: low
substance_abuse_level: high
noncompliance_level: high
social_environment_level: medium
vocational_level: medium
```

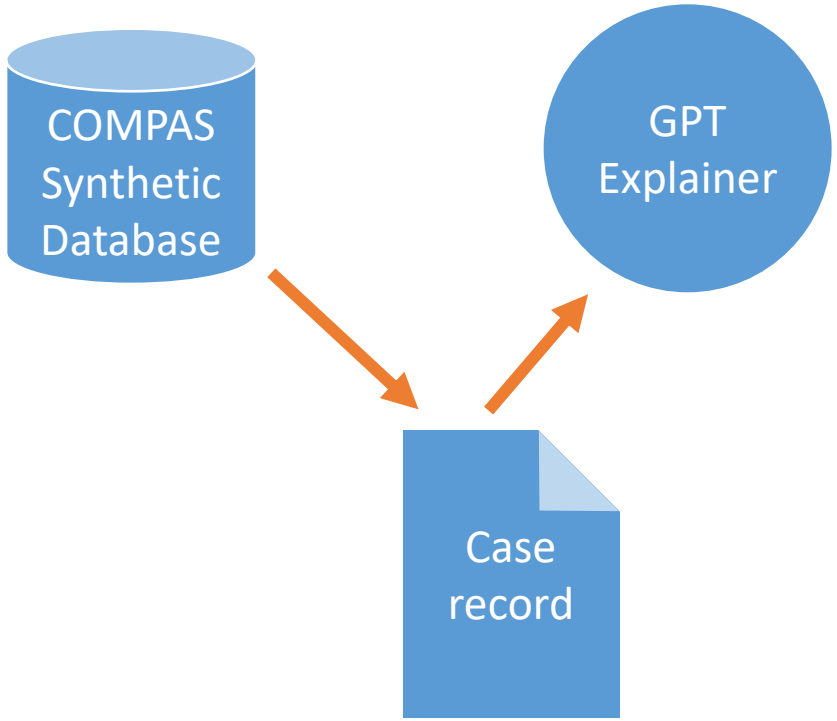
```
# Case Info
id: 142
name: richard veach
decile_score: 8
decile_score_level: high

# Features
priors_count: 9
c_charge_degree: F
juv_fel_count: 0
family_criminality: 5
criminal_attitude: 8
criminal_associates: 7
financial_problems: 3
substance_abuse: 8
noncompliance: 6
social_environment: 5
vocational: 4
```

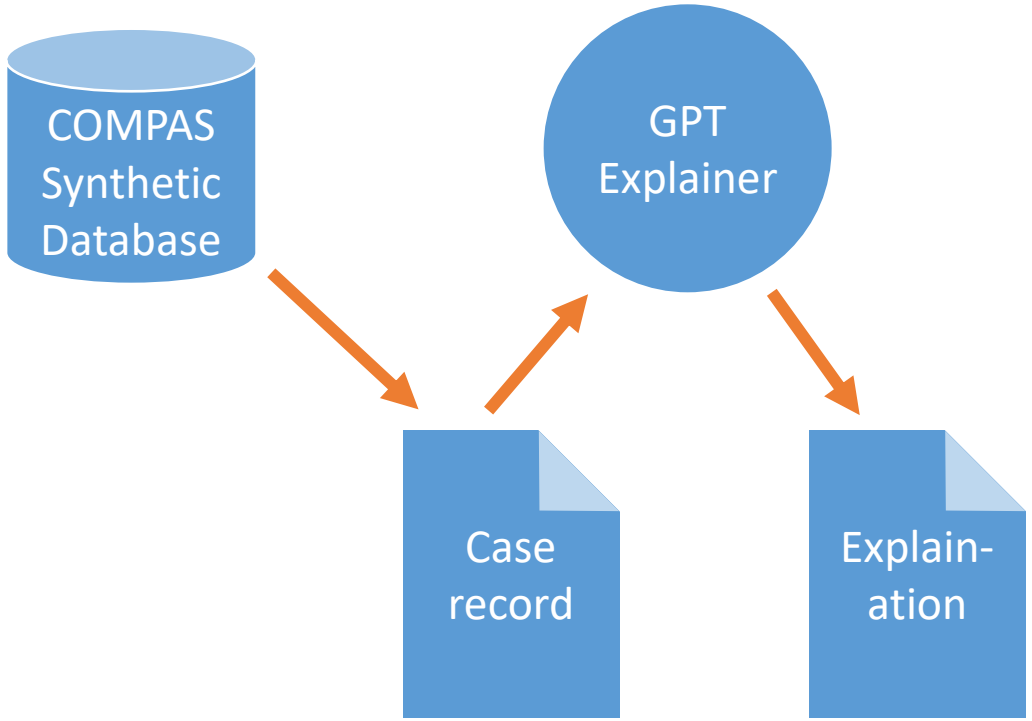
```
# Feature Levels
priors_count_level: high
c_charge_degree_level: high
juv_fel_count_level: low
family_criminality_level: medium
criminal_attitude_level: high
criminal_associates_level: high
financial_problems_level: low
substance_abuse_level: high
noncompliance_level: high
social_environment_level: medium
vocational_level: medium

# Feature Importance Levels
noncompliance_importance_level: high
substance_abuse_importance_level: high
criminal_associates_importance_level: medium
priors_count_importance_level: medium
financial_problems_importance_level: low
vocational_importance_level: low
social_environment_importance_level: low
c_charge_degree_importance_level: low
juv_fel_count_importance_level: low
family_criminality_importance_level: low
criminal_attitude_importance_level: low
```

LLM-based NLEs



LLM-based NLEs



Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for criminal attitude (8), a high risk score for noncompliance (6), and a high risk score for substance abuse (8). These were the three most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a medium risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for **criminal attitude** (**8**), a high risk score for noncompliance (6), and a high risk score for substance abuse (8). These were the three most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a **medium** risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

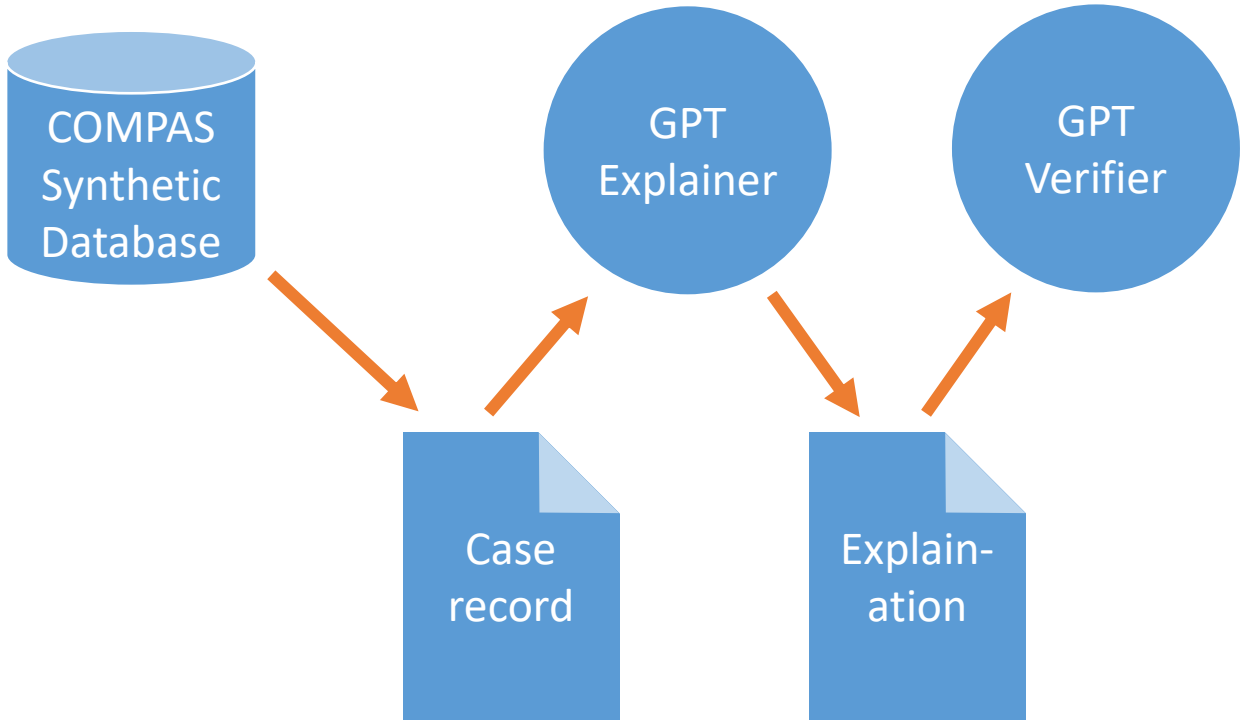
The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

LLM-based NLEs



Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for criminal attitude (8), a high risk score for noncompliance (6), and a high risk score for substance abuse (8). These were the three most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a medium risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for criminal attitude (8), a high risk score for noncompliance (6), and a high risk score for substance abuse (8). These were the three most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a medium risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

```
# Case Info
id: 142
name: richard
decile_score: 8
decile_score_level: high

# Features
priors_count: 9
c_charge_degree:
juv_fel_count:
family_criminality:
criminal_attitude:
criminal_associates: 7
financial_problems: 3
substance_abuse: 8
noncompliance: 6
social_environment: 5
vocational: 4
```

```
# Feature Levels
priors_count_level: high
c_charge_degree_level:
juv_fel_count_level:
family_criminality_level:
criminal_attitude_level:
criminal_associates_level: high
financial_problems_level: low
substance_abuse_level: high
noncompliance_level: high
social_environment_level: medium
vocational_level: medium

# Feature Importance
noncompliance_importance_level: high
substance_abuse_importance_level: high
criminal_associates_importance_level: medium
priors_count_importance_level: medium
financial_problems_importance_level: low
social_environment_importance_level: low
c_charge_degree_importance_level: low
juv_fel_count_importance_level: low
family_criminality_importance_level: low
criminal_attitude_importance_level: low
vocational_importance_level: low
```

```
# Case Info
id: 142
name: richard
decile_score: 8
decile_score_level: high

# Features
priors_count: 9
c_charge_degree: [redacted]
juv_fel_count: [redacted]
family_criminality: [redacted]
criminal_attitude: [redacted]
criminal_associates: 7
financial_problems: 3
substance_abuse: 8
noncompliance: 6
social_environment: 5
vocational: 4
```

```
# Feature Levels
priors_count_level: high
c_charge_degree_level: [redacted]
juv_fel_count_level: [redacted]
family_criminality_level: [redacted]
criminal_attitude_level: [redacted]
criminal_associates_level: high
financial_problems_level: low
substance_abuse_level: high
noncompliance_level: high
social_environment_level: medium
vocational_level: medium

# Feature Importance
noncompliance_importance_level: high
substance_abuse_importance_level: high
criminal_associates_importance_level: medium
priors_count_importance_level: medium
financial_problems_importance_level: low
social_environment_importance_level: low
c_charge_degree_importance_level: low
juv_fel_count_importance_level: low
family_criminality_importance_level: low
criminal_attitude_importance_level: low
vocational_importance_level: low
```

Errors

priors_count_level: medium -> high

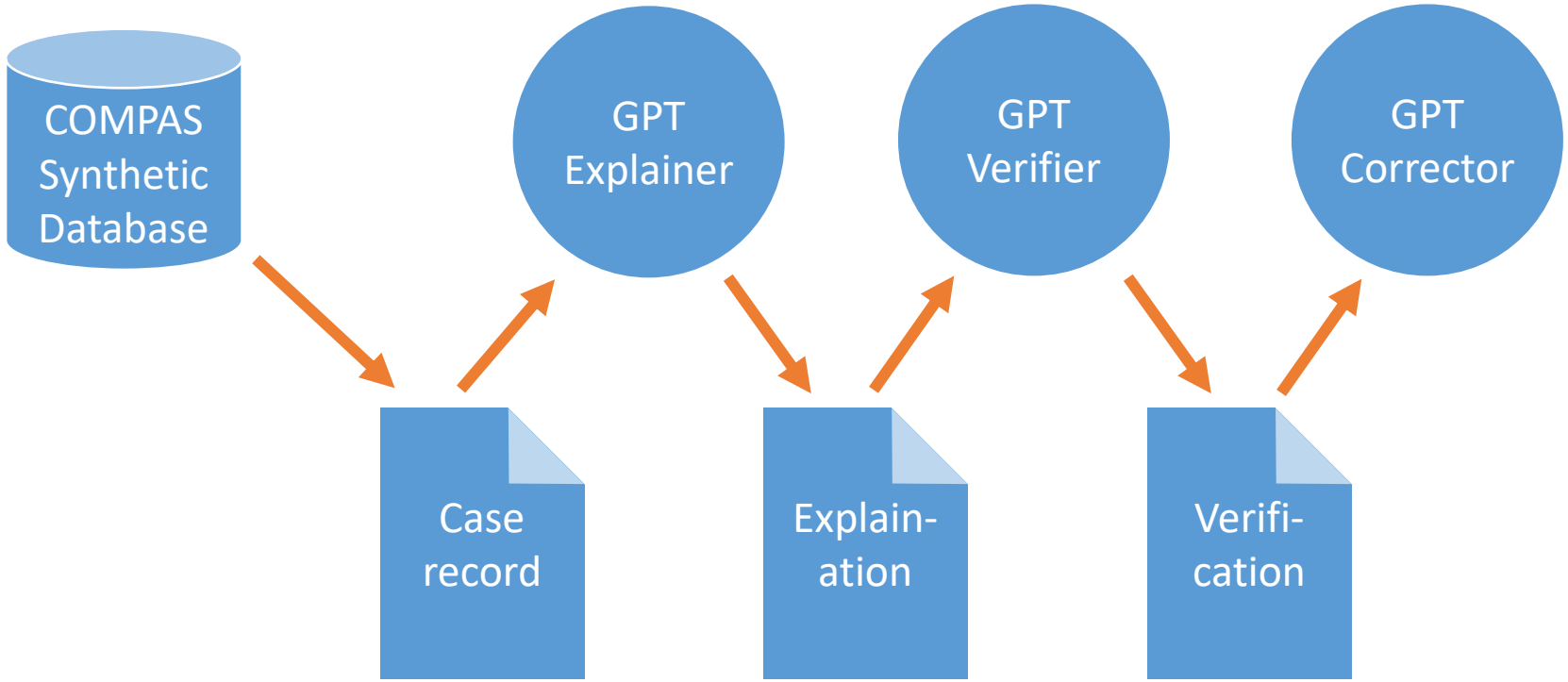
vocational_level: low -> medium

criminal_attitude_importance_level: high -> low

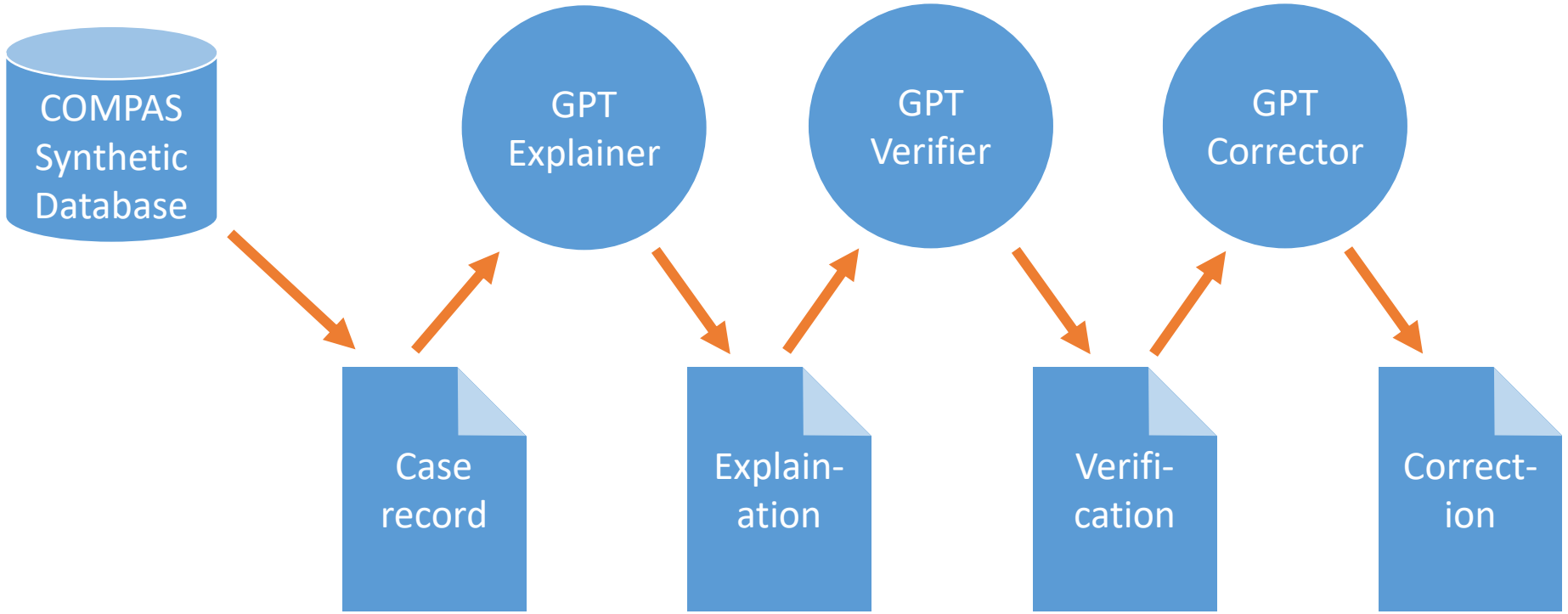
```
# Errors
priors_count_level: medium -> high
vocational_level: low -> medium
criminal_attitude_importance_level: high -> low
```

```
# Errors
priors_count_level: medium -> high
vocational_level: low -> medium
criminal_attitude_importance_level: high -> low
```

LLM-based NLEs



LLM-based NLEs



Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for **criminal attitude** (**8**), a high risk score for noncompliance (6), and a high risk score for substance abuse (8). These were the three most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a medium risk score for number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for noncompliance (6) and a high risk score for substance abuse (8). These were the two most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a high risk score for the number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Richard,

Thank you for sharing your case record for our review.

...

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Corrections
- Line 26: Fixed priors_count_level: medium -> high
- Line 27: Fixed vocational_level: low -> medium
- Line 32: Fixed criminal_attitude_importance_level: high -> low

Explanation
- The priors_count feature had a value of 9, which is high, so the feature level was corrected from medium to high.
- The vocational feature had a value of 4, which is medium, so the feature level was corrected from low to medium.
- The criminal_attitude feature had a high risk score of 8, but it was incorrectly listed as having high importance. It was corrected to low importance.

Richard,

Thank you for sharing your case record for our review.

...

Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Corrections
- Line 26: Fixed priors_count_level: medium -> high
- Line 27: Fixed vocational_level: low -> medium
- Line 32: Fixed criminal_attitude_importance_level: high -> low

Explanation
- The priors_count feature had a value of 9, which is high, so the feature level was corrected from medium to high.
- The vocational feature had a value of 4, which is medium, so the feature level was corrected from low to medium.
- The criminal_attitude feature had a high risk score of 8, but it was incorrectly listed as having high importance. It was corrected to low importance.

Richard,

Thank you for sharing your case record for our review.

...

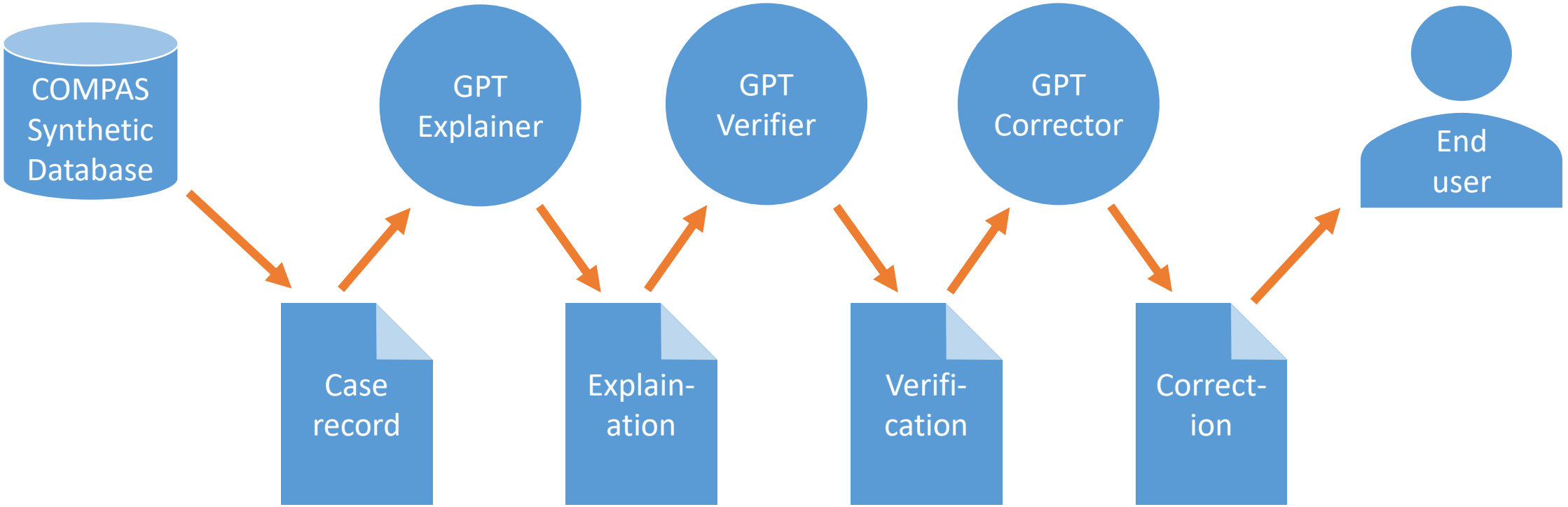
Please let us know if you have any questions or concerns.

Sincerely,
Explainable AI

Corrections
- Line 26: Fixed priors_count_level: medium -> high
- Line 27: Fixed vocational_level: low -> medium
- Line 32: Fixed criminal_attitude_importance_level: high -> low

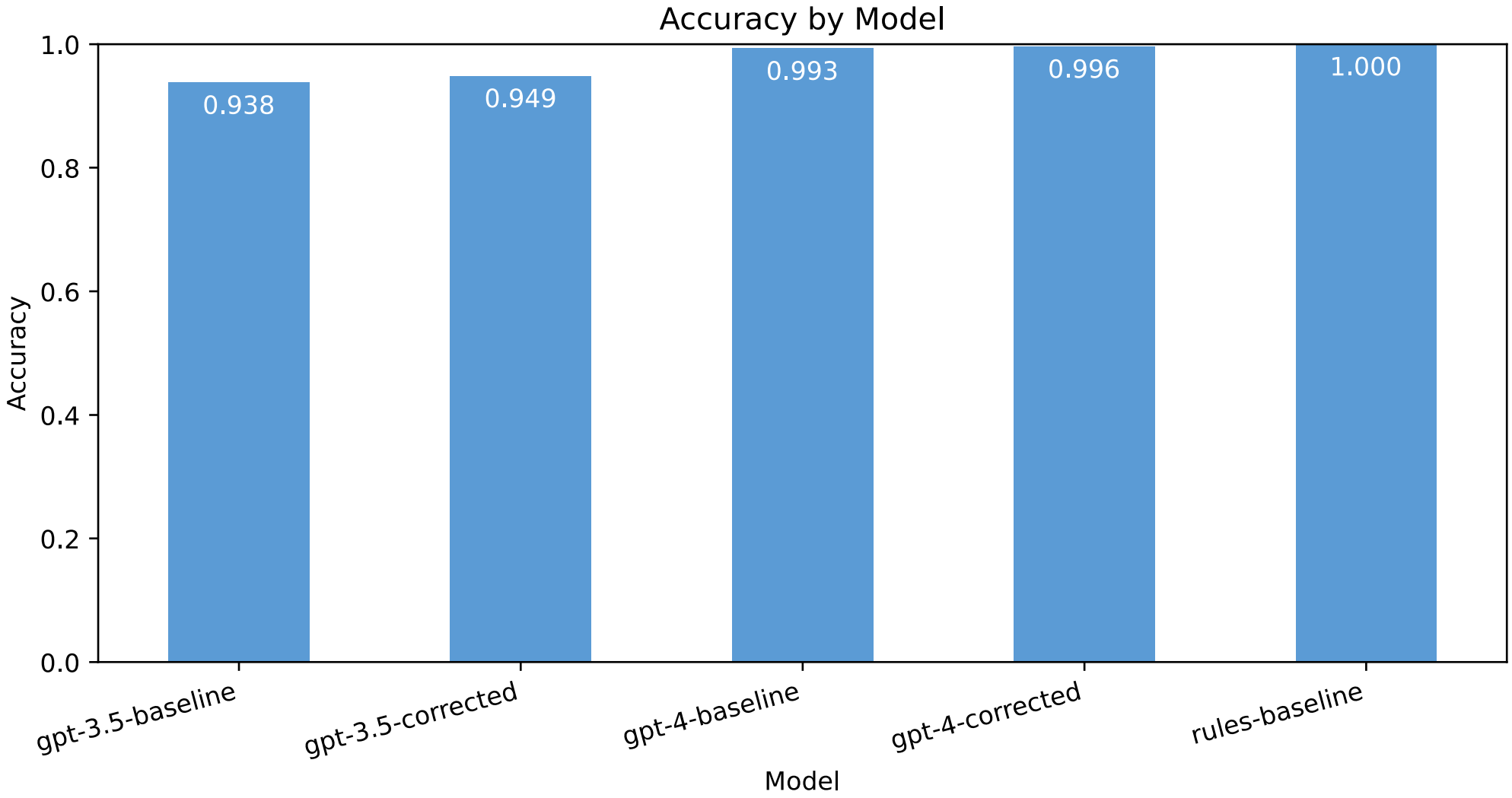
Explanation
- The priors_count feature had a value of 9, which is high, so the feature level was corrected from medium to high.
- The vocational feature had a value of 4, which is medium, so the feature level was corrected from low to medium.
- The criminal_attitude feature had a high risk score of 8, but it was incorrectly listed as having high importance. It was corrected to low importance.

LLM-based NLEs

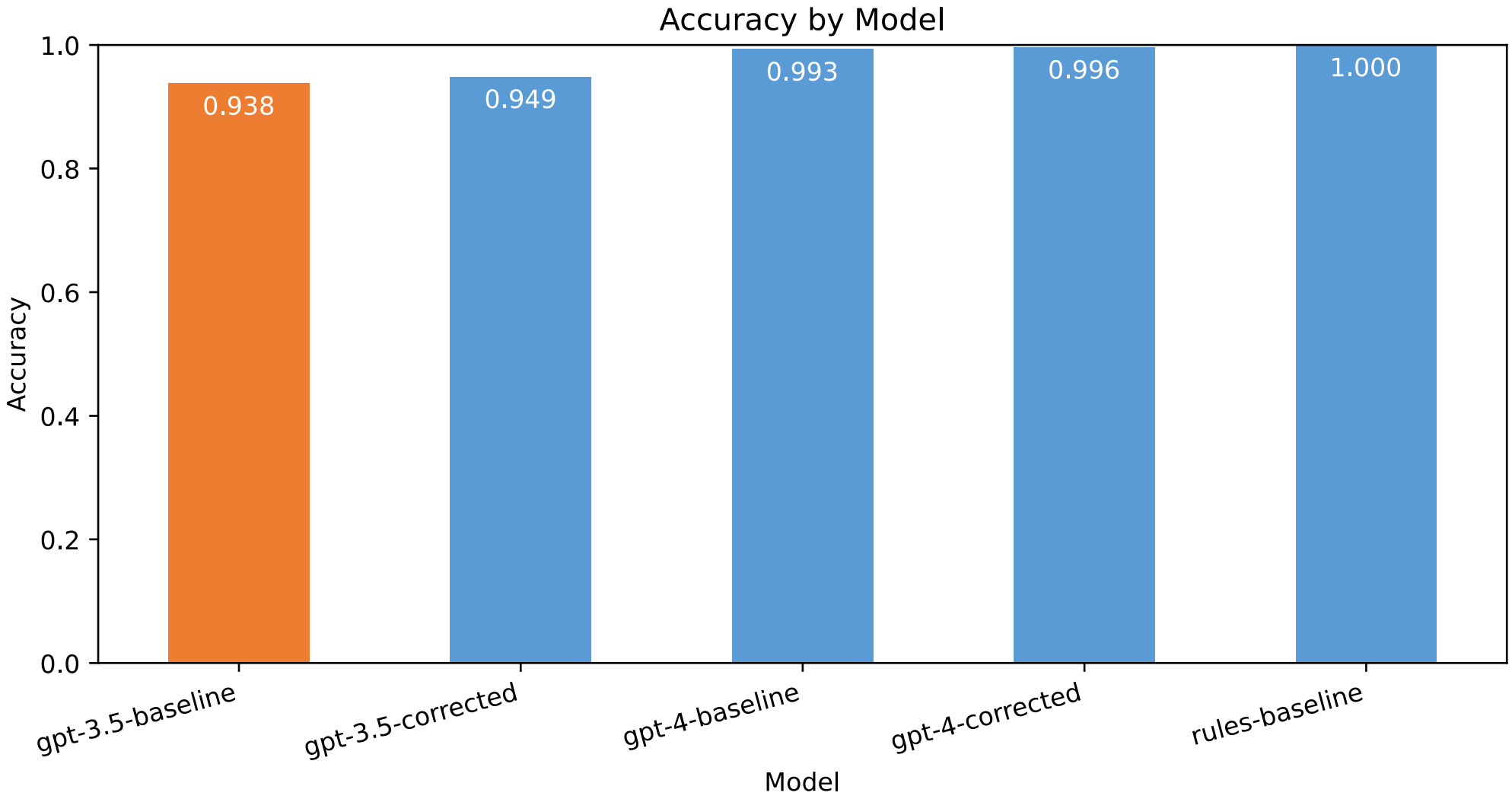


Results

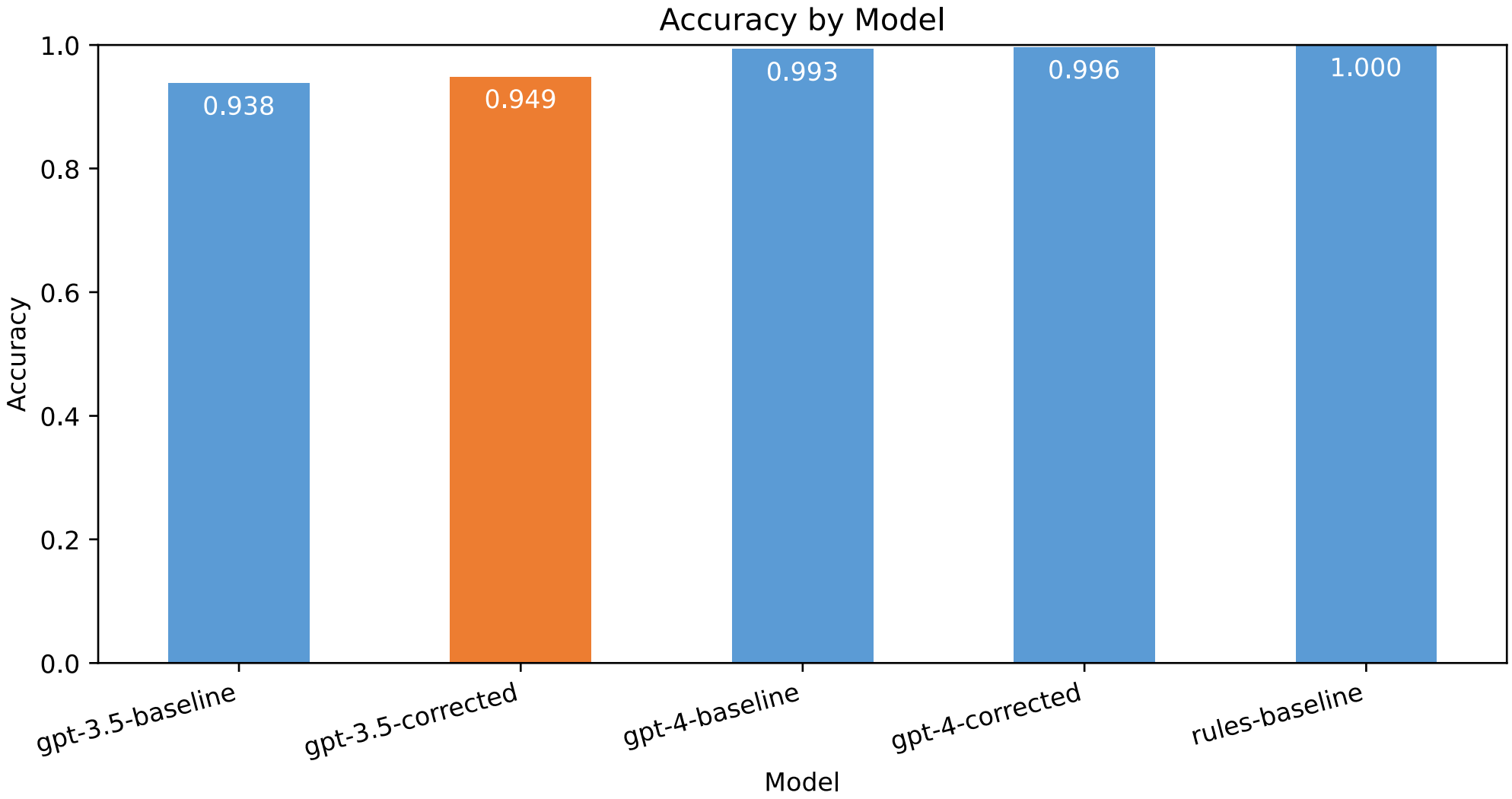
Factual Accuracy



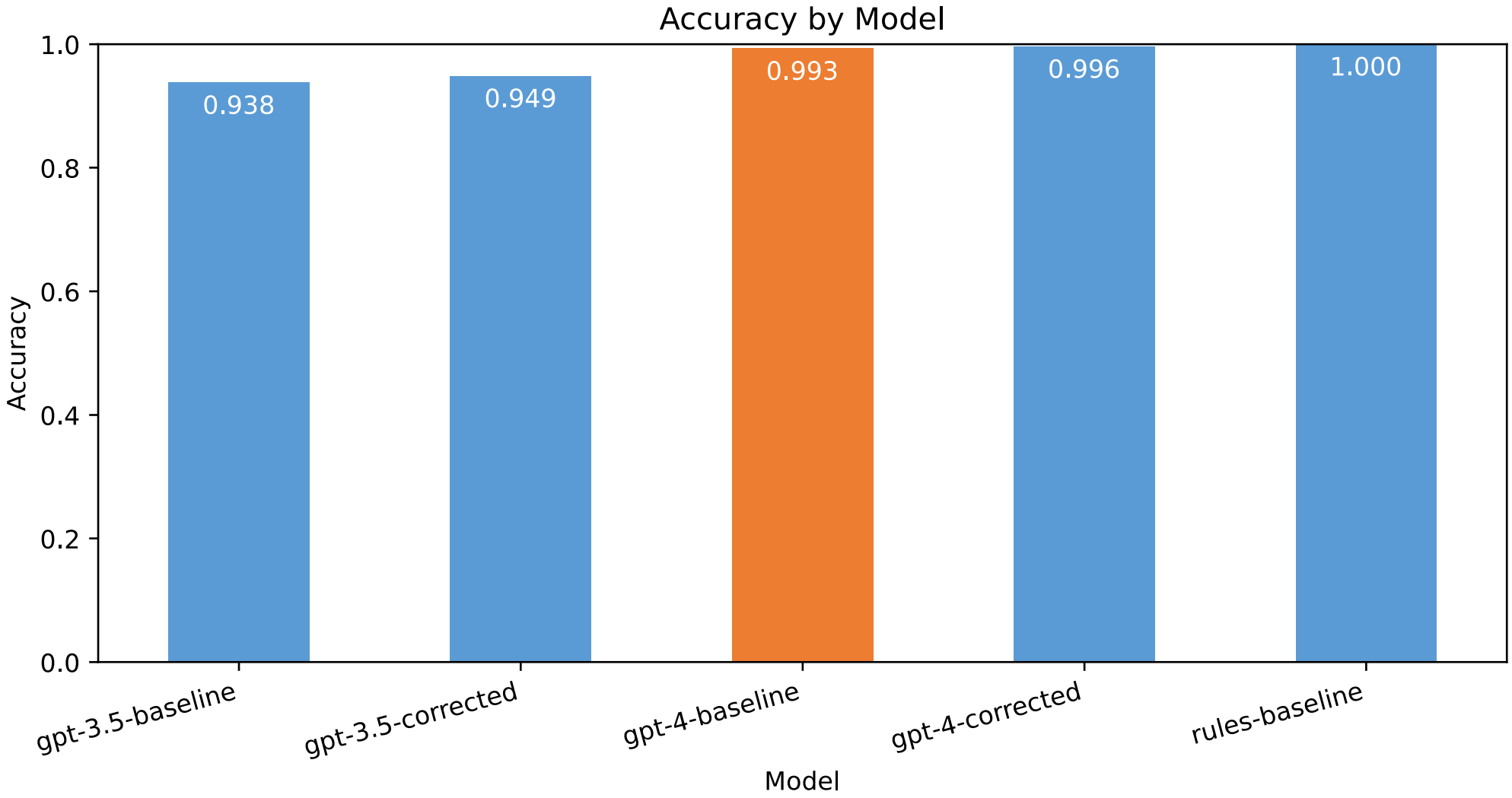
Factual Accuracy



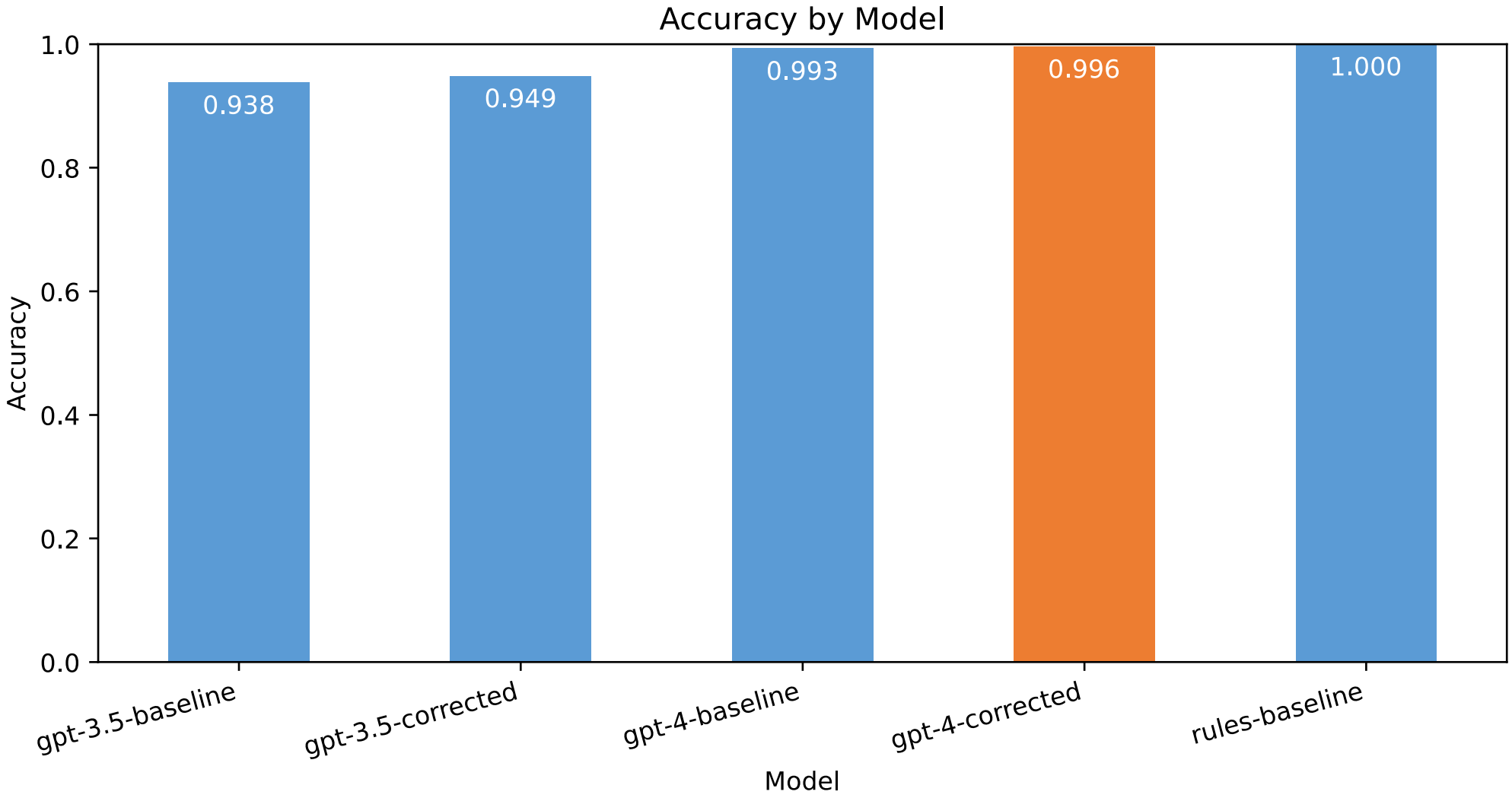
Factual Accuracy



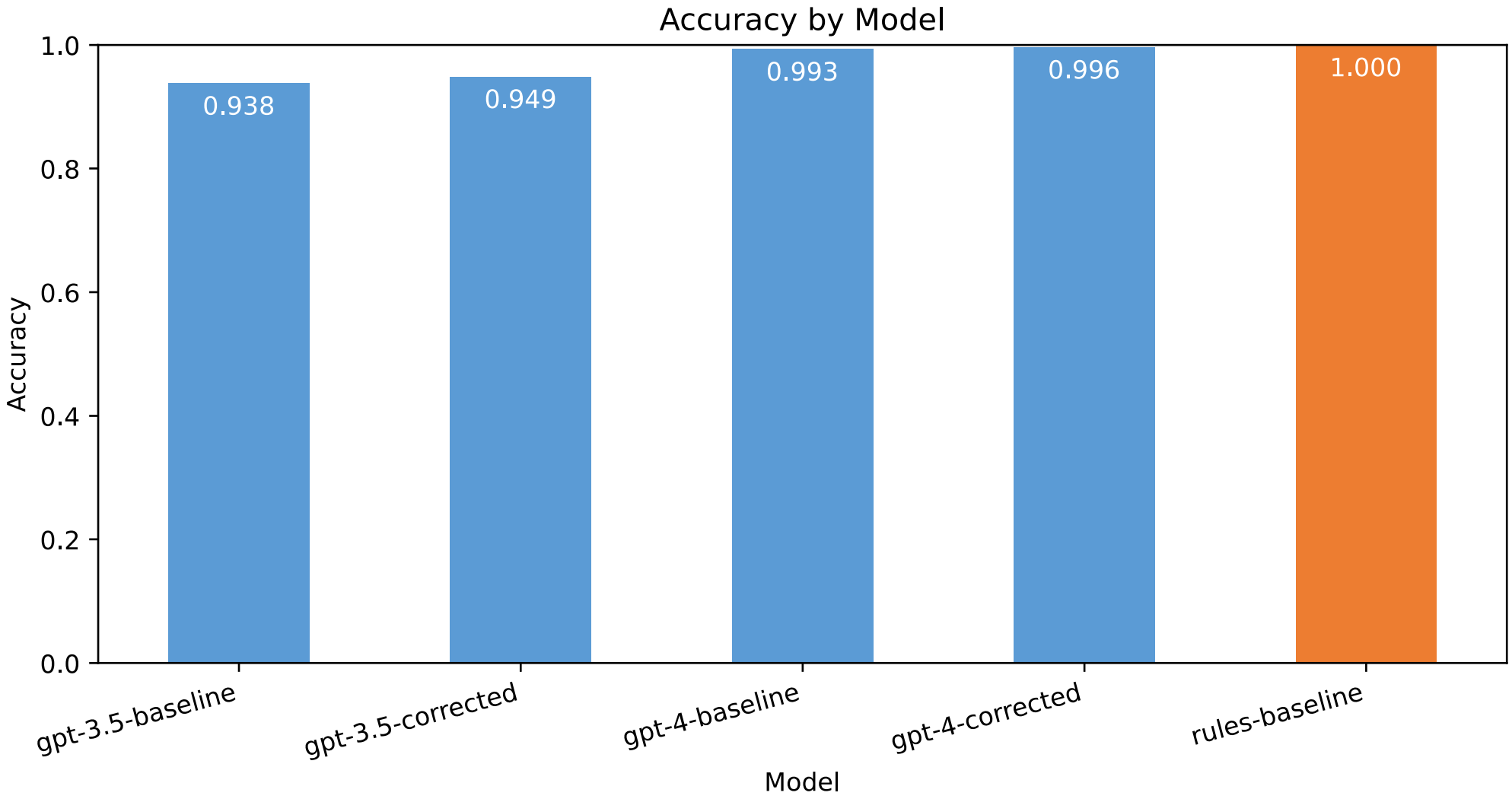
Factual Accuracy



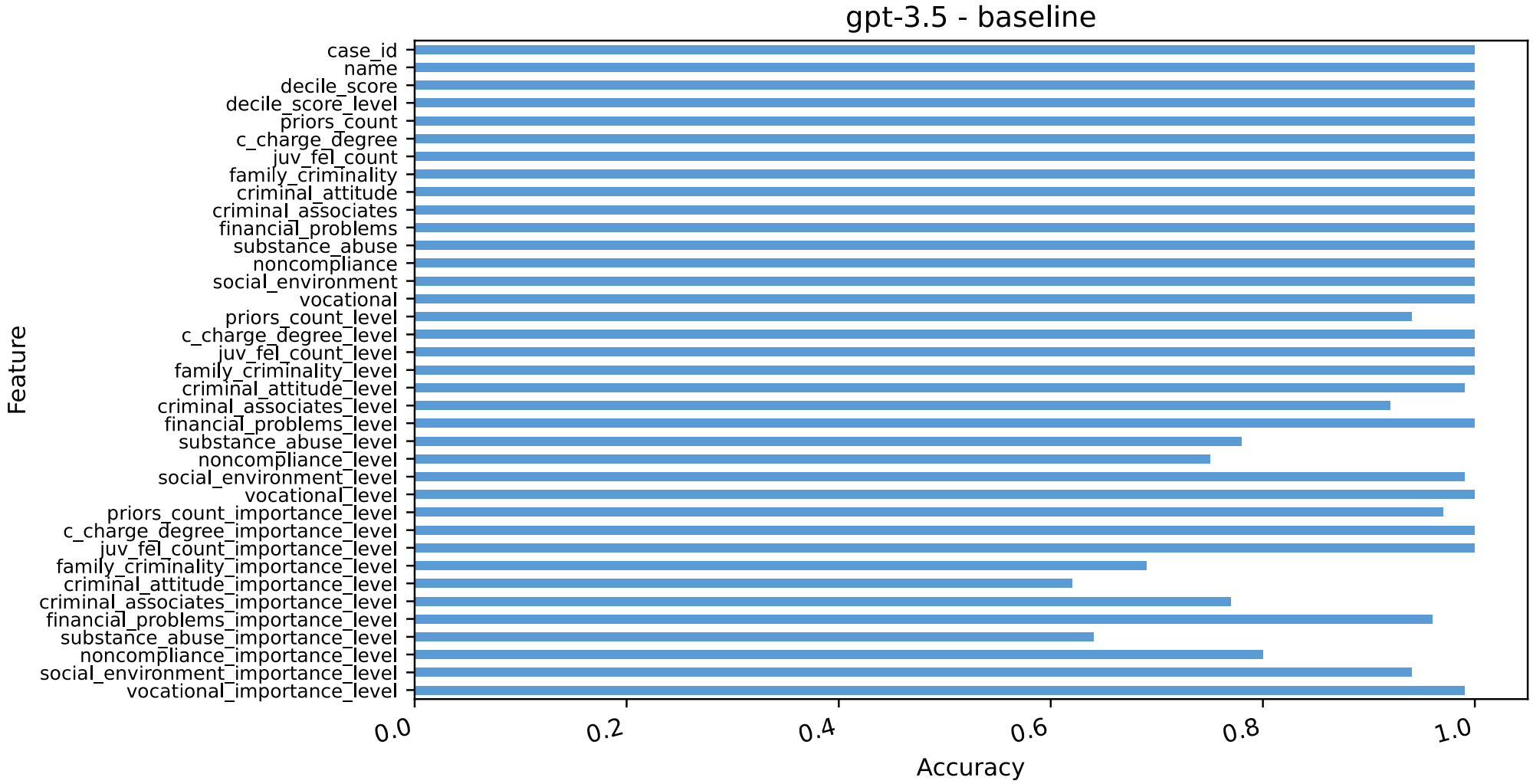
Factual Accuracy



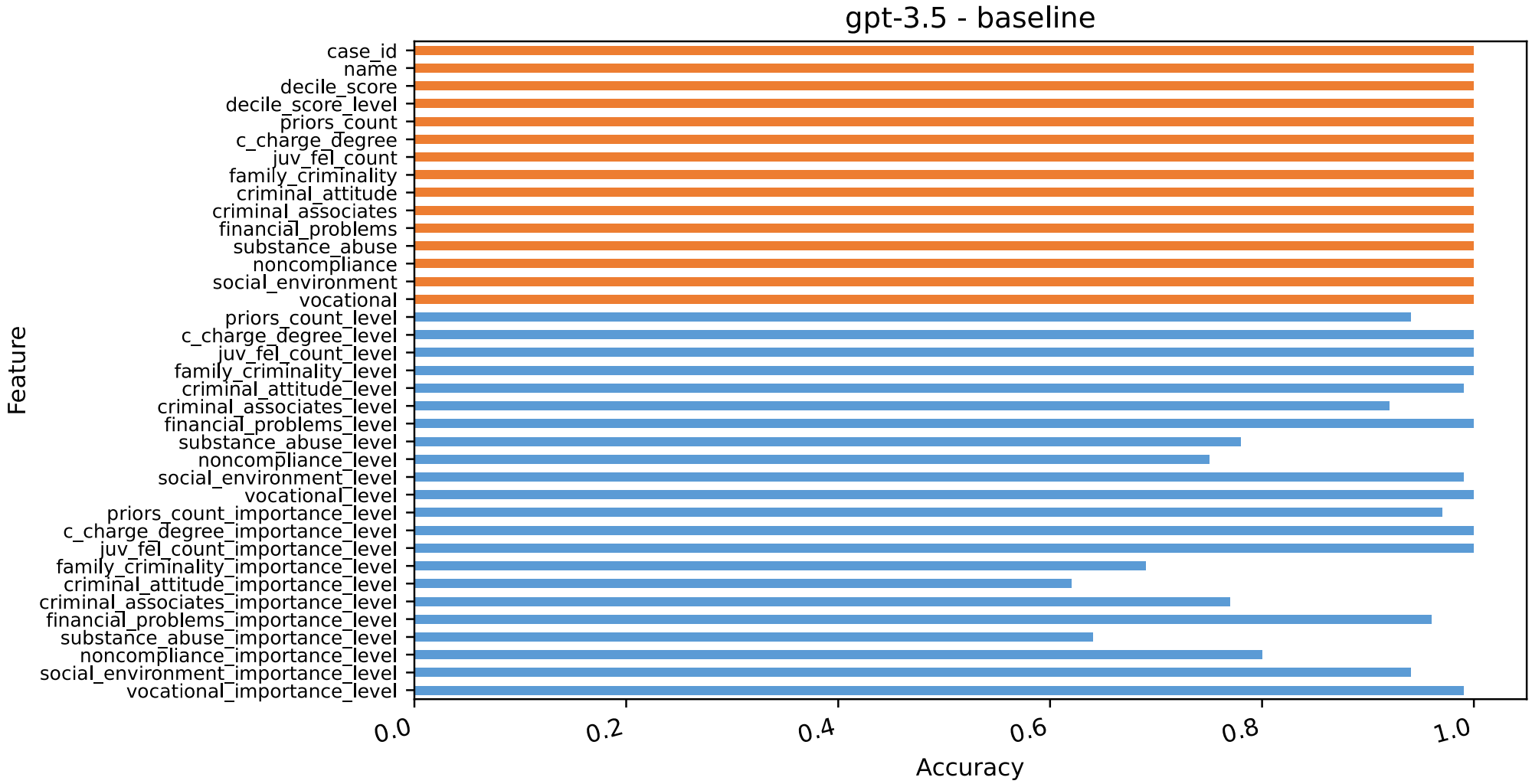
Factual Accuracy



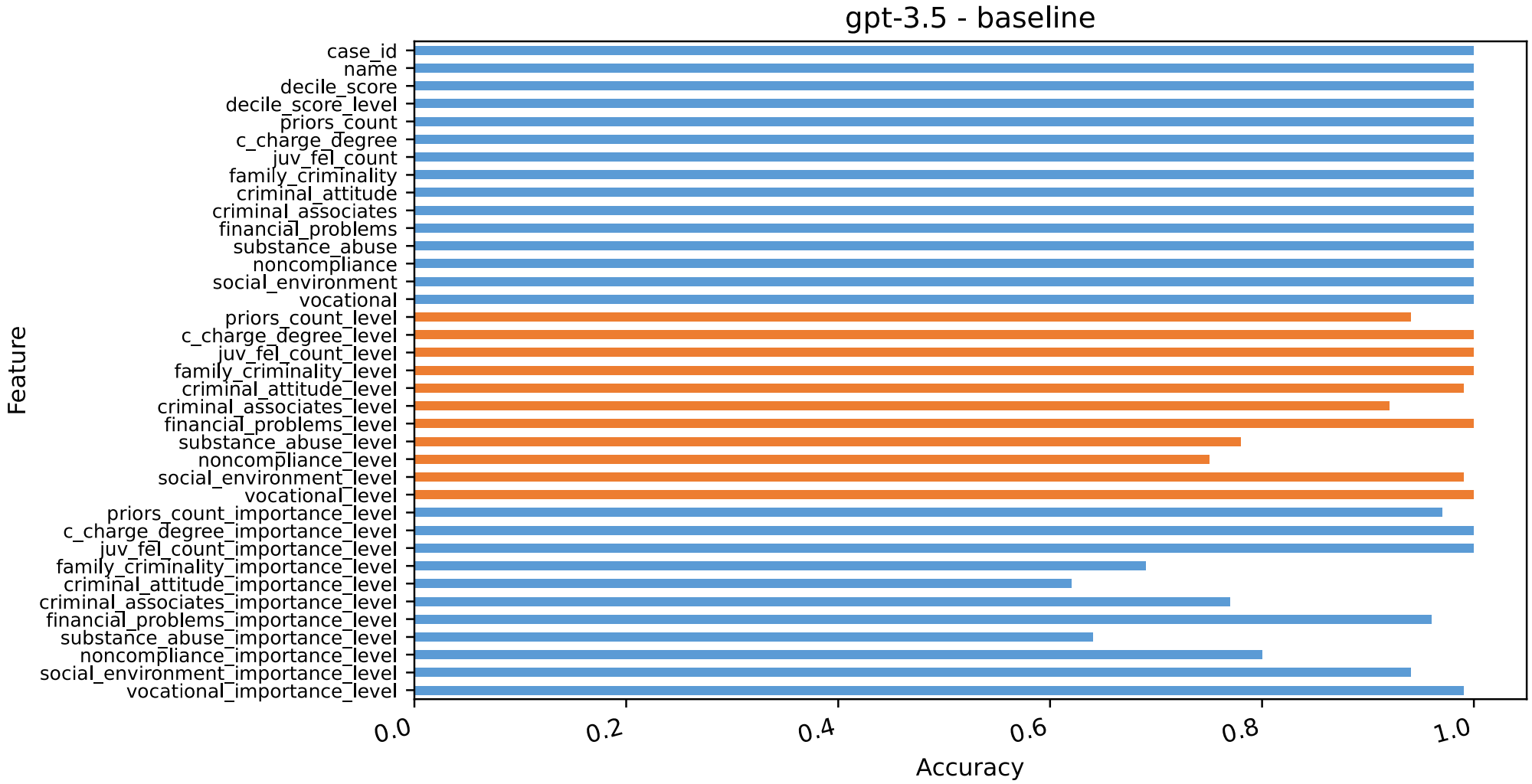
Detailed Analysis (Worst Case)



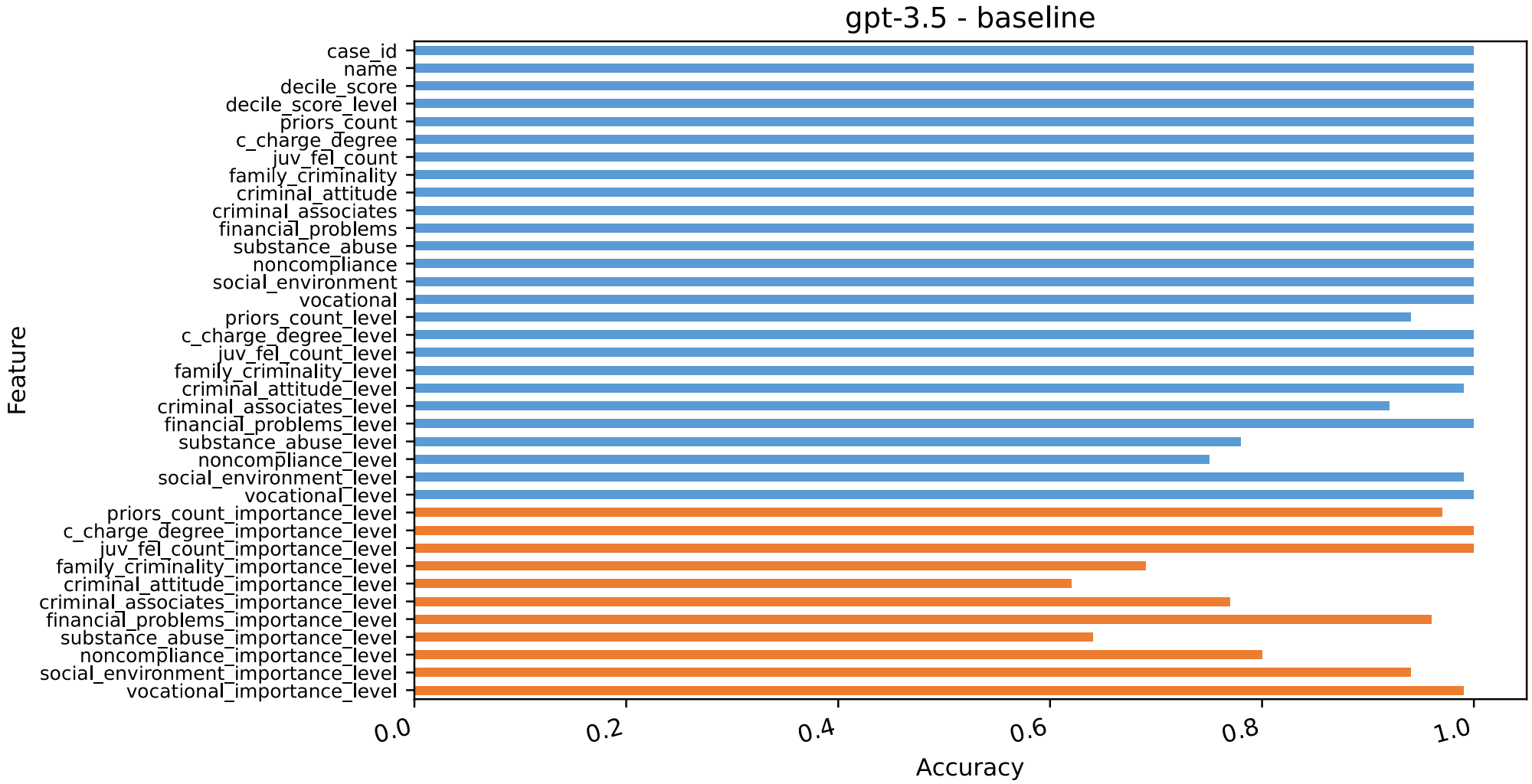
Detailed Analysis (Worst Case)



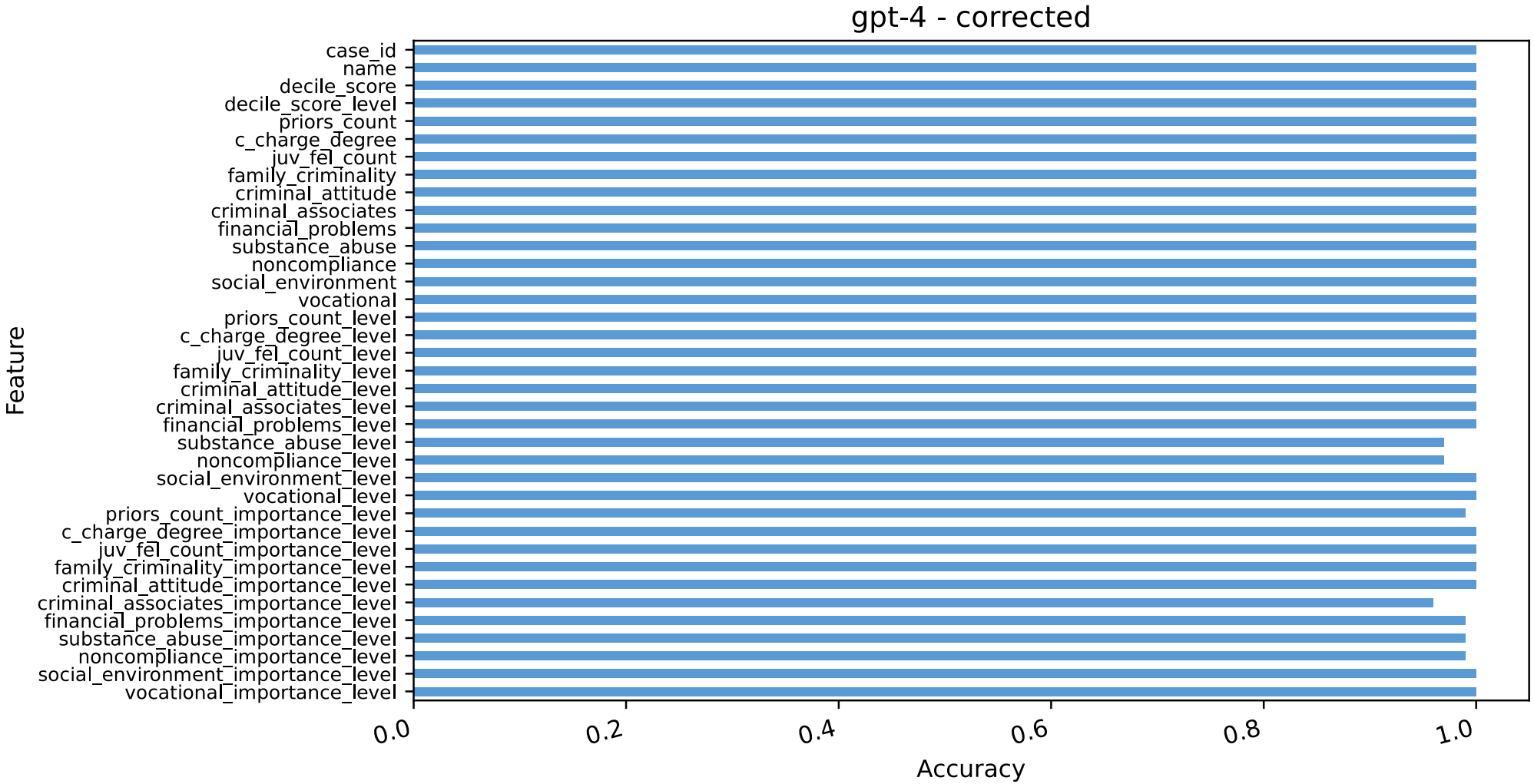
Detailed Analysis (Worst Case)



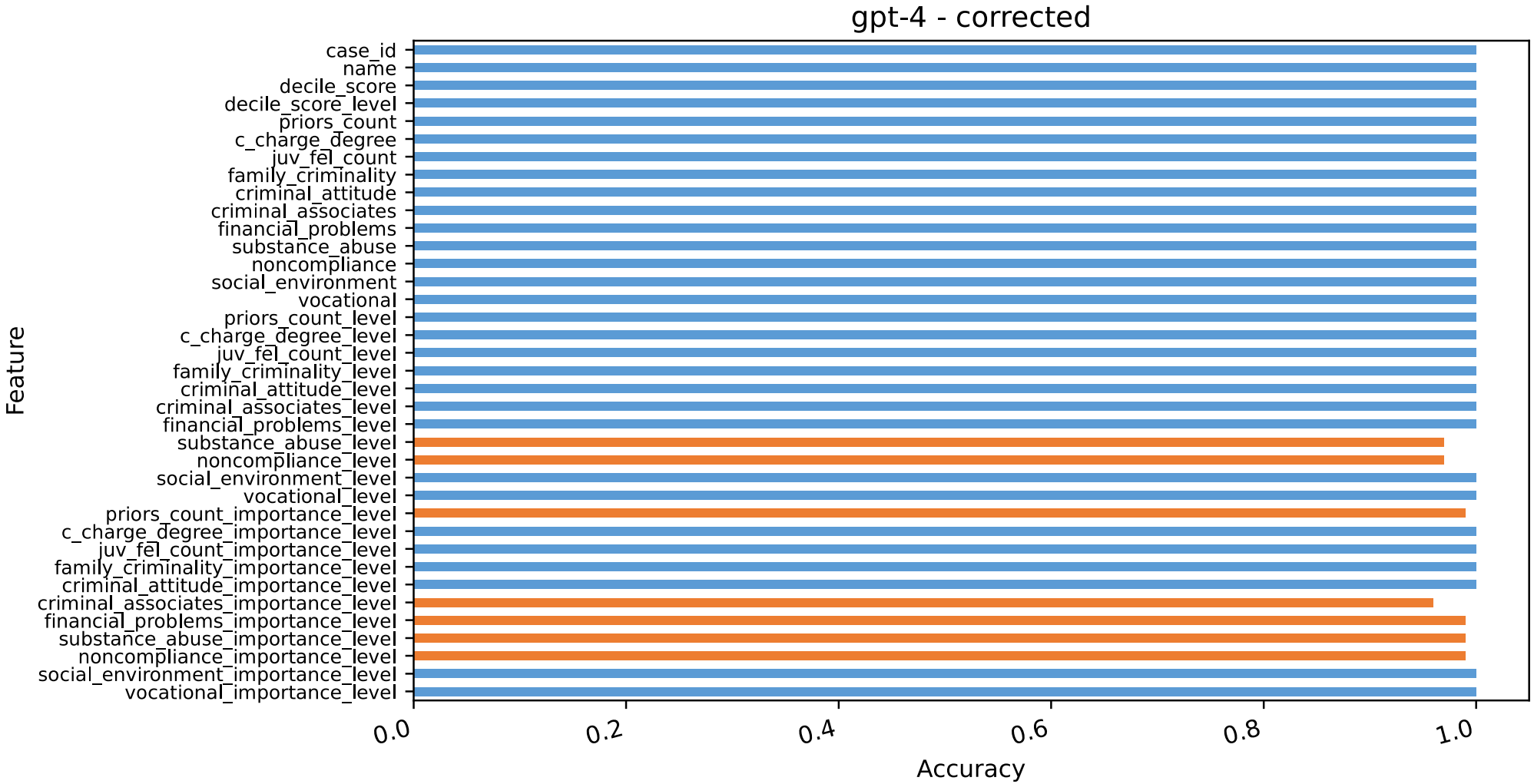
Detailed Analysis (Worst Case)



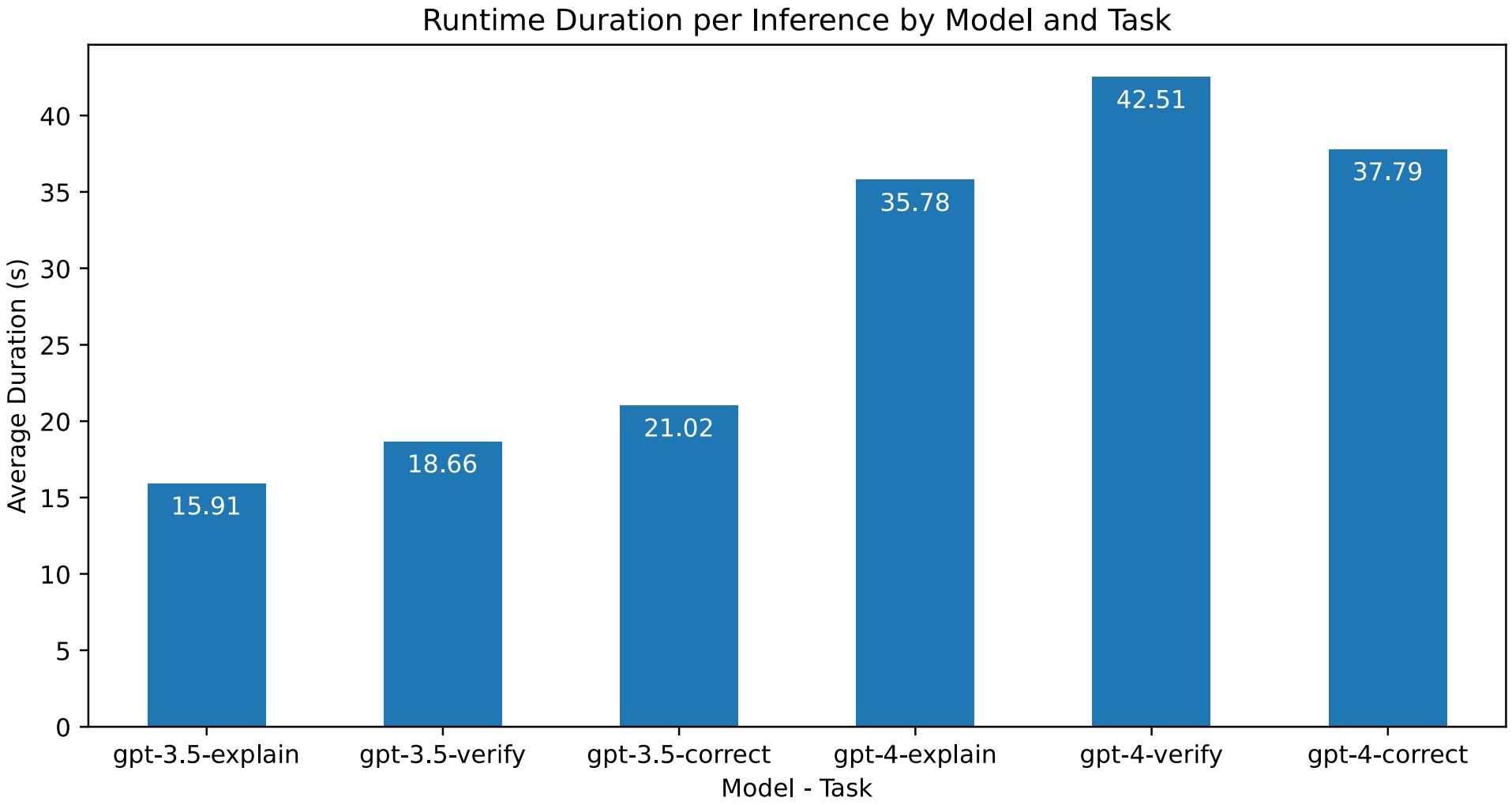
Detailed Analysis (Best Case)



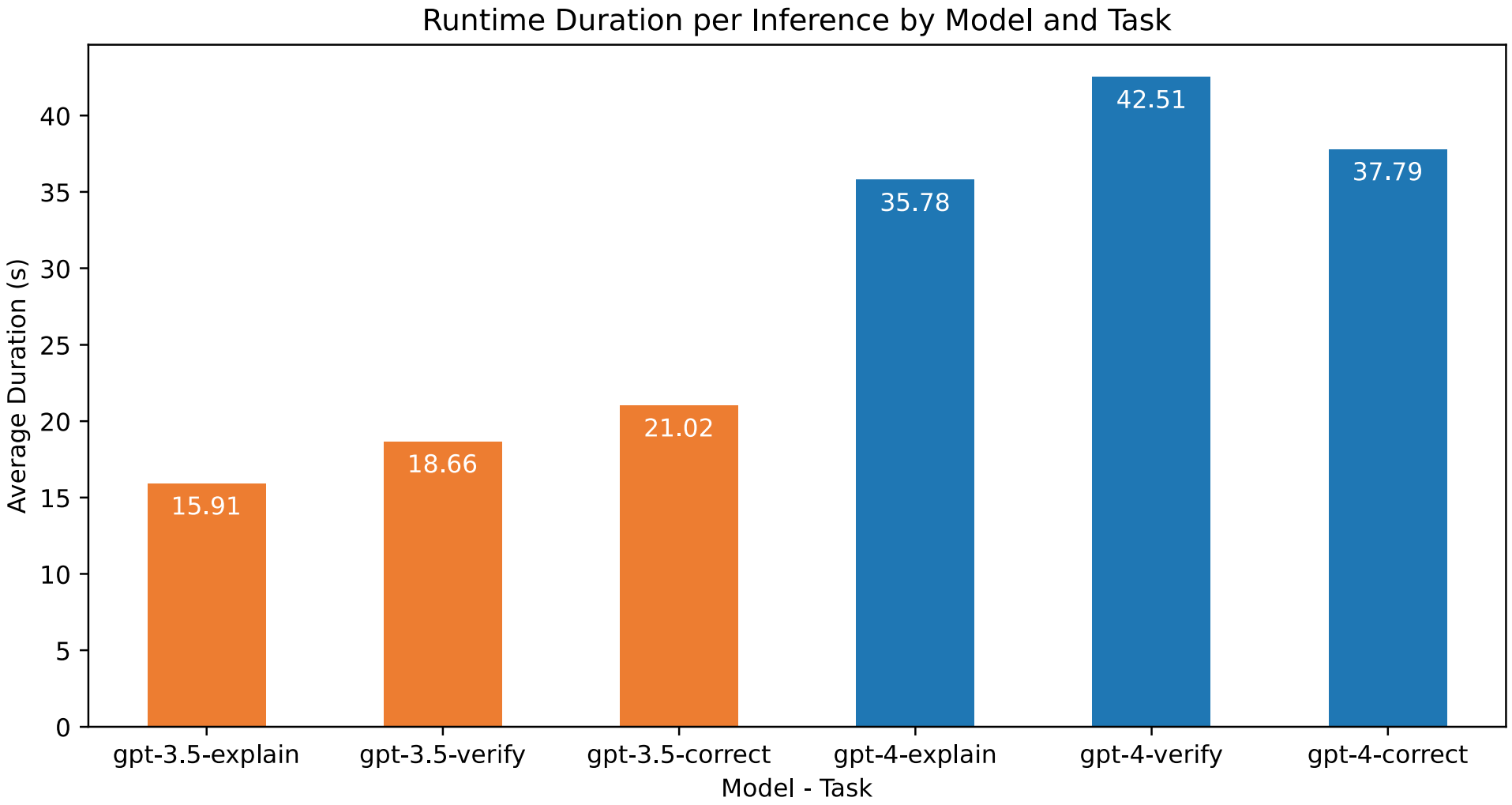
Detailed Analysis (Best Case)



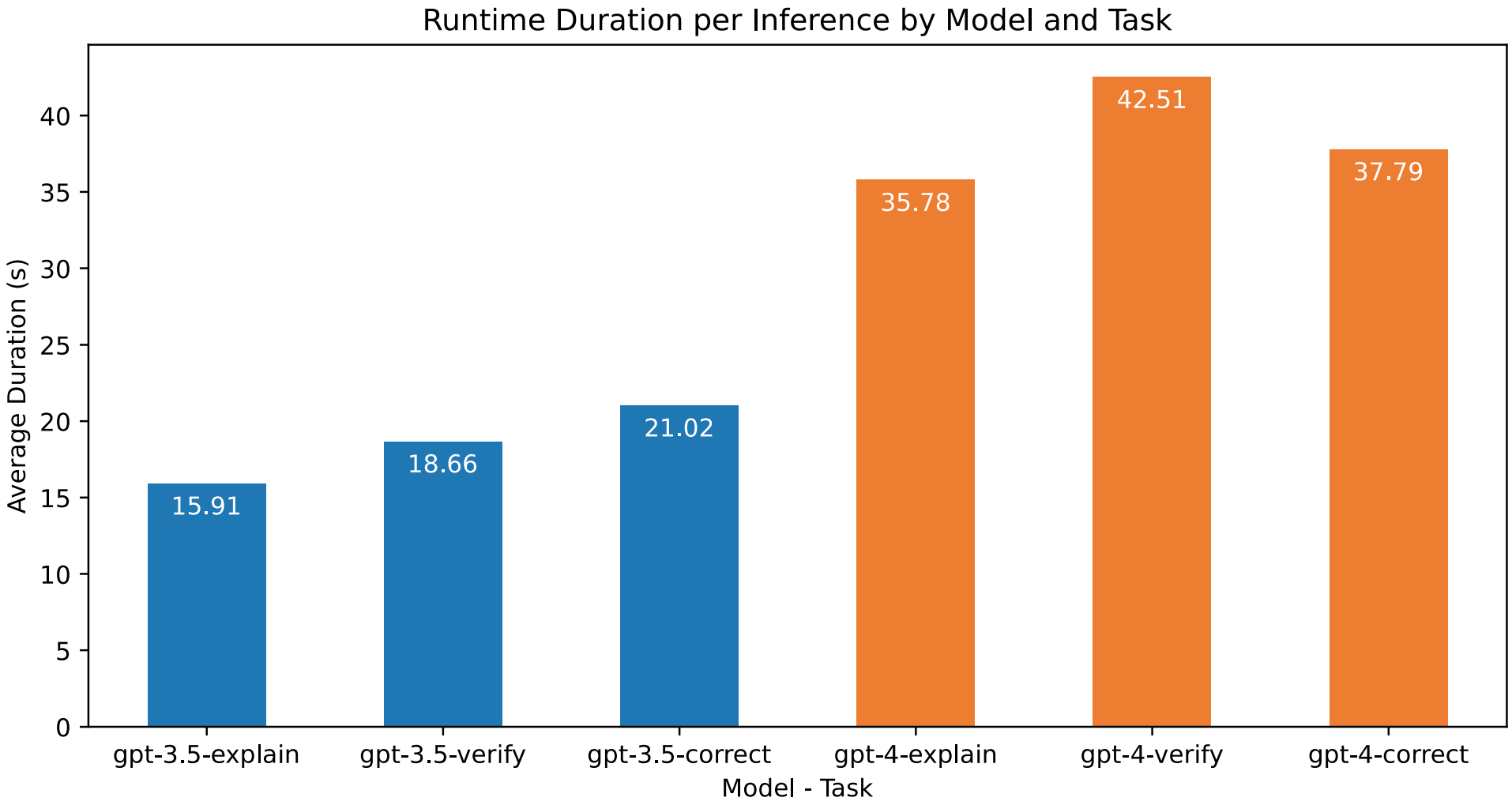
Runtime Analysis



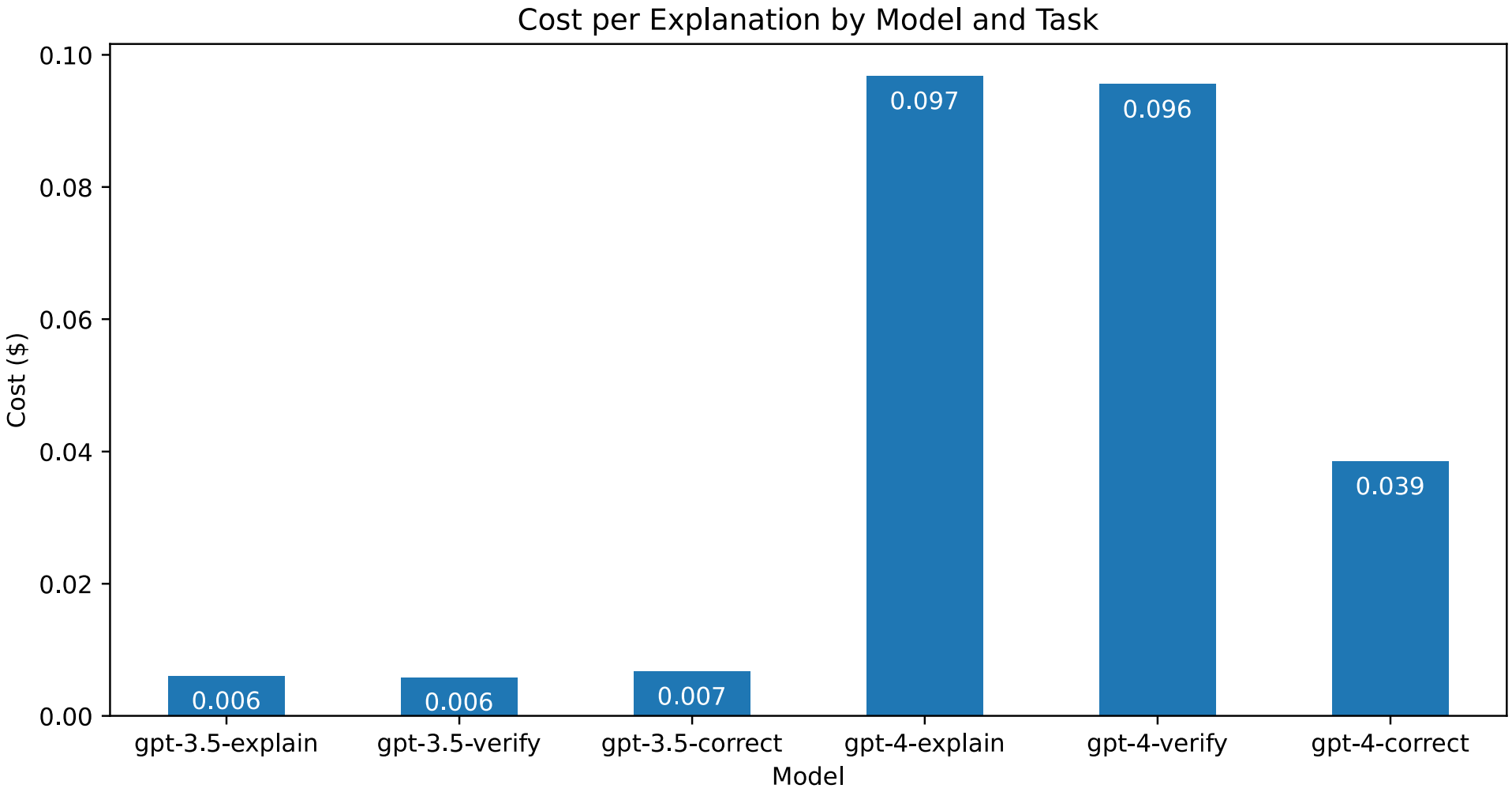
Runtime Analysis



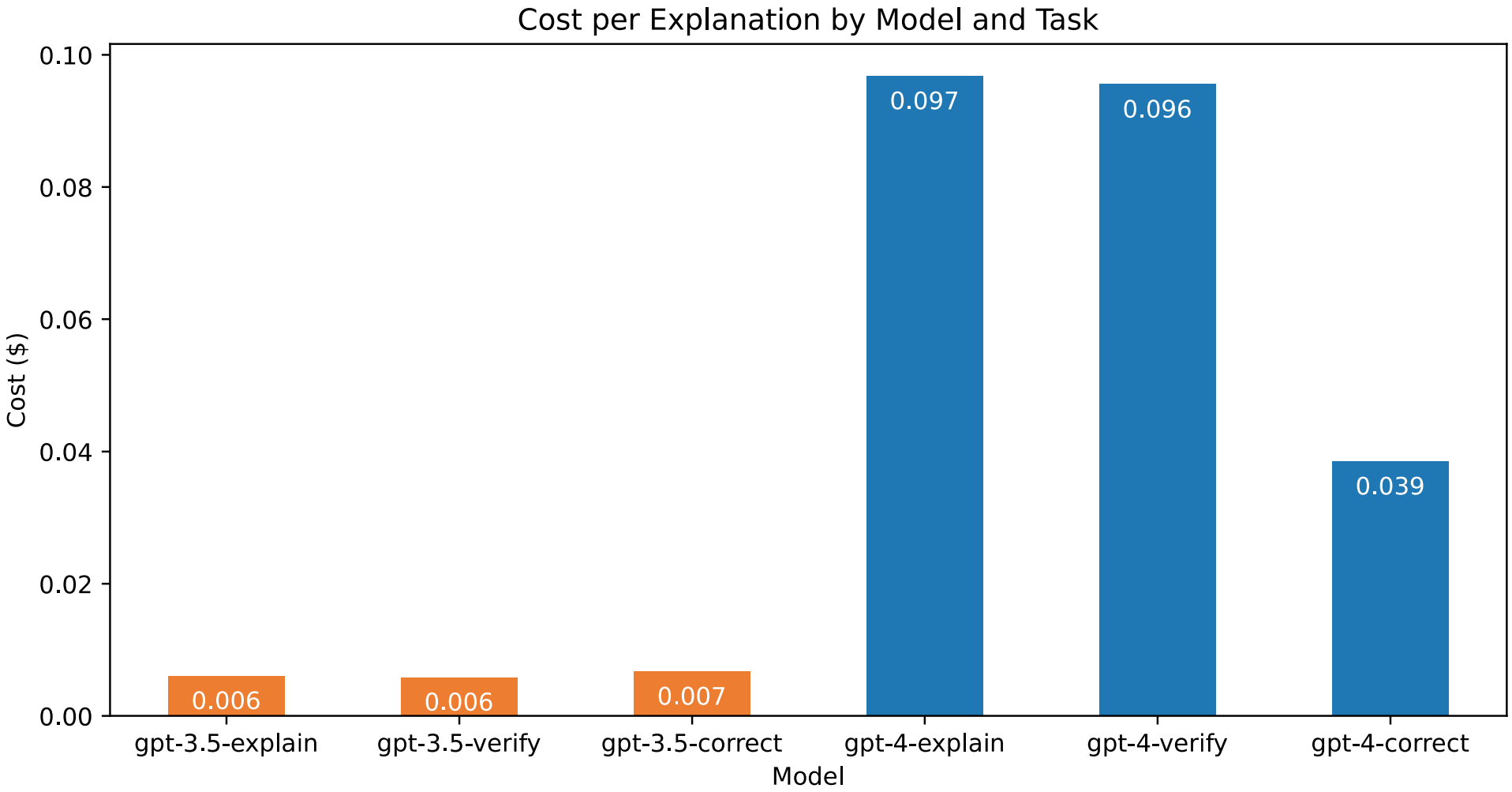
Runtime Analysis



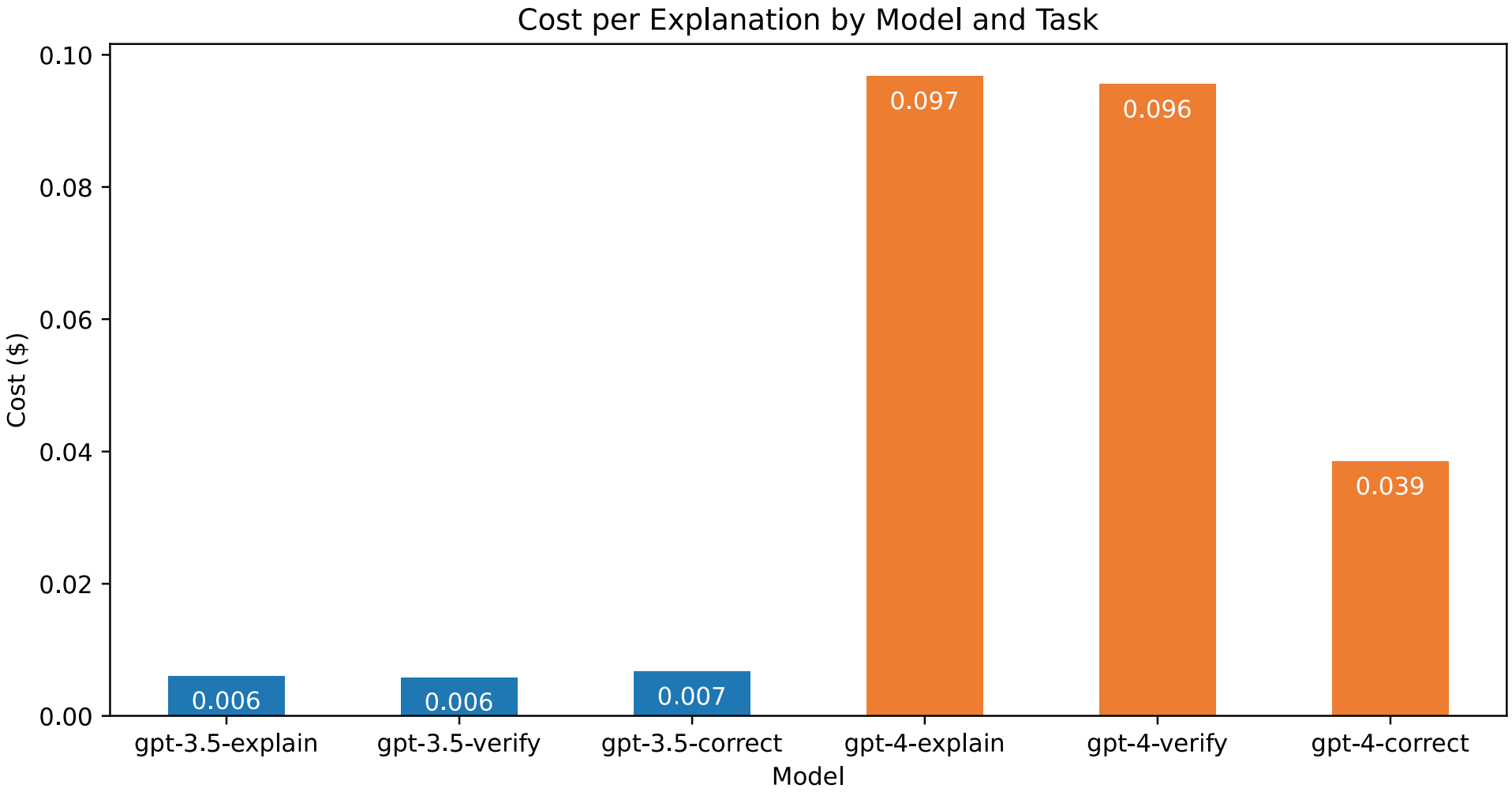
Cost Analysis



Cost Analysis



Cost Analysis



Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for noncompliance (6) and a high risk score for substance abuse (8). These were the two most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a high risk score for the number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm’s scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm’s score.

Demographic data including age, sex, race, etc., were not included in the algorithm’s decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,

Explainable AI

Richard,

Thank you for sharing your case record for our review.

Regarding your case record #142, you had a COMPAS risk score of 8 on a scale of 1-10. This means you were scored as a high risk for reoffending.

You received this score primarily because you had a high risk score for noncompliance (6) and a high risk score for substance abuse (8). These were the two most important factors in receiving a high overall risk score.

You also had a high risk score for criminal associates (7) and a high risk score for the number of prior offenses (9). However, these factors only contributed medium importance to the COMPAS algorithm's scoring of your overall risk.

The remaining risk factors including financial problems (3), social environment risk (5), and vocational risk factors (4), etc. were of low importance to the COMPAS algorithm's score.

Demographic data including age, sex, race, etc., were not included in the algorithm's decision because they are protected by law.

Please let us know if you have any questions or concerns.

Sincerely,

Explainable AI

Discussion

Limitations

Limitations

Used a single data source

Limitations

Used a single data source

Limited context length

Limitations

Used a single data source

Limited context length

Required hand-picked examples

Limitations

Used a single data source

Limited context length

Required hand-picked examples

Needed text labels (low, med., high)

Limitations

Used a single data source

Limited context length

Required hand-picked examples

Needed text labels (low, med., high)

Had to separate verifier and corrector

Future Research

Future Research

Use XAI benchmark dataset

Future Research

Use XAI benchmark dataset

Improve self-corrections

Future Research

Use XAI benchmark dataset

Improve self-corrections

Use raw feature scores

Future Research

Use XAI benchmark dataset

Improve self-corrections

Use raw feature scores

Condense data format

Future Research

Use XAI benchmark dataset

Improve self-corrections

Use raw feature scores

Condense data format

Survey for quality

Implications

Implications

Demonstrates feasibility

Implications

Demonstrates feasibility

Increases trust in ML

Implications

Demonstrates feasibility

Increases trust in ML

Frees up XAI experts

Implications

Demonstrates feasibility

Increases trust in ML

Frees up XAI experts

Allows for interactive Q&A

Conclusion

Summary

Summary

LLMs can generate NLEs for XAI

Summary

LLMs can generate NLEs for XAI

GPT-4 with corrections was best

Summary

LLMs can generate NLEs for XAI

GPT-4 with corrections was best

GPT-3.5 better for runtime / cost

Summary

LLMs can generate NLEs for XAI

GPT-4 with corrections was best

GPT-3.5 better for runtime / cost

Limitations need to be overcome

Summary

LLMs can generate NLEs for XAI

GPT-4 with corrections was best

GPT-3.5 better for runtime / cost

Limitations need to be overcome

Further research is warranted

Summary

LLMs can generate NLEs for XAI

GPT-4 with corrections was best

GPT-3.5 better for runtime / cost

Limitations need to be overcome

Further research is warranted

Could lead to more trust in AI/ML

Thank You!

Learn more at:
https://matthewrenze.com