

Building Effective AI-Powered Data Systems



Shreya Shankar

UC Berkeley EECS

November 2025

docetl.org



A system for LLM-powered data processing

Email: shreyashankar@berkeley.edu

Website: sh-reya.com

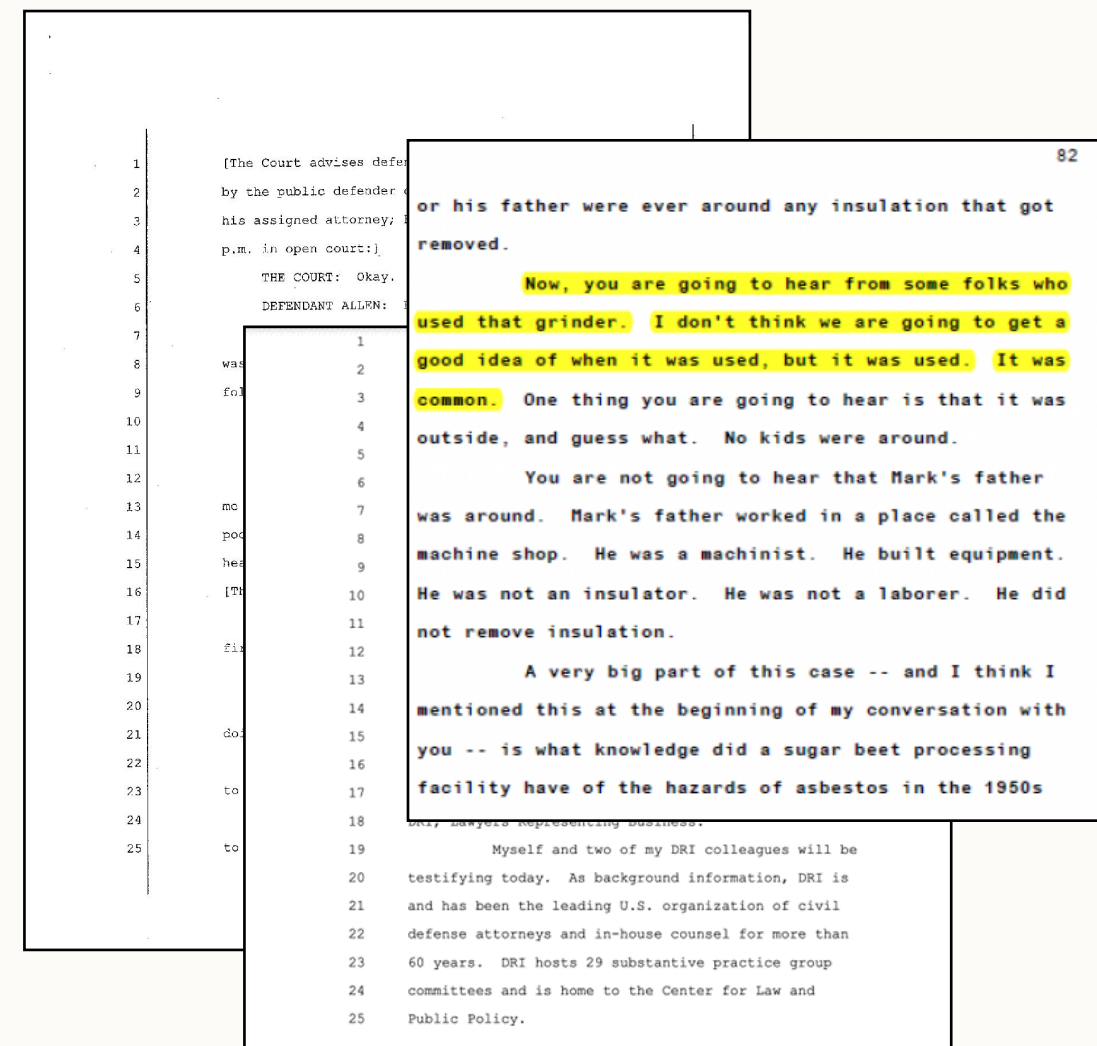


Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.

Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.



Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.

1

[The Court advises defe

2

by the public defender

3

his assigned attorney;

4

p.m. in open court].

5

THE COURT: Okay.

6

DEPENDANT ALLAN:

7

1

8

2

9

3

10

4

11

5

12

6

13

7

14

8

15

9

16

10

17

11

18

12

19

13

20

14

21

15

22

16

23

17

24

18

25

19

82

or his father were ever around any insulation that got removed.

Now, you are going to hear from some folks who used that grinder. I don't think we are going to get a good idea of when it was used, but it was used. It was common. One thing you are going to hear is that it was outside, and guess what. No kids were around.

You are not going to hear that Mark's father was around. Mark's father worked in a place called the machine shop. He was a machinist. He built equipment. He was not an insulator. He was not a laborer. He did not remove insulation.

A very big part of this case -- and I think I mentioned this at the beginning of my conversation with you -- is what knowledge did a sugar beet processing facility have of the hazards of asbestos in the 1950s

Myself and two of my DRI colleagues will be testifying today. As background information, DRI is and has been the leading U.S. organization of civil defense attorneys and in-house counsel for more than 60 years. DRI hosts 29 substantive practice group committees and is home to the Center for Law and Public Policy.



THERAPY PROGRESS NOTE			
DATE OF SESSION	8/9/2023	DURATION OF SESSION	45 minutes
LOCATION OF SESSION	VIRTUAL		
PATIENT DEMOGRAPHIC INFORMATION			
CLIENT NAME	JOHN DOE	DOB: 10-10-1970	
CURRENT PSYCHIATRIST	JON JOHNS, MD	NPI: 1234567890	
PRESENTING PROBLEM	Client reports feeling overwhelmed over the past two weeks. Client discussed recent interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.		
SESSION CONTENT	Client discussed recent interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.		
INTERVENTIONS	Utilized cognitive restructuring techniques. Became tearful appreciation for the support.		
CLIENT RESPONSE	The client is motivated and patterns of behavior and th will develop effective coping.		
THERAPIST'S OBSERVATIONS	The client is motivated and patterns of behavior and th will develop effective coping.		
MENTAL STATUS			
Client presents in casual attire, appearing states age. Normal cooperative, though somewhat reserved. Speech is of normal mood as 'okay' and affect is congruent. restricted but euthym content is without any overt delusions, hallucinations, suicidal oriented to person, place, and time. Client demonstrates good partially limited, and judgment appears adequate.			
RISK ASSESS			
No indications of suicidal or homicidal ideation. No concerns.			
DIAGNOSTIC			
Major Depressive Disorder, Recurrent, Moderate Generalized Anxiety Disorder.			
PLAN FOR NEXT			
Continue to delve into past experiences and their connection for effective communication and assertiveness at work. Review restructuring.			
RETURN TO CLINIC			
Every week			
Instructed the client to call our office immediately if symptoms worsen. Instructed to seek immediate medical attention if suicidal thoughts, homicidal thoughts, or any other medical emergency.			
8/9/2023			
JOHN DOE, MFT, Psychotherapist			



Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

[The Court advises defe
by the public defender
his assigned attorney;
p.m. in open court].

THE COURT: Okay.

DEPENDANT ALLON:

or his father were ever around any insulation that got removed.

Now, you are going to hear from some folks who used that grinder. I don't think we are going to get a good idea of when it was used, but it was used. It was common. One thing you are going to hear is that it was outside, and guess what. No kids were around.

You are not going to hear that Mark's father was around. Mark's father worked in a place called the machine shop. He was a machinist. He built equipment. He was not an insulator. He was not a laborer. He did not remove insulation.

A very big part of this case -- and I think I mentioned this at the beginning of my conversation with you -- is what knowledge did a sugar beet processing facility have of the hazards of asbestos in the 1950s

Myself and two of my DRI colleagues will be testifying today. As background information, DRI is and has been the leading U.S. organization of civil defense attorneys and in-house counsel for more than 60 years. DRI hosts 29 substantive practice group committees and is home to the Center for Law and Public Policy.

THERAPY PROGRESS NOTE			
DATE OF SESSION	8/9/2023	DURATION OF SESSION	45 minutes
LOCATION OF SESSION	VIRTUAL		
PATIENT DEMOGRAPHIC INFORMATION			
CLIENT NAME	JOHN DOE	DOB	10/10/1980
CURRENT PSYCHIATRIST	JON JOHNS, MD	NP	7AM-11PM
PRESENTING PROBLEM	Client reports feeling overwhelmed over the past two weeks. Client discussed recent interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.		
SESSION CONTENT	Client discussed recent interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.		
INTERVENTIONS	Utilized cognitive restructuring techniques. Became tearful. The client was receptive to restructuring exercises. Did appreciate for the support.		
CLIENT RESPONSE	The client is motivated and patterns of behavior and th will develop effective coping.		
THERAPIST'S OBSERVATIONS	The client is motivated and patterns of behavior and th will develop effective coping.		
MENTAL STATUS			
Client presents in casual attire, appearing stressed. Normal mood as 'okay' and affect is congruent. Restricted but euthym content is without any overt delusions, hallucinations, suicidal oriented to person, place, and time. Client demonstrates good partially limited, and judgment appears adequate.			
RISK ASSESS			
No indications of suicidal or homicidal ideation. No concerns.			
DIAGNOSTIC			
Major Depressive Disorder, Recurrent, Moderate Generalized Anxiety Disorder.			
PLAN FOR NEXT			
Continue to delve into past experiences and their connection for effective communication and assertiveness at work. Review restructuring.			
RETURN TO CLINIC			
Every week			
Instructed the client to call our office immediately if symptoms worsen. Instructed to seek immediate medical attention if suicidal thoughts, homicidal thoughts, or any other medical emergency arise.			
8/9/2023			
JOHN DOE, MFT, Psychotherapist			

← Search Amazon

Customer reviews

★★★★☆ 4.1 out of 5

969 global ratings

5 star

4 star

3 star

2 star

1 star

Write a review

How customer review

Reviews with image

Top reviews

Filter »

969 total ratings, 247 with reviews

Hi Joe,

I hope all is well. If you get a moment I'd really appreciate it if you could write us a short testimonial?

Thank you so much!

Write testimonial

Hi Joe, do you have a few minutes to leave a short testimonial? Tap this link to get started: [onetap.reviews/abc](#)



Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.

1

[The Court advises defe

2

by the public defender

3

his assigned attorney,

4

p.m. in open court].

5

THE COURT: Okay.

6

DEPENDANT ALLPH:

7

1

8

2

9

3

10

4

11

5

12

6

13

7

14

8

15

9

16

10

17

11

18

12

19

13

20

14

21

15

22

16

23

17

24

18

25

19

26

20

27

21

28

22

29

23

30

24

31

25

32

26

33

27

34

28

35

29

36

30

37

31

38

32

39

33

40

34

41

35

42

36

43

37

44

38

45

39

46

40

47

41

48

42

49

43

50

44

51

45

52

46

53

47

54

48

55

49

56

50

57

51

58

52

59

53

60

54

61

55

62

56

63

57

64

58

65

59

66

60

67

61

68

62

69

63

70

64

71

65

72

66

73

67

74

68

75

69

76

70

77

71

78

72

79

73

80

74

81

75

82

76

83

77

84

78

85

79

86

80

87

81

88

82

89

83

90

84

91

85

92

86

93

87

94

88

95

89

96

90

97

91

98

92

99

93

100

94

101

95

102

96

103

97

104

98

105

99

106

100

107

101

108

102

109

103

110

104

111

105

112

106

113

107

114

108

115

109

116

110

117

111

118

112

119

113

120

114

121

115

122

116

123

117

124

118

125

119

126

120

127

121

128

122

129

123

130

124

131

125

132

126

133

127

134

128

135

129

136

130

137

131

138

132

139

133

140

134

141

135

142

136

143

137

144

138

145

139

146

140

147

141

148

142

149

143

150

144

151

145

152

146

153

147

154

148

155

149

156

150

157

151

158

152

159

153

160

154

161

155

162

156

163

157

164

158

165

159

166

160

167

161

168

162

169

163

170

164

171

165

172

166

173

167

174

168

175

169

176

170

177

171

178

172

179

173

180

174

181

175

182

176

183

177

184

178

185

179

186

180

187

181

188

182

189

183

190

184

191

185

192

186

193

187

194

188

195

189

196

190

197

191

198

192

199

193

200

194

201

195

202

196

203

197

204

198

205

199

206

200

207

201

208

202

209

203

210

204

211

205

212

206

213

207

214

208

215

209

216

210

217

211

218

212

219

213

220

214

221

215

222

216

223

217

224

218

225

219

226

220

227

221

228

222

229

223

230

224

231

225

232

226

233

227

234

228

235

229

236

230

237

231

238

232

239

233

240

234

241

235

242

236

243

237

244

238

245

239

246

240

247

241

248

242

249

243

250

244

251

245

252

246

253

247

254

248

255

249

256

250

257

251

258

252

259

253

260

254

261

255

262

256

263

257

264

258

265

259

266

260

267

261

268

262

269

263

270

264

271

265

272

266

273

267

274

268

275

269

276

270

277

271

278

272

279

273

280

274

281

275

282

276

283

277

284

278

285

279

286

280

287

281

288

282

289

283

290

284

291

285

292

286

293

287

294

288

295

289

296

290

297

291

298

292

299

293

300

294

301

295

302

296

303

297

304

298

305

299

306

300

307

301

308

302

309

303

310

304

311

305

312

306

313

307

314

308

315

309

316

310

317

311

318

312

319

313

320

314

321

315

322

316

323

317

324

318

325

319

326

320

327

321

328

322

329

323

330

324

331

325

332

326

333

327

334

328

335

329

336

330

337

331

338

332

339

333

340

334

341

335

342

336

343

337

344

338

345

339

346

340

347

341

348

342

349

343

350

344

351

345

352

346

353

347

354

348

355

349

356

350

357

351

358

352

359

353

360

354

361

355

362

356

363

357

364

358

365

359

366

360

367

361

368

362

369

363

370

364

371

365

372

366

373

367

374

368

375

369

376

370

377

371

378

372

379

373

380

374

381

375

382

376

383

377

384

378

385

379

386

380

387

381

388

382

389

383

390

384

391

385

392

386

393

387

394

388

395

389

396

390

397

391

398

392

399

393

400

394

401

395

402

396

403

397

404

398

405

399

406

400

407

401

408

402

409

403

410

404

411

405

412

406

413

407

414

408

415

409

416

410

417

411

418

412

419

413

420

414

421

415

422

416

423

417

424

418

425

419

426

420

427

421

428

422

429

423

430

424

431

425

432

426

433

427

434

428

435

429

436

430

437

431

438

432

439

433

440

434

441

435

442

436

443

437

444

438

445

439

446

440

447

441

448

442

449

443

450

444

451

445

452

446

453

447

454

448

455

449

456

450

457

451

458

452

459

453

460

454

461

455

462

456

463

457

464

458

465

459

466

460

467

461

468

462

469

463

470

464

471

465

472

466

473

467

474

468

475

469

476

470

477

471

478

472

479

473

480

474

481

475

482

476

483

477

484

478

485

479

486

480

487

481

488

482

489

483

490

484

491

485

492

486

493

487

494

488

495

489

496

490

497

491

498

492

499

493

500

494

501

495

502

496

503

497

504

498

505

499

506

500

507

501

508

502

509

503

510

504

511

505

512

506

513

507

514

508

515

509

516

510

517

511

518

512

519

513

520

514

521

515

522

516

523

517

524

518

525

519

526

520

527

521

528

522

529

523

530

524

531

525

532

526

533

527

534

528

535

529

536

530

537

531

538

532

539

533

540

534

541

535

542

536

543

537

544

538

545

539

546

540

547

541

548

542

549

543

550

544

551

545

552

546

553

547

554

548

555

549

556

550

557

551

558

552

559

553

560

554

561

555

562

556

563

557

564

558

565

559

566

560

567

561

568

562

569

563

570

564

571

565

572

566

573

567

574

568

575

569

576

570

577

571

578

572

579

573

580

574

581

575

582

576

583

577

584

578

585

579

586

580

587

581

588

582

589

583

590

584

591

585

592

586

593

587

594

588

595

589

596

590

597

591

598

592

599

593

600

594

601

595

602

596

603

597

604

598

605

599

606

600

607

601

608

602

609

603

610

604

611

605

612

606

613

607

614

608

615

609

616

610

617

611

618

612

619

613

620

614

621

615

622

616

623

617

624

618

<

Unstructured Data Problems

LLMs unlock the ability to reason over unstructured data.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

[The Court advises defe
by the public defender
his assigned attorney;
p.m. in open court].

THE COURT: Okay.

DEPENDANT ALLON:

or his father were ever around any insulation that got removed.

Now, you are going to hear from some folks who used that grinder. I don't think we are going to get a good idea of when it was used, but it was used. It was common. One thing you are going to hear is that it was outside, and guess what. No kids were around.

You are not going to hear that Mark's father was around. Mark's father worked in a place called the machine shop. He was a machinist. He built equipment. He was not an insulator. He was not a laborer. He did not remove insulation.

A very big part of this case -- and I think I mentioned this at the beginning of my conversation with you -- is what knowledge did a sugar beet facility have of the hazards of asbestos i

Myself and two of my DRI colleagues will be testifying today. As background information, DRI is and has been the leading U.S. organization of civil defense attorneys and in-house counsel for more than 60 years. DRI hosts 29 substantive practice group committees and is home to the Center for Law and Public Policy.

82

or his father were ever around any insulation that got removed.

Now, you are going to hear from some folks who used that grinder. I don't think we are going to get a good idea of when it was used, but it was used. It was common. One thing you are going to hear is that it was outside, and guess what. No kids were around.

You are not going to hear that Mark's father was around. Mark's father worked in a place called the machine shop. He was a machinist. He built equipment. He was not an insulator. He was not a laborer. He did not remove insulation.

A very big part of this case -- and I think I mentioned this at the beginning of my conversation with you -- is what knowledge did a sugar beet facility have of the hazards of asbestos i

Myself and two of my DRI colleagues will be testifying today. As background information, DRI is and has been the leading U.S. organization of civil defense attorneys and in-house counsel for more than 60 years. DRI hosts 29 substantive practice group committees and is home to the Center for Law and Public Policy.

DATE OF SESSION

8/9/2023

LOCATION OF SESSION

VIRTUAL

DURATION OF SESSION

45 minutes

PATIENT DEMOGRAPHIC INFORMATION

CLIENT NAME

JOHN DOE

CURRENT PSYCHIATRIST

JON JOHNS, MD

PRESENTING PROBLEM

Client reports feeling over the past two weeks. Client discussed recent ch interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.

SESSION CONTENT

Utilized cognitive restructuring feelings of inadequacy. Tack acute anxiety. Explored past anxiety and fear of judgment. The client was receptive to restructuring exercises. Del techniques. Became tearful appreciation for the support.

CLIENT RESPONSE

The client is motivated and patterns of behavior and th will develop effective coping.

THERAPIST'S OBSERVATIONS

The client is motivated and patterns of behavior and th will develop effective coping.

MENTAL STATUS

Client presents in casual attire, appearing stated age. Normal cooperative, though somewhat reserved. Speech is of normal mood as 'okay' and affect is congruent, restricted but euthym content is without any overt delusions, hallucinations, suicidal oriented to person, place, and time. Client demonstrates good partially limited, and judgment appears adequate.

RISK ASSE

No indications of suicidal or homicidal ideation. No concerns.

DIAGNOSTIC

Major Depressive Disorder, Recurrent, Moderate Generalized Anxiety Disorder.

DATE OF SESSION

8/9/2023

LOCATION OF SESSION

VIRTUAL

DURATION OF SESSION

45 minutes

PATIENT DEMOGRAPHIC INFORMATION

CLIENT NAME

JOHN DOE

CURRENT PSYCHIATRIST

JON JOHNS, MD

PRESENTING PROBLEM

Client reports feeling over the past two weeks. Client discussed recent ch interpersonal conflicts with making mistakes. Described lack day. Briefly touched on feelings of anxiety.

SESSION CONTENT

Utilized cognitive restructuring feelings of inadequacy. Tack acute anxiety. Explored past anxiety and fear of judgment. The client was receptive to restructuring exercises. Del techniques. Became tearful appreciation for the support.

CLIENT RESPONSE

The client is motivated and patterns of behavior and th will develop effective coping.

THERAPIST'S OBSERVATIONS

The client is motivated and patterns of behavior and th will develop effective coping.

MENTAL STATUS

Client presents in casual attire, appearing stated age. Normal cooperative, though somewhat reserved. Speech is of normal mood as 'okay' and affect is congruent, restricted but euthym content is without any overt delusions, hallucinations, suicidal oriented to person, place, and time. Client demonstrates good partially limited, and judgment appears adequate.

RISK ASSE

No indications of suicidal or homicidal ideation. No concerns.

DIAGNOSTIC

Major Depressive Disorder, Recurrent, Moderate Generalized Anxiety Disorder.

Customer reviews

★★★★☆ 4.1 out of 5

969 global ratings

5 star

4 star

3 star

2 star

1 star

Write a review

Write testimonial

Hi Joe, I hope all is well. If you get a moment I'd really appreciate it if you could write us a short testimonial? Tap this link to get started: onetap.reviews/abc

Building systems to do this reliably and efficiently is hard!

Filter »



A Hard Unstructured Data Task



User's Goal:

Study the climate initiatives undertaken by Scottish companies.

*Status: This version of this Act contains provisions that are prospective.
Changes to legislation: There are currently no known outstanding effects for the Climate Change (Scotland) Act 2009. (See end of Document for details)*

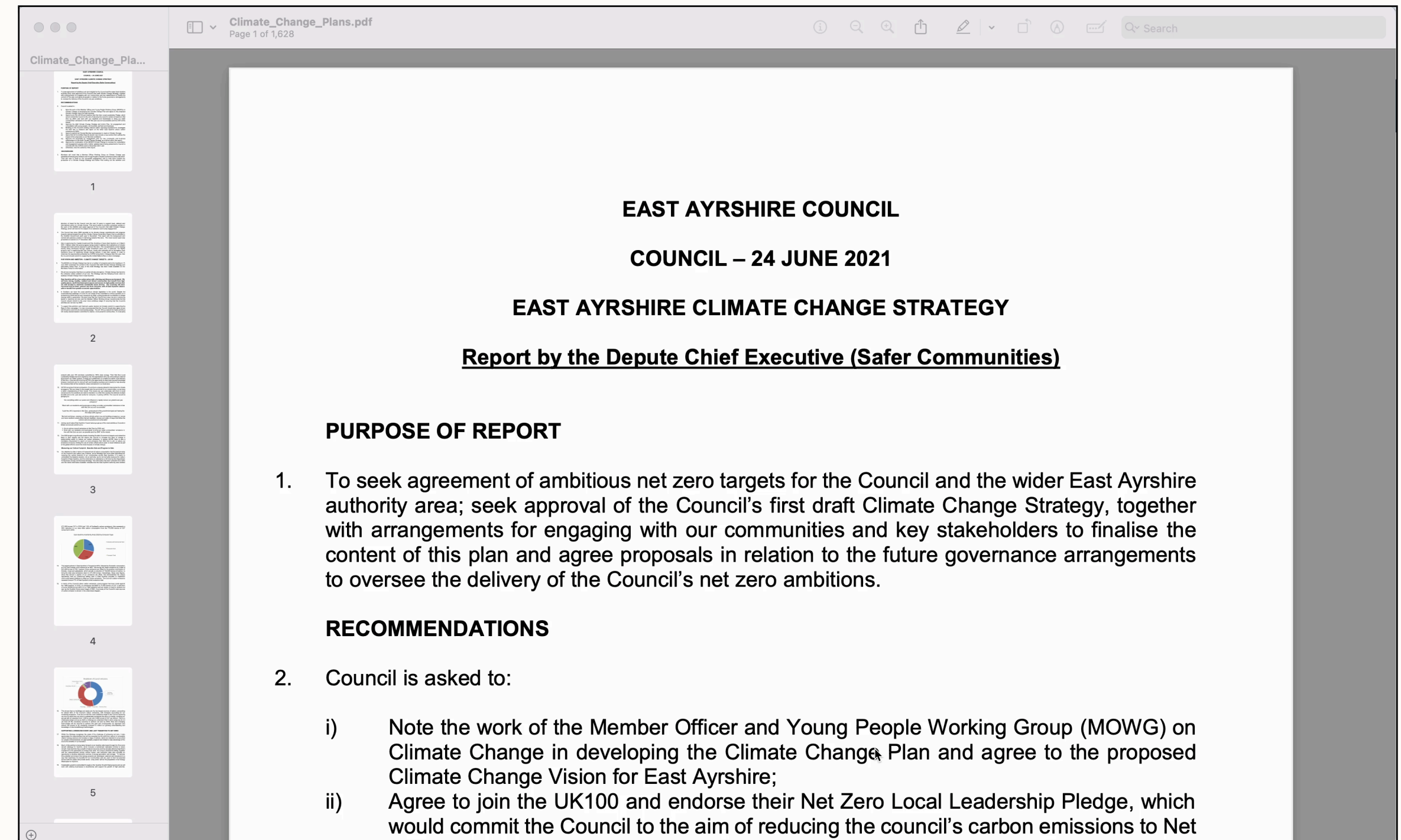


Climate Change (Scotland) Act 2009 2009 asp 12

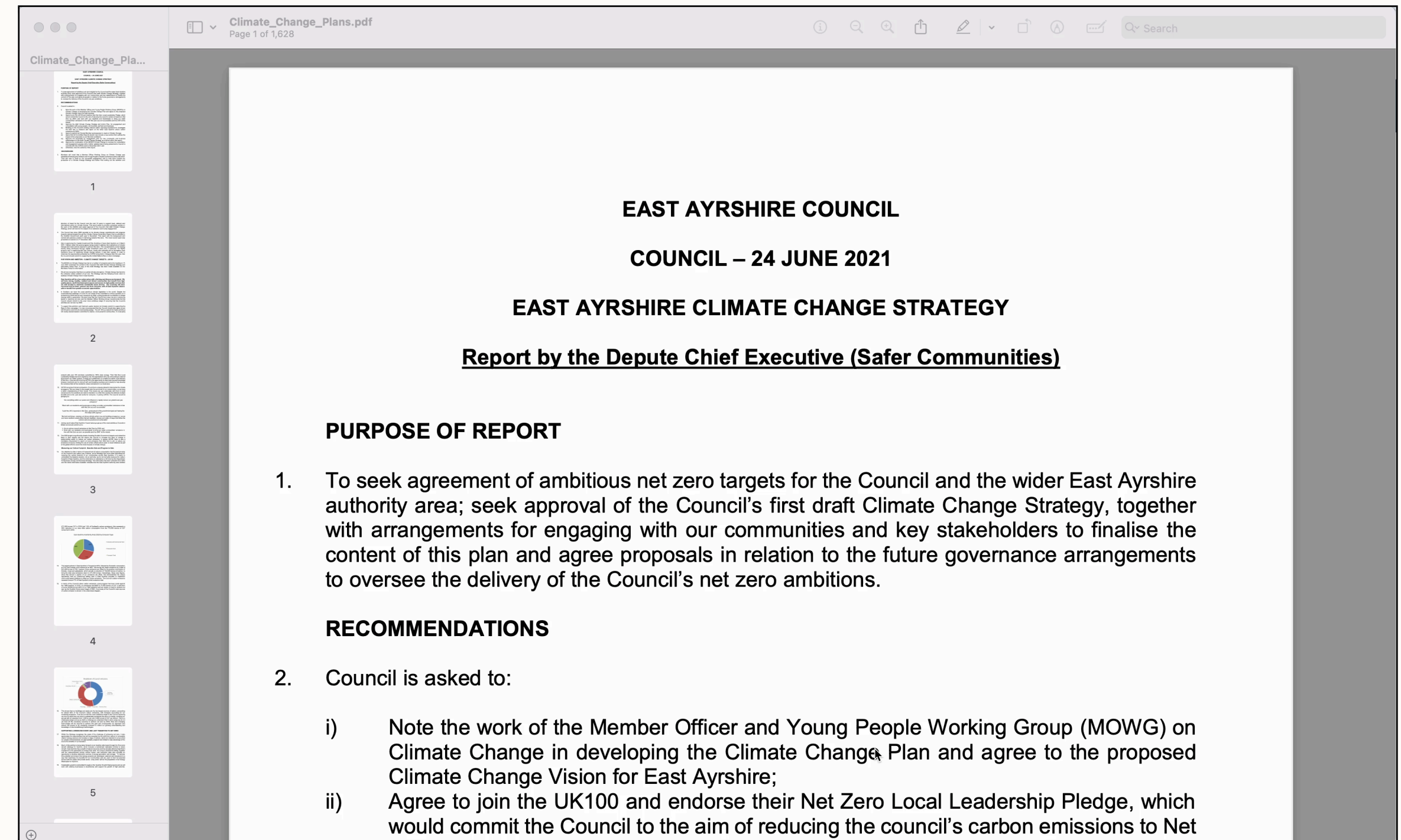
The Bill for this Act of the Scottish Parliament was passed by the Parliament on 24th June 2009 and received Royal Assent on 4th August 2009

An Act of the Scottish Parliament to set a target for the year 2050, an interim target for the year 2020, and to provide for annual targets, for the reduction of greenhouse gas emissions; to provide about the giving of advice to the Scottish Ministers relating to climate change; to confer power on Ministers to impose climate change duties on public bodies; to make further provision about mitigation of and adaptation to climate change; to make provision about energy efficiency, including provision enabling council tax discounts; to make provision about the reduction and recycling of waste; and for connected purposes.

A Hard Unstructured Data Task

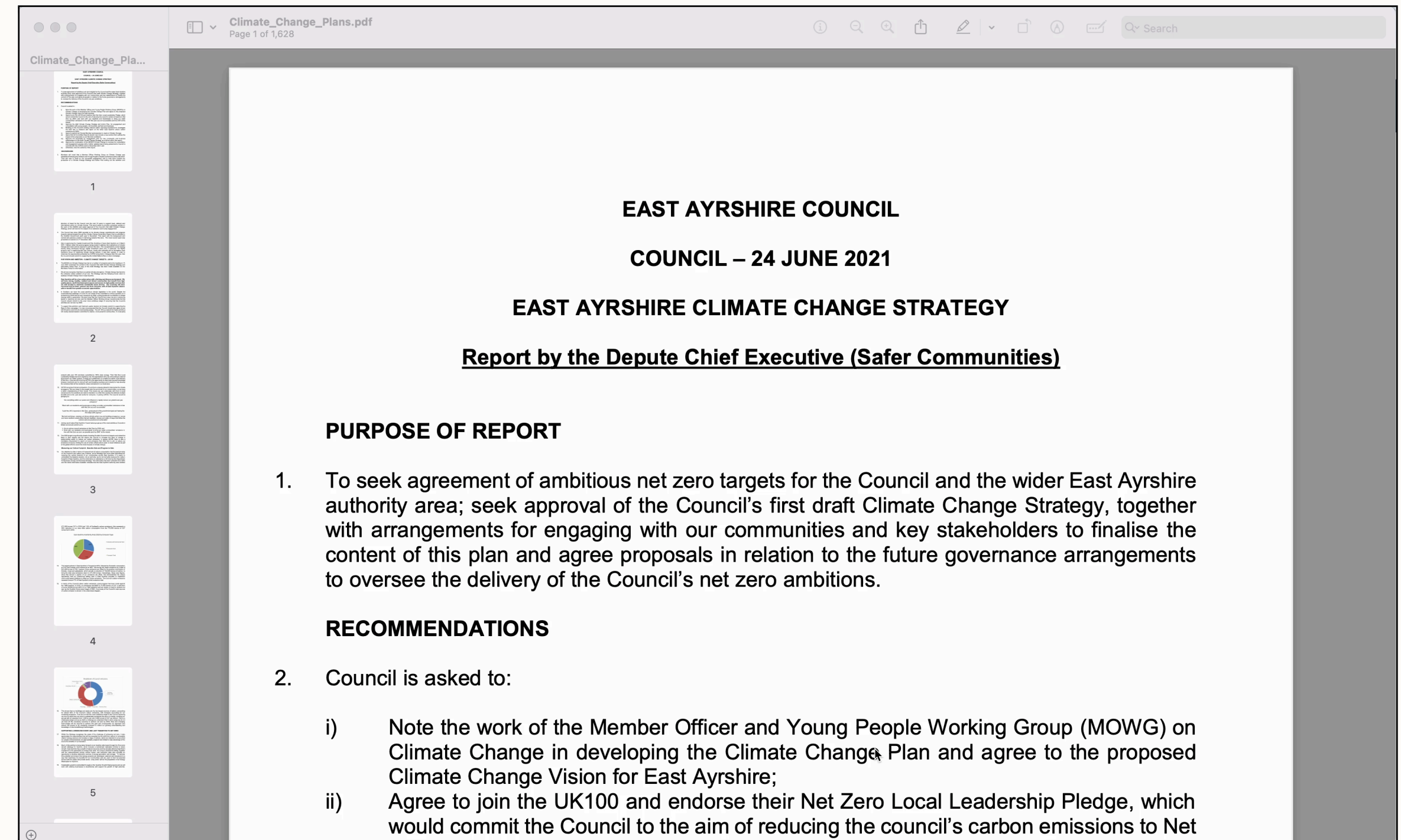


A Hard Unstructured Data Task



A Hard Unstructured Data Task

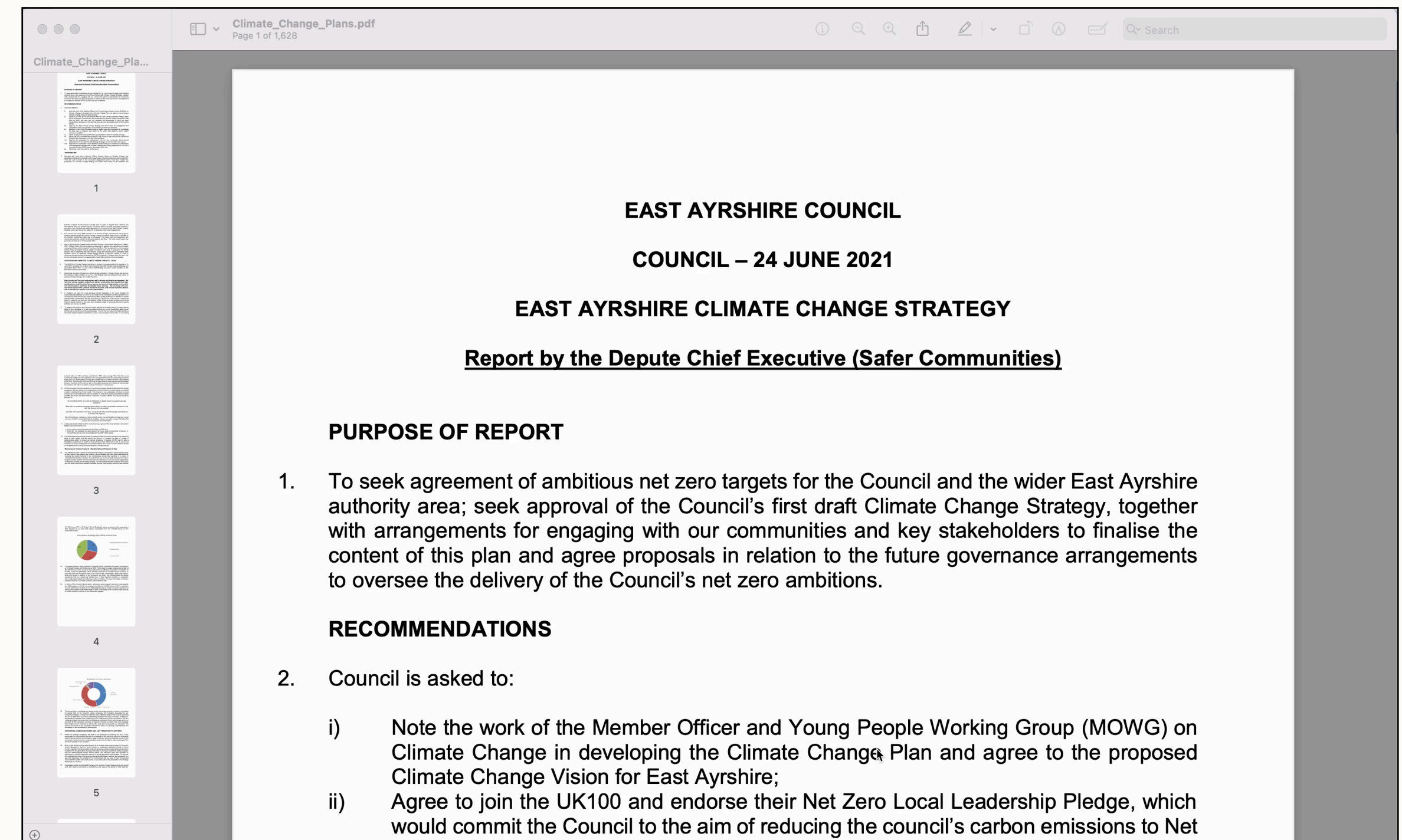
Task: Extract, summarize, & aggregate proposed climate interventions (e.g., "EV subsidy") from corporate sustainability reports, meeting minutes, etc.



A Hard Unstructured Data Task

Task: Extract, summarize, & aggregate proposed climate interventions (e.g., "EV subsidy") from corporate sustainability reports, meeting minutes, etc.

Data Volume: 32 PDFs (~1500 pages each) per month dating back last 10 years

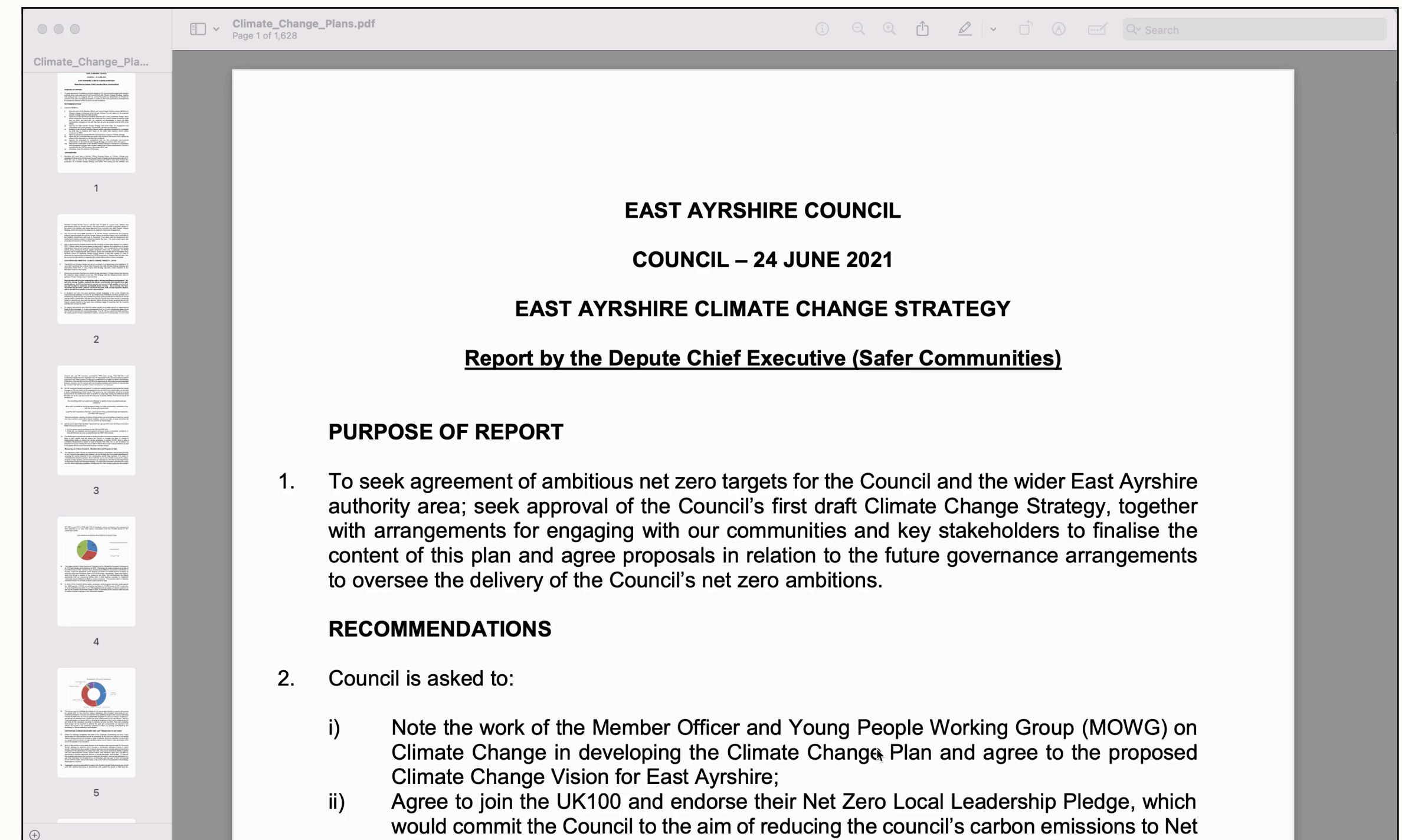


A Hard Unstructured Data Task

Task: Extract, summarize, & aggregate proposed climate interventions (e.g., "EV subsidy") from corporate sustainability reports, meeting minutes, etc.

Data Volume: 32 PDFs (~1500 pages each) per month dating back last 10 years

Cost: >\$5k to process each monthly batch



Another Hard Unstructured Data Task

User's Goal:

Statistical analyses of explicit and implicit biases in court transcripts, police reports, news articles, etc.

 [HOME](#) [ABOUT US](#) [CLIENTS](#) [MEDIA + PUBLIC](#) [CAREERS](#)

MEDIA CONTACT: PDR-MediaRelations@sfgov.org

****PRESS RELEASE****

SF Public Defenders Win First Racial Justice Act Motion in San Francisco

Judge grants motion finding implicit bias with police officer's testimony

SAN FRANCISCO – San Francisco Public Defenders have won the first California Racial Justice Act (RJA) motion in San Francisco since the law was enacted in 2021. An RJA hearing for a young Black man revealed that a police officer exhibited implicit bias during the man's arrest and during trial testimony. This led the judge to reduce certain felony convictions to misdemeanors as a remedy for that discrimination, as provided under the RJA.

The California Racial Justice Act [states that](#), "Implicit bias, although often unintentional and unconscious, may inject racism and unfairness into proceedings similar to intentional bias. The intent of the Legislature is not to punish this type of bias, but rather to remedy the harm to the defendant's case and to the integrity of the judicial system."

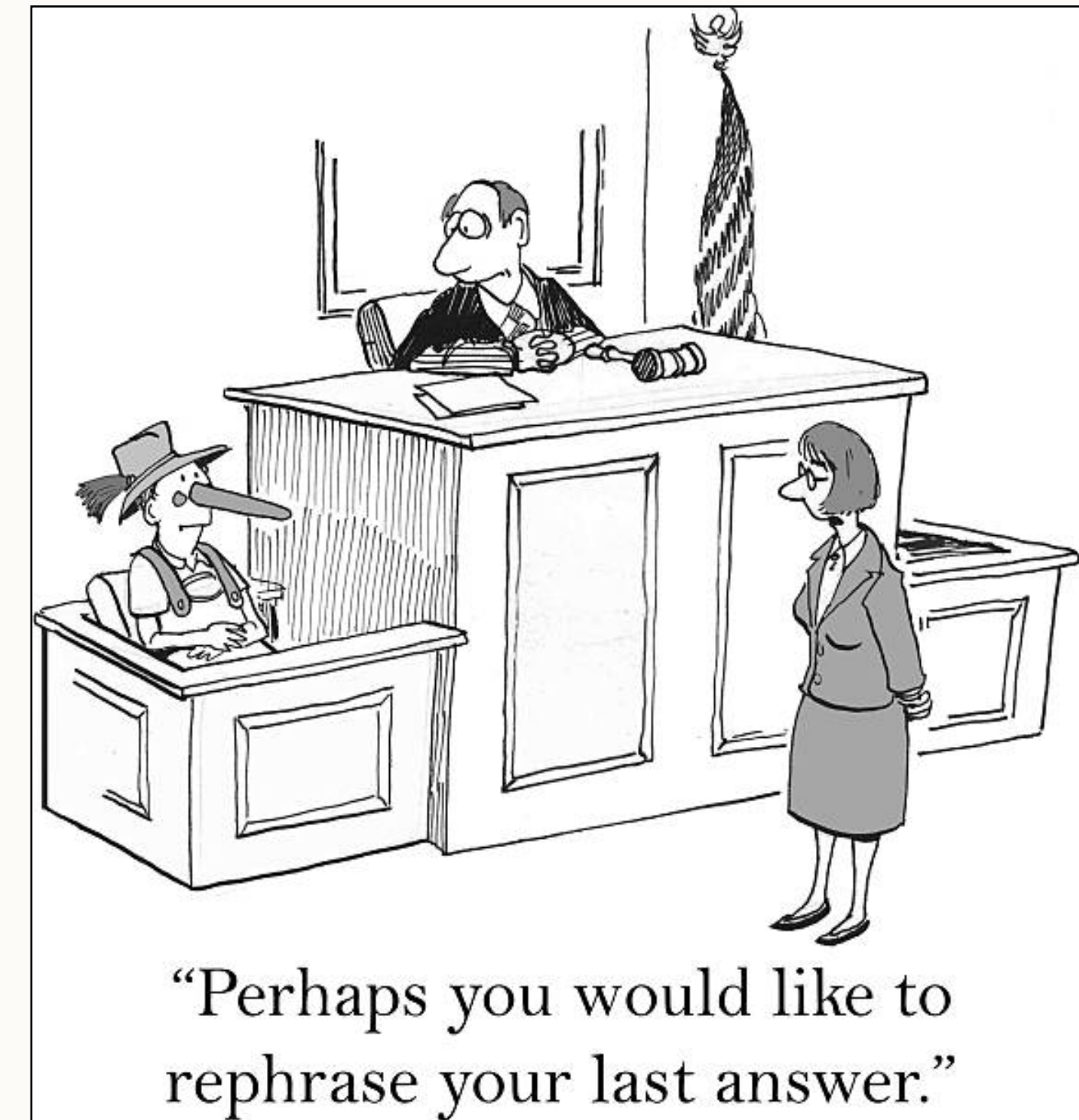
"Implicit bias plays a huge role in our legal system—from police to prosecutors to judges—and has historically resulted in the over-policing, over-charging, and over-sentencing of people of color," said **Deputy Public Defender Diamond Ward**, who represented Adonte Bailey, who was granted relief under the RJA.

Bailey, a 22-year-old Black man, was on trial for an incident involving a report of someone holding a gun while standing on the street. Evidence at trial showed that when police eventually arrested Bailey, the officer said on body-worn camera that he was "a little surprised he didn't run." When the officer later testified at trial, he painted an untrue picture of Bailey as acting evasive—telling the jury that Bailey was "ducking" and "bobbing" and had "darted"—which was contradicted by video footage of Bailey being cooperative and by other witnesses' testimony. The officer also ignored the court's orders and made unsolicited statements that he had been notified were inadmissible in front of the jury.

Bailey's attorneys filed a motion arguing that the arresting officer violated the RJA by his comments and actions at the arrest scene and in his trial testimony.

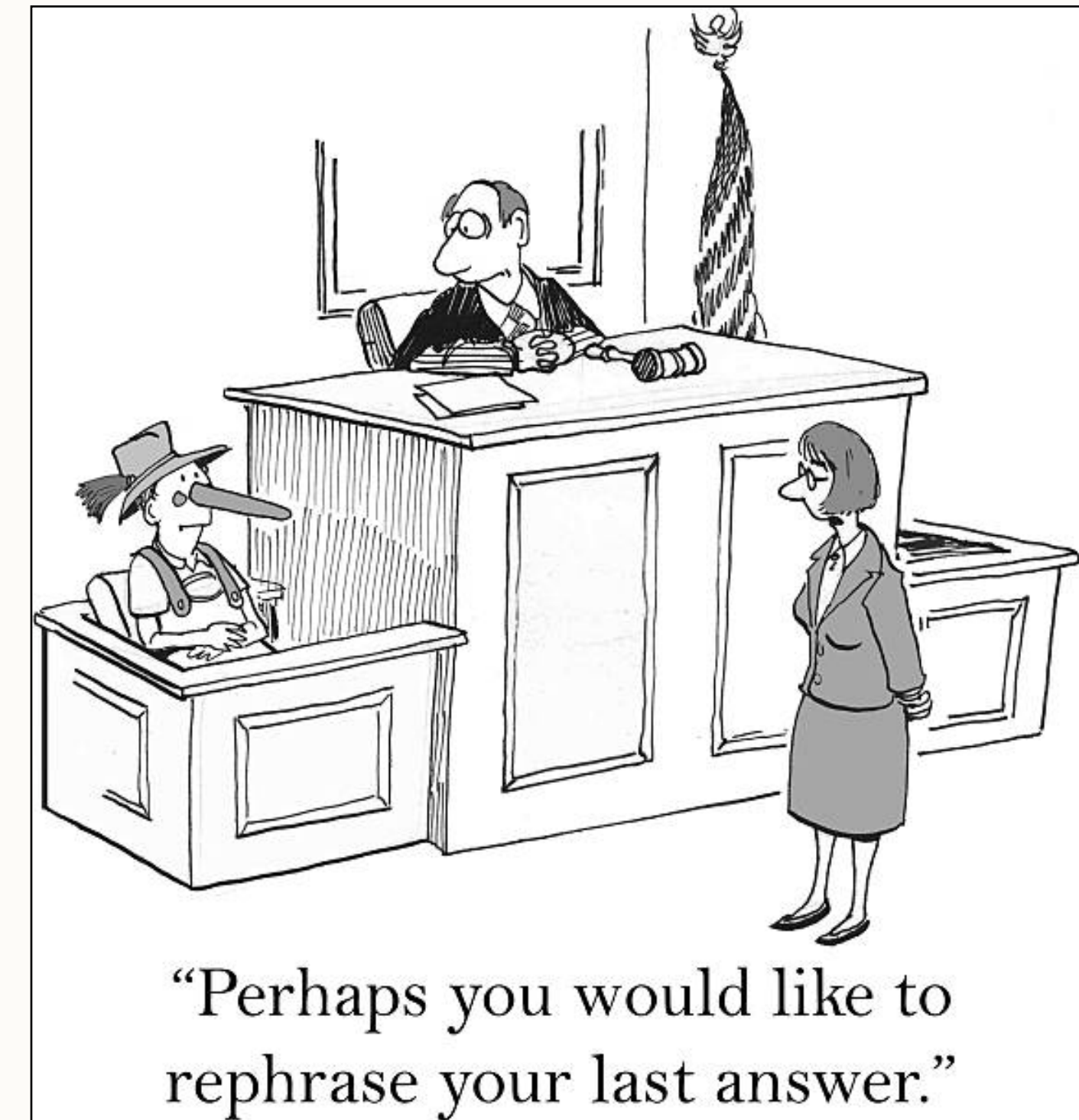
During the RJA hearing, an expert witness on race and the criminal legal system testified that the officer exhibited implicit

Another Hard Unstructured Data Task



Another Hard Unstructured Data Task

Task: Detect and compare racially coded language and implicit bias across similar court cases.



Another Hard Unstructured Data Task

Task: Detect and compare racially coded language and implicit bias across similar court cases.

Data Volume: 1,500 pages per case; 300 related cases; 135 million words per analysis + images.



Another Hard Unstructured Data Task

Task: Detect and compare racially coded language and implicit bias across similar court cases.

Data Volume: 1,500 pages per case; 300 related cases; 135 million words per analysis + images.

Cost: \$11 million a year if analyzing data for all cases.



Unifying Characteristics

Unifying Characteristics

Unstructured data tasks often require, at scale:

Unifying Characteristics

Unstructured data tasks often require, at scale:

- ◆ Complex reasoning

Unifying Characteristics

Unstructured data tasks often require, at scale:

- ◆ Complex reasoning
- ◆ over *all* documents

Unifying Characteristics

Unstructured data tasks often require, at scale:

- ◆ Complex reasoning
- ◆ over *all* documents
- ◆ to generate open-ended, subjective outputs

Unifying Characteristics

Unstructured data tasks often require, at scale:

- ◆ Complex reasoning
- ◆ over *all* documents
- ◆ to generate open-ended, subjective outputs
- ◆ that feed into downstream aggregations and summaries.

Challenges in Unstructured Data Analysis

Challenges in Unstructured Data Analysis

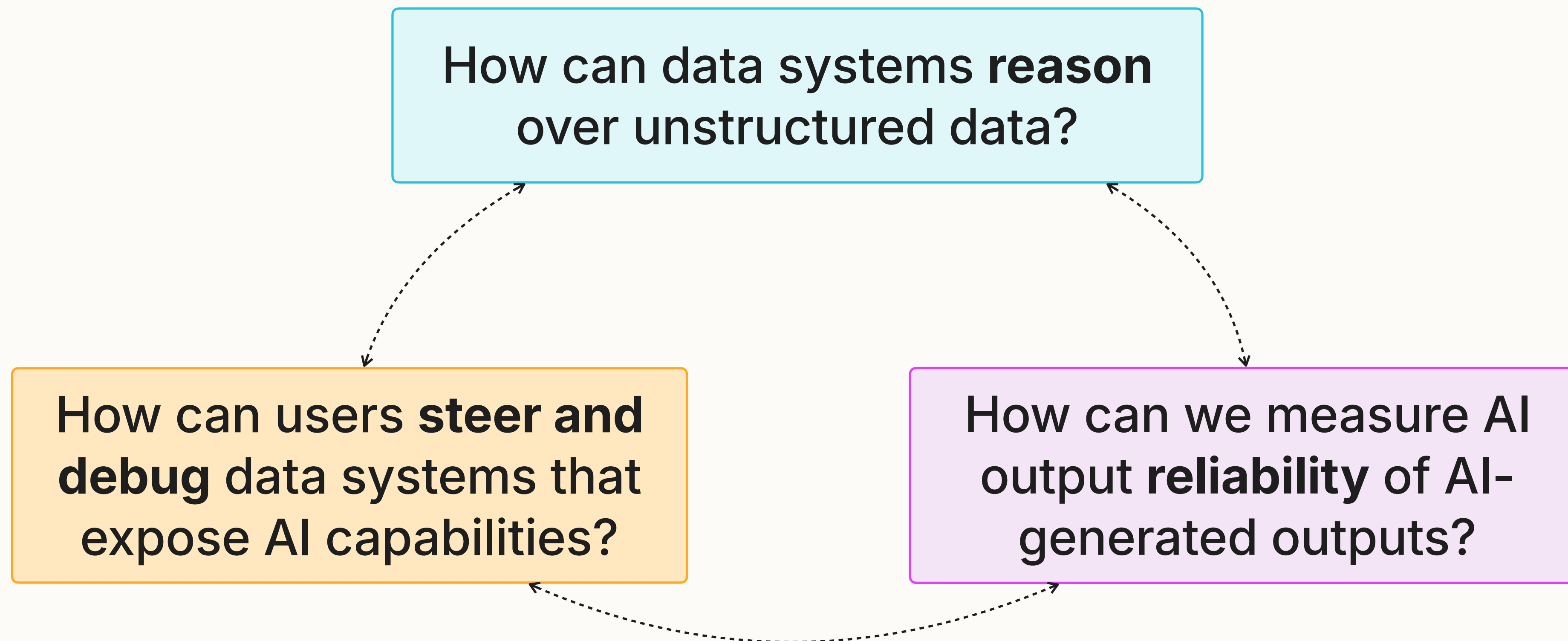
**How can data systems reason
over unstructured data?**

Challenges in Unstructured Data Analysis

How can data systems **reason** over unstructured data?

How can users **steer and debug** data systems that expose AI capabilities?

Challenges in Unstructured Data Analysis



My Work

How can data systems **reason** over unstructured data?

How can users **steer and debug** data systems that expose AI capabilities?

How can we measure the **reliability** of AI-generated outputs?

My Work

How can data systems **reason** over unstructured data?

How can users **steer and debug** data systems that expose AI capabilities?

How can we measure the **reliability** of AI-generated outputs?

My Work

How can data systems **reason** over unstructured data?

Multi-Objective Agentic Rewrites. Wei*, **Shankar*** et al. *Under Review*.
Task Cascades. **Shankar** et al. *Under Revision at SIGMOD '26*.
BARGAIN. Zeighami, **Shankar** et al. *SIGMOD '26*.
DocETL. **Shankar** et al. *VLDB '25*.
ZenDB. Lin, Hulsebos, Ma, **Shankar** et al. *ICDE '25*.

How can users **steer and debug** data systems that expose AI capabilities?

How can we measure the **reliability** of AI-generated outputs?

*Joint first author work.

My Work

How can data systems **reason** over unstructured data?

Multi-Objective Agentic Rewrites. Wei*, **Shankar*** et al. *Under Review*.
Task Cascades. **Shankar** et al. *Under Revision at SIGMOD '26*.
BARGAIN. Zeighami, **Shankar** et al. *SIGMOD '26*.
DocETL. **Shankar** et al. *VLDB '25*.
ZenDB. Lin, Hulsebos, Ma, **Shankar** et al. *ICDE '25*.

How can users **steer and debug** data systems that expose AI capabilities?

DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆
RAGGY. Romero Lauro*, **Shankar*** et al. *Under Revision at CHI '26*.
DataScout. Lin*, Chopra*, Lin, **Shankar** et al. *UIST '25*.
Operationalizing ML, an Interview Study. **Shankar***, Garcia* et al. *CSCW '24*.
MLTrace. **Shankar** et al. *VLDB 2023*.
NBSlicer. **Shankar***, Macke* et al. *VLDB 2023*.

How can we measure the **reliability** of AI-generated outputs?

*Joint first author work.

My Work

How can data systems **reason** over unstructured data?

Multi-Objective Agentic Rewrites. Wei*, **Shankar*** et al. *Under Review*.
Task Cascades. **Shankar** et al. *Under Revision at SIGMOD '26*.
BARGAIN. Zeighami, **Shankar** et al. *SIGMOD '26*.
DocETL. **Shankar** et al. *VLDB '25*.
ZenDB. Lin, Hulsebos, Ma, **Shankar** et al. *ICDE '25*.

How can users **steer and debug** data systems that expose AI capabilities?

DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆
RAGGY. Romero Lauro*, **Shankar*** et al. *Under Revision at CHI '26*.
DataScout. Lin*, Chopra*, Lin, **Shankar** et al. *UIST '25*.
Operationalizing ML, an Interview Study. **Shankar***, Garcia* et al. *CSCW '24*.
MLTrace. **Shankar** et al. *VLDB 2023*.
NBSlicer. **Shankar***, Macke* et al. *VLDB 2023*.

How can we measure the **reliability** of AI-generated outputs?

AI Evals for Engineers and PMs. **Shankar** and Husain. *O'Reilly '26*.
PromptEvals. Vir*, **Shankar*** et al. *NAACL 2025*.
Who Validates the Validators? **Shankar** et al. *UIST '24*.
SPADE. **Shankar** et al. *VLDB '24*.
Automatic and Precise Data Validation for ML. **Shankar** et al. *CIKM 2023*.

*Joint first author work.

Today's Talk

How can data systems **reason** over unstructured data?

Multi-Objective Agentic Rewrites. Wei*, **Shankar*** et al. *Under Review*.
Task Cascades. **Shankar** et al. *Under Revision at SIGMOD '26*.

DocETL. **Shankar** et al. *VLDB '25*.

How can users **steer and debug** data systems that expose AI capabilities?

DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆

How can we measure the **reliability** of AI-generated outputs?

AI Evals for Engineers and PMs. **Shankar** and Husain. *O'Reilly '26*.

Who Validates the Validators? **Shankar** et al. *UIST '24*.

*Joint first author work.

Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities?
DocWrangler (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)

Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)

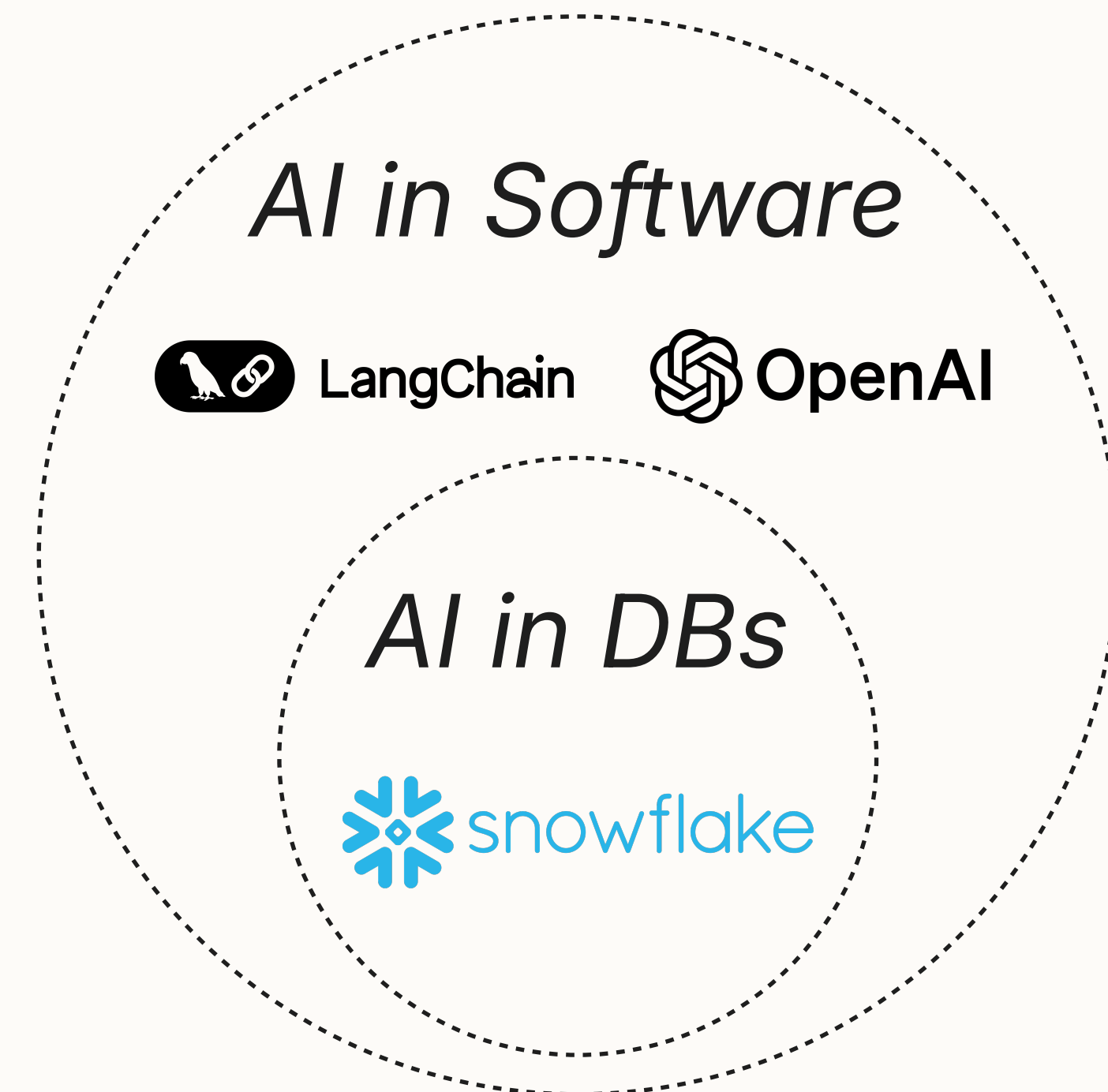


Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)

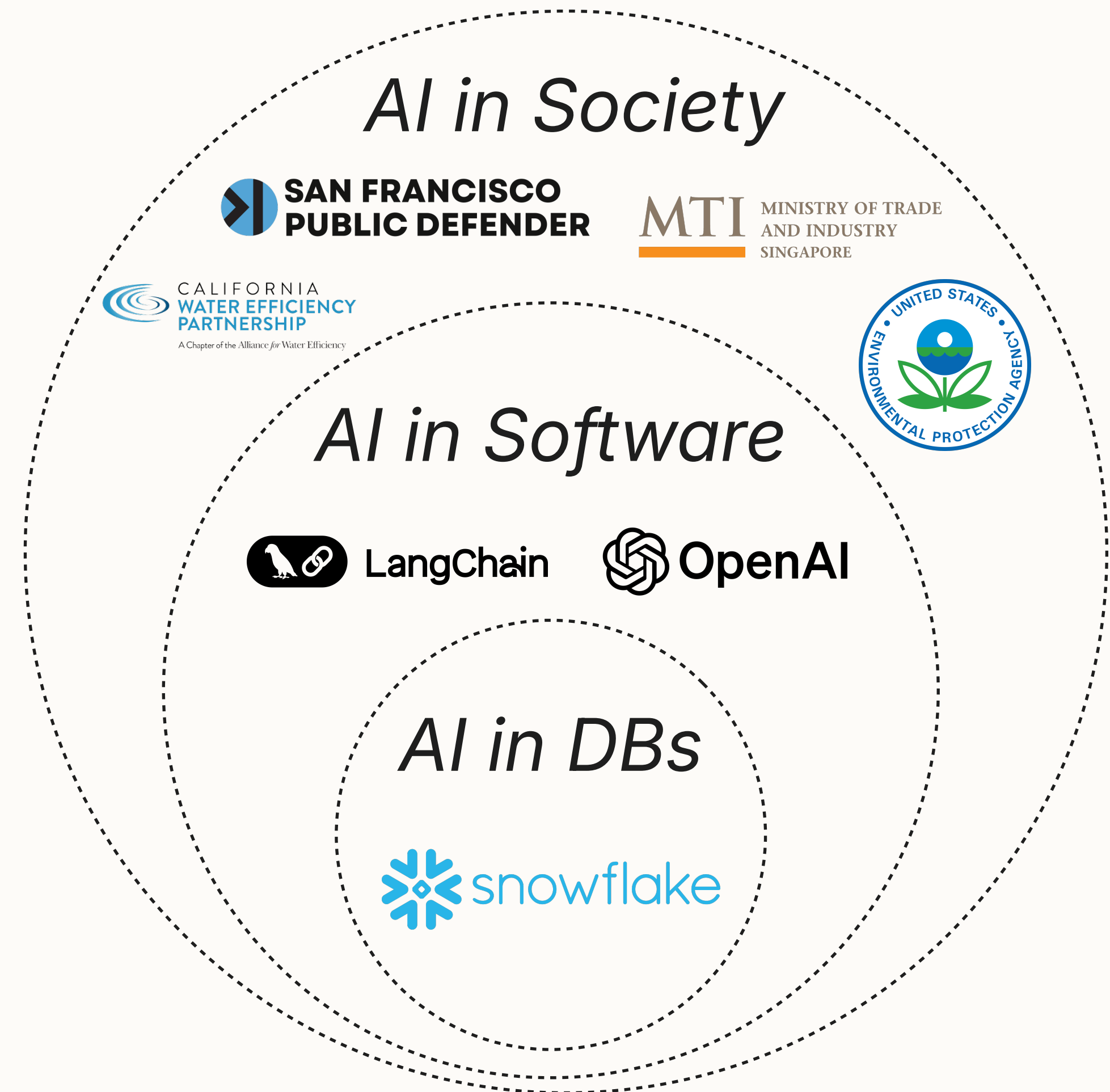


Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)



Programming Model

Programming Model

Background

In data systems, users build pipelines of operators (e.g., filter, map, reduce, join).

Programming Model

Background

In data systems, users build pipelines of operators (e.g., filter, map, reduce, join).

LLM Analogy

Programming Model

Background

In data systems, users build pipelines of operators (e.g., filter, map, reduce, join).

LLM Analogy

Expose LLMs through *semantic operators**, each defined by

*coined by Patel et al. VLDB 2025

Programming Model

Background

In data systems, users build pipelines of operators (e.g., filter, map, reduce, join).

LLM Analogy

Expose LLMs through *semantic operators**, each defined by

- ◆ Natural language description of the task
- ◆ Operator type (e.g., map, filter)
- ◆ Desired output schema

*coined by Patel et al. VLDB 2025

Programming Model

Background

In data systems, users build pipelines of operators (e.g., filter, map, reduce, join).

LLM Analogy

Expose LLMs through *semantic operators**, each defined by

- ♦ Natural language description of the task
- ♦ Operator type (e.g., map, filter)
- ♦ Desired output schema

Goal: run a pipeline of semantic (and possibly relational) operators, accurately and cheaply.

*coined by Patel et al. VLDB 2025

DocETL: A Declarative System for Semantic Operators Over Text

Semantic Map

Semantic Map

Dataset

```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 2,  
    "transcript":  
    "Last time we..."  
  },  
  ...  
]
```

Semantic Map

Dataset

```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 2,  
    "transcript":  
    "Last time we..."  
  },  
  ...  
]
```

Pipeline (Single Operator)

```
type: map  
prompt: "Extract all  
statements made by the  
judge from {{transcript}}  
that indicate implicit  
bias, and explain why")  
output:  
  schema:  
    statements: list[string]  
    explanation: string
```

Semantic Map

Dataset

```
[
  {
    "id": 1,
    "transcript":
      "This is the..."
  },
  {
    "id": 2,
    "transcript":
      "Last time we..."
  },
  ...
]
```

Pipeline (Single Operator)

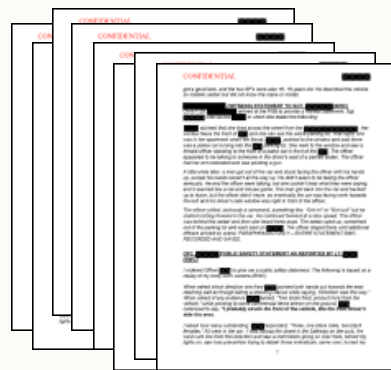
```
type: map
prompt: "Extract all
statements made by the
judge from {{transcript}}
that indicate implicit
bias, and explain why")
output:
  schema:
    statements: list[string]
    explanation: string
```

Output

```
[
  {
    "id": 1,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  {
    "id": 2,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  ...
]
```


Semantic Map

Dataset



```
[
  {
    "id": 1,
    "transcript":
      "This is the..."
  },
  {
    "id": 2,
    "transcript":
      "Last time we..."
  },
  ...
]
```

Pipeline (Single Operator)

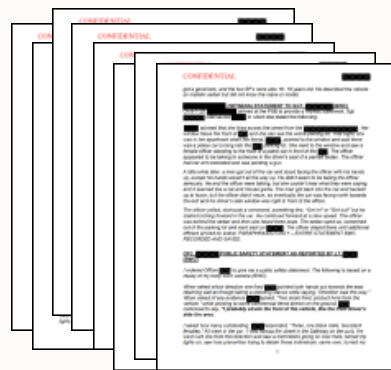
```
type: map
prompt: "Extract all
statements made by the
judge from {{transcript}}
that indicate implicit
bias, and explain why")
output:
  schema:
    statements: list[string]
    explanation: string
```

Output

```
[
  {
    "id": 1,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  {
    "id": 2,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  ...
]
```

Semantic Map

Dataset

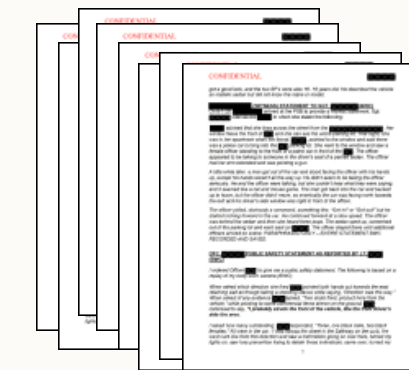


```
[
  {
    "id": 1,
    "transcript":
      "This is the..."
  },
  {
    "id": 2,
    "transcript":
      "Last time we..."
  },
  ...
]
```

Pipeline (Single Operator)

```
type: map
prompt: "Extract all
statements made by the
judge from {{transcript}}
that indicate implicit
bias, and explain why")
output:
  schema:
    statements: list[string]
    explanation: string
```

Output



```
[
  {
    "id": 1,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  {
    "id": 2,
    "transcript": ...,
    "statements": [...],
    "explanation": ...,
  },
  ...
]
```

Semantic Filter

Semantic Filter

Dataset

```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 2,  
    "transcript":  
    "Last time we..."  
  },  
  ... 10,000 more ...  
]
```

Semantic Filter

Dataset

```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 2,  
    "transcript":  
    "Last time we..."  
  },  
  ... 10,000 more ...  
]
```

Pipeline (Single Operator)

```
type: filter  
prompt: "Did the judge in  
{{transcript}} say  
anything that indicates  
implicit bias?")  
output:  
  schema:  
    violation: bool
```


Semantic Filter

Dataset

```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 2,
    "transcript":
    "Last time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

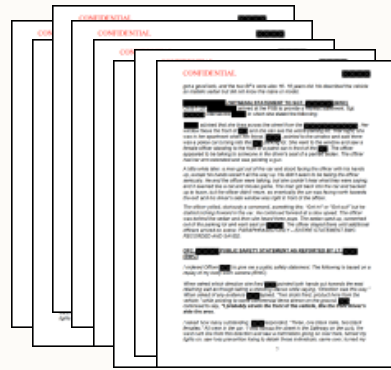
```
type: filter
prompt: "Did the judge in
{{transcript}} say
anything that indicates
implicit bias?")
output:
  schema:
    violation: bool
```

Output

```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 87,
    "transcript":
    "Once we start..."
  },
  ... 138 more ...
]
```

Semantic Filter

Dataset



```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 2,
    "transcript":
    "Last time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

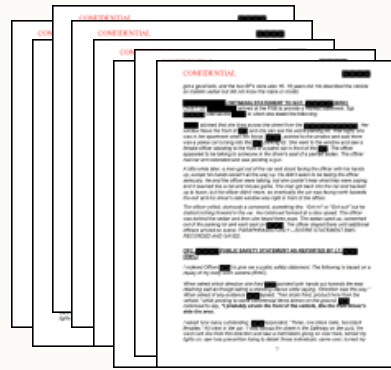
```
type: filter
prompt: "Did the judge in
{{transcript}} say
anything that indicates
implicit bias?")
output:
  schema:
    violation: bool
```

Output

```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 87,
    "transcript":
    "Once we start..."
  },
  ... 138 more ...
]
```

Semantic Filter

Dataset

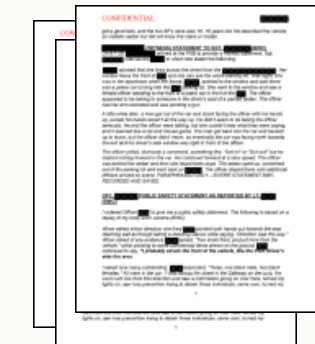


```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 2,
    "transcript":
    "Last time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

```
type: filter
prompt: "Did the judge in
{{transcript}} say
anything that indicates
implicit bias?")
output:
  schema:
    violation: bool
```

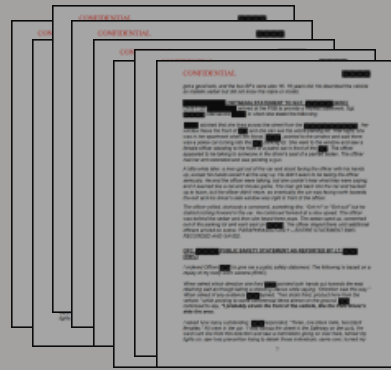
Output



```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 87,
    "transcript":
    "Once we start..."
  },
  ... 138 more ...
]
```

Semantic Filter

Dataset

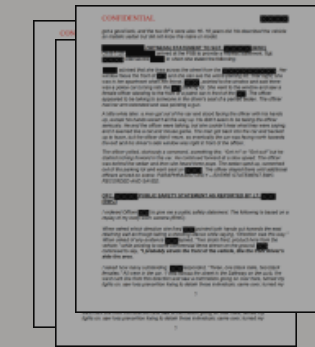


```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 2,
    "transcript":
    "Last time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

```
type: filter
prompt: "Did the judge in
{{transcript}} say
anything that indicates
implicit bias?")
output:
  schema:
    violation: bool
```

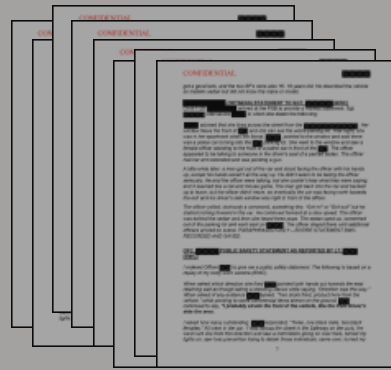
Output



```
[
  {
    "id": 1,
    "transcript":
    "This is the..."
  },
  {
    "id": 87,
    "transcript":
    "Once we start..."
  },
  ... 138 more ...
]
```

Semantic Filter

Dataset



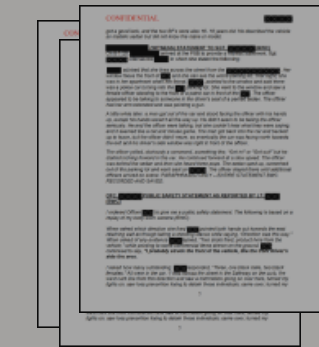
```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 2,  
    "transcript":  
    "Last time we  
    ... 10,000 more ...  
  }  
]
```

Pipeline (Single Operator)

```
type: filter  
prompt: "Did the judge in  
{{transcript}} say  
anything that indicates  
implicit bias?")
```

Like a semantic map, but with filter semantics.

Output



```
[  
  {  
    "id": 1,  
    "transcript":  
    "This is the..."  
  },  
  {  
    "id": 87,  
    "transcript":  
    "Once we start..."  
    ... 138 more ...  
  }  
]
```


Semantic Reduce

Semantic Reduce

Dataset

```
[  
  {  
    "id": 1,  
    "name": "Johnny...",  
    "transcript": "This  
is the..."  
  },  
  {  
    "id": 2,  
    "name": "Jake...",  
    "transcript": "Last  
time we..."  
  },  
  ... 10,000 more ...  
]
```

Semantic Reduce

Dataset

```
[
  {
    "id": 1,
    "name": "Johnny...",
    "transcript": "This
is the..."
  },
  {
    "id": 2,
    "name": "Jake...",
    "transcript": "Last
time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

```
type: reduce
reduce_key: name
prompt: "Summarize the
common implicit biases shown
by this judge across all
their trial transcripts:
{{transcripts}}")
output:
  schema:
    common_biases: str
```

Semantic Reduce

Dataset

```
[
  {
    "id": 1,
    "name": "Johnny...",
    "transcript": "This
is the..."
  },
  {
    "id": 2,
    "name": "Jake...",
    "transcript": "Last
time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

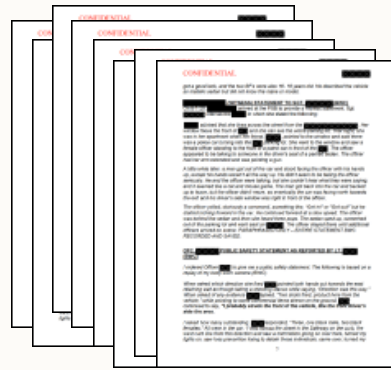
```
type: reduce
reduce_key: name
prompt: "Summarize the
common implicit biases shown
by this judge across all
their trial transcripts:
{{transcripts}}")
output:
  schema:
    common_biases: str
```

Output

```
[
  {
    "name": "Johnny...",
    "common_biases": ...
  },
  {
    "name": "Jake...",
    "common_biases": ...
  },
  ... 250 more ...
]
```

Semantic Reduce

Dataset



```
[
  {
    "id": 1,
    "name": "Johnny...",
    "transcript": "This
is the..."
  },
  {
    "id": 2,
    "name": "Jake...",
    "transcript": "Last
time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

```
type: reduce
reduce_key: name
prompt: "Summarize the
common implicit biases shown
by this judge across all
their trial transcripts:
{{transcripts}}")
output:
  schema:
    common_biases: str
```

Output

```
[
  {
    "name": "Johnny...",
    "common_biases": ...
  },
  {
    "name": "Jake...",
    "common_biases": ...
  },
  ... 250 more ...
]
```


Semantic Reduce

Dataset



```
[
  {
    "id": 1,
    "name": "Johnny...",
    "transcript": "This
is the..."
  },
  {
    "id": 2,
    "name": "Jake...",
    "transcript": "Last
time we..."
  },
  ... 10,000 more ...
]
```

Pipeline (Single Operator)

```
type: reduce
reduce_key: name
prompt: "Summarize the
common implicit biases shown
by this judge across all
their trial transcripts:
{{transcripts}}")
output:
  schema:
    common_biases: str
```


Output



```
[
  {
    "name": "Johnny...",
    "common_biases": ...
  },
  {
    "name": "Jake...",
    "common_biases": ...
  },
  ... 250 more ...
]
```

Semantic Data Pipeline

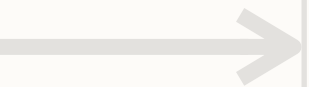
```
type: map  
prompt: "Determine the name of  
the judge in this trial"  
output:  
  schema:  
    name: str
```



```
type: reduce  
reduce_key: name  
prompt: "Summarize the common implicit  
biases shown by this judge across all  
their trial transcripts"  
output:  
  schema:  
    common_biases: str
```

Semantic Data Pipeline

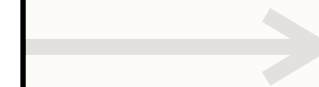
```
type: map
prompt: "Determine the name of
the judge in this trial"
output:
  schema:
    name: str
```



```
type: reduce
reduce_key: name
prompt: "Summarize the common implicit
biases shown by this judge across all
their trial transcripts"
output:
  schema:
    common_biases: str
```

Semantic Data Pipeline

```
type: map
prompt: "Determine the name of
the judge in this trial"
output:
  schema:
    name: str
```

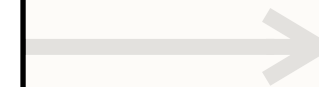


```
type: reduce
reduce_key: name
prompt: "Summarize the common implicit
biases shown by this judge across all
their trial transcripts"
output:
  schema:
    common_biases: str
```

id	transcript
...	...

Semantic Data Pipeline

```
type: map
prompt: "Determine the name of
the judge in this trial"
output:
  schema:
    name: str
```



```
type: reduce
reduce_key: name
prompt: "Summarize the common implicit
biases shown by this judge across all
their trial transcripts"
output:
  schema:
    common_biases: str
```

id	transcript	name
...

Semantic Data Pipeline

```
type: map  
prompt: "Determine the name of  
the judge in this trial"  
output:  
  schema:  
    name: str
```



```
type: reduce  
reduce_key: name  
prompt: "Summarize the common implicit  
biases shown by this judge across all  
their trial transcripts"  
output:  
  schema:  
    common_biases: str
```

id	transcript	name
...

Semantic Data Pipeline

```
type: map  
prompt: "Determine the name of  
the judge in this trial"  
output:  
  schema:  
    name: str
```



```
type: reduce  
reduce_key: name  
prompt: "Summarize the common implicit  
biases shown by this judge across all  
their trial transcripts"  
output:  
  schema:  
    common_biases: str
```

id	transcript	name
...

name
...

Semantic Data Pipeline

```
type: map
prompt: "Determine the name of
the judge in this trial"
output:
  schema:
    name: str
```



```
type: reduce
reduce_key: name
prompt: "Summarize the common implicit
biases shown by this judge across all
their trial transcripts"
output:
  schema:
    common_biases: str
```

id	transcript	name
...

name	common_biases
...	...

Semantic Data Pipeline

```
type: map  
prompt: "Determine the name of  
the judge in this trial"  
output:  
  schema:  
    name: str
```



```
type: reduce  
reduce_key: name  
prompt: "Summarize the common implicit  
biases shown by this judge across all  
their trial transcripts"  
output:  
  schema:  
    common_biases: str
```

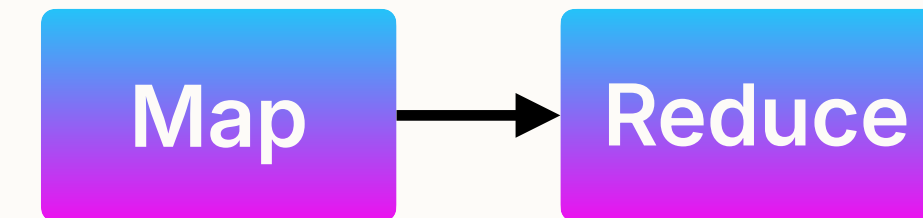
id	transcript	name
...

name	common_biases
...	...

Optimizing Semantic Data Pipelines

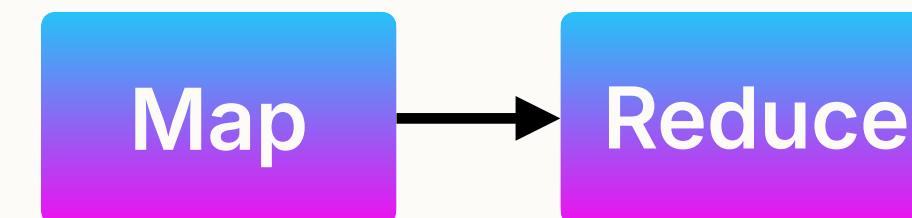
Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline



Optimizing Semantic Data Pipelines

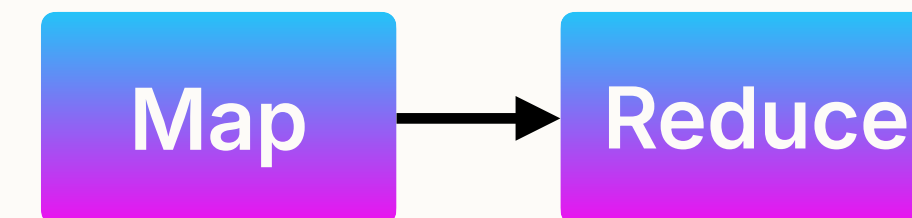
Interface: Semantic operator pipeline



Query Optimizers:

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

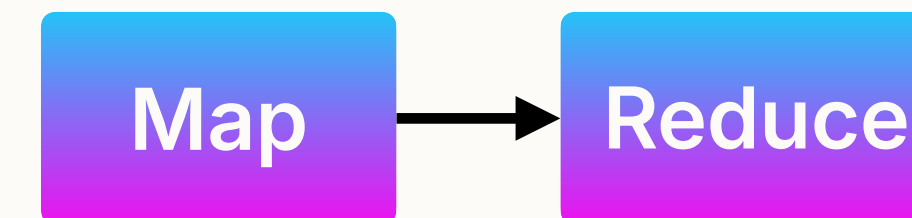


Query Optimizers:

1. Define the plan space

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

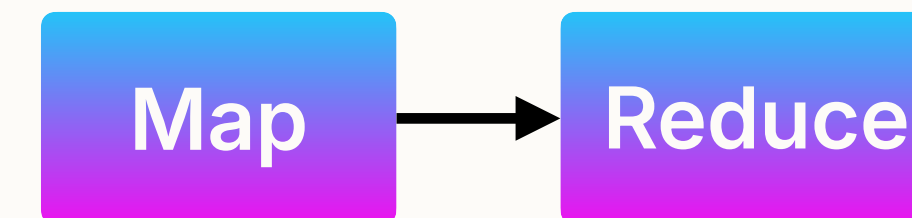


Query Optimizers:

1. Define the plan space
2. Estimate costs

Optimizing Semantic Data Pipelines

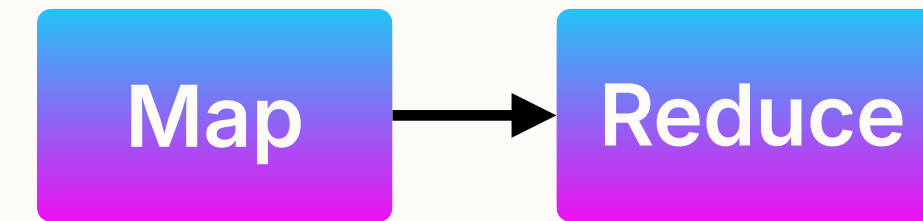
Interface: Semantic operator pipeline



Query Optimizers:

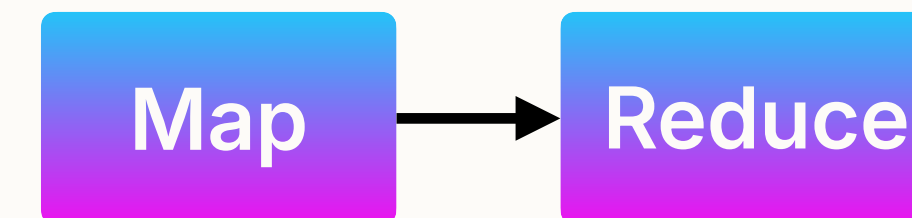
1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Optimizing Semantic Data Pipelines



Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

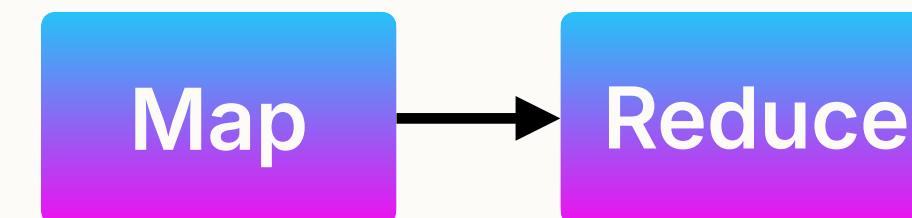


Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline



Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Prompting
strategies

Different
LLMs

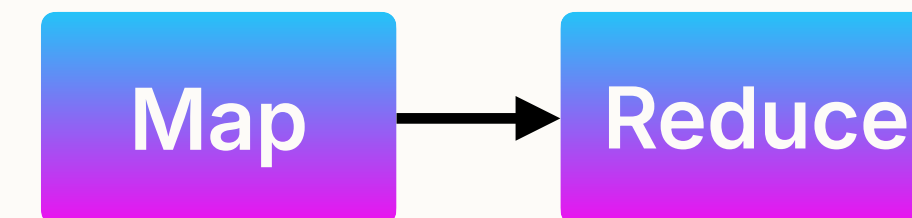
Model
ensembles

Document
pruning

...

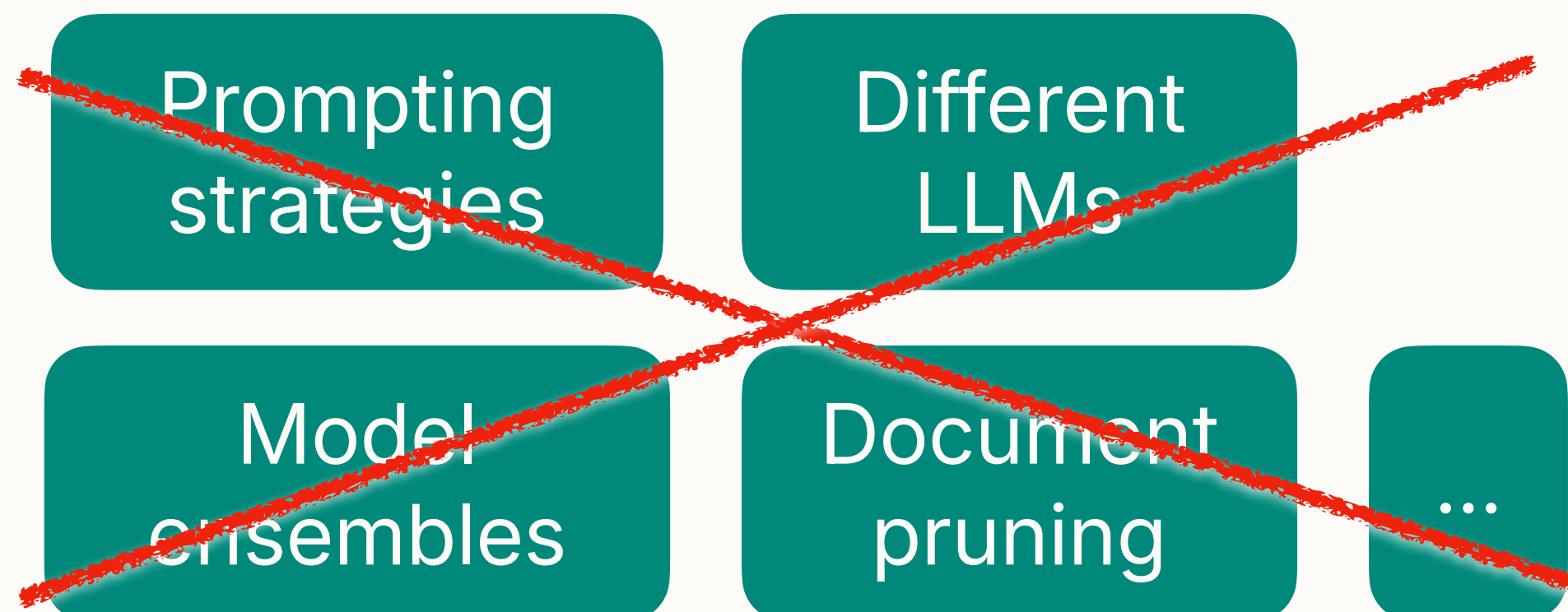
Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline



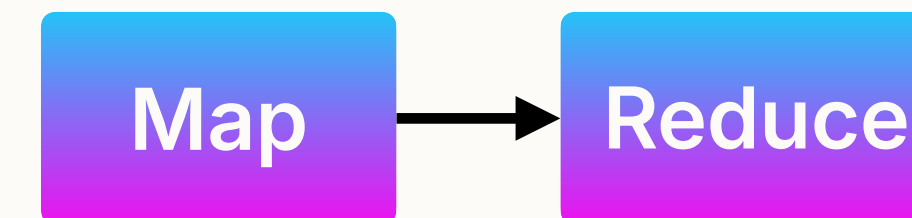
Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



Optimizing Semantic Data Pipelines

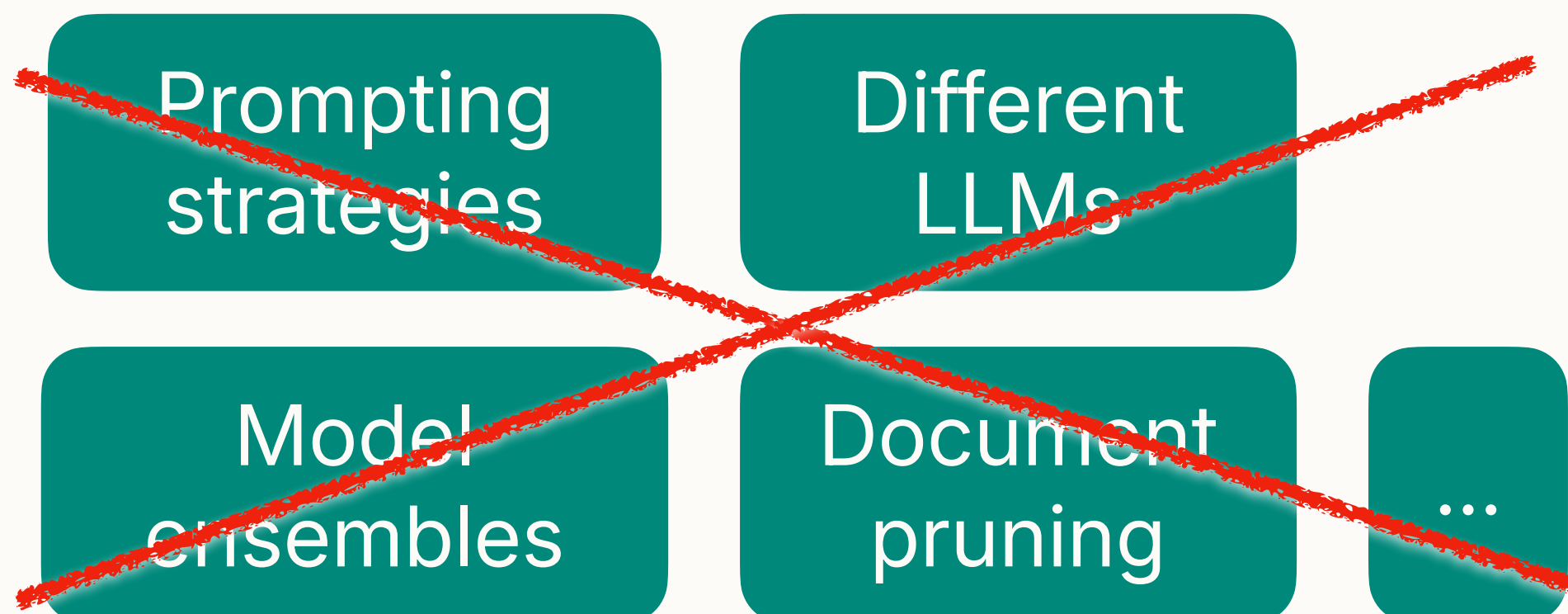
Interface: Semantic operator pipeline



Query Optimizers:

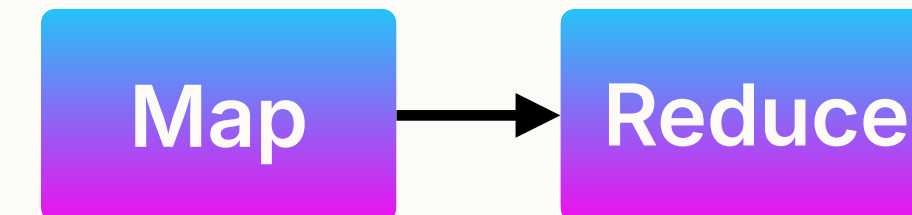
1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Errors:



Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

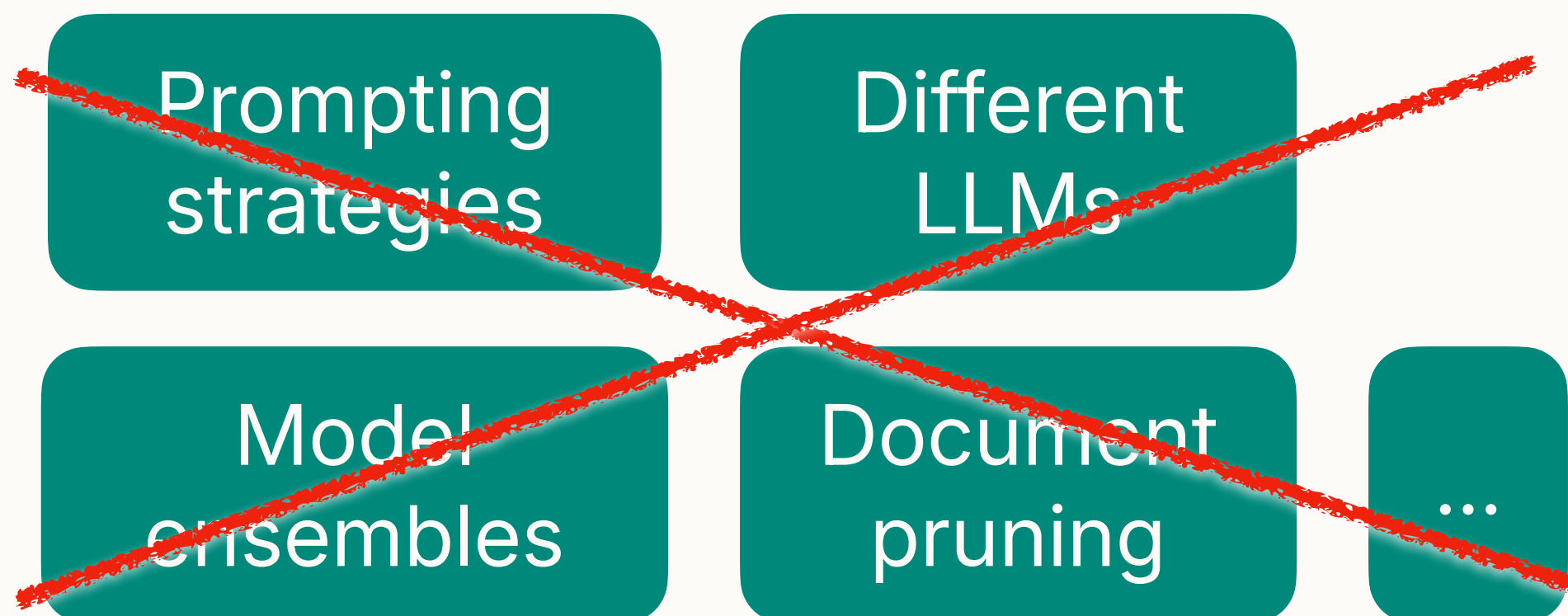


Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

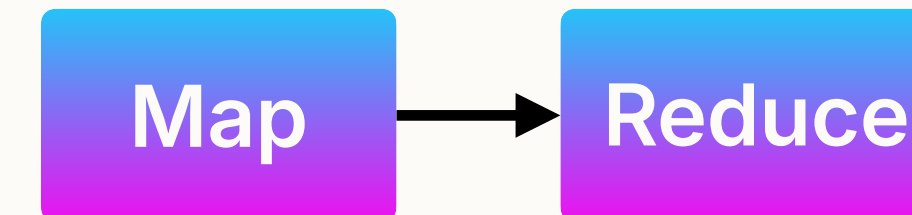
Errors:

* Outlier errors



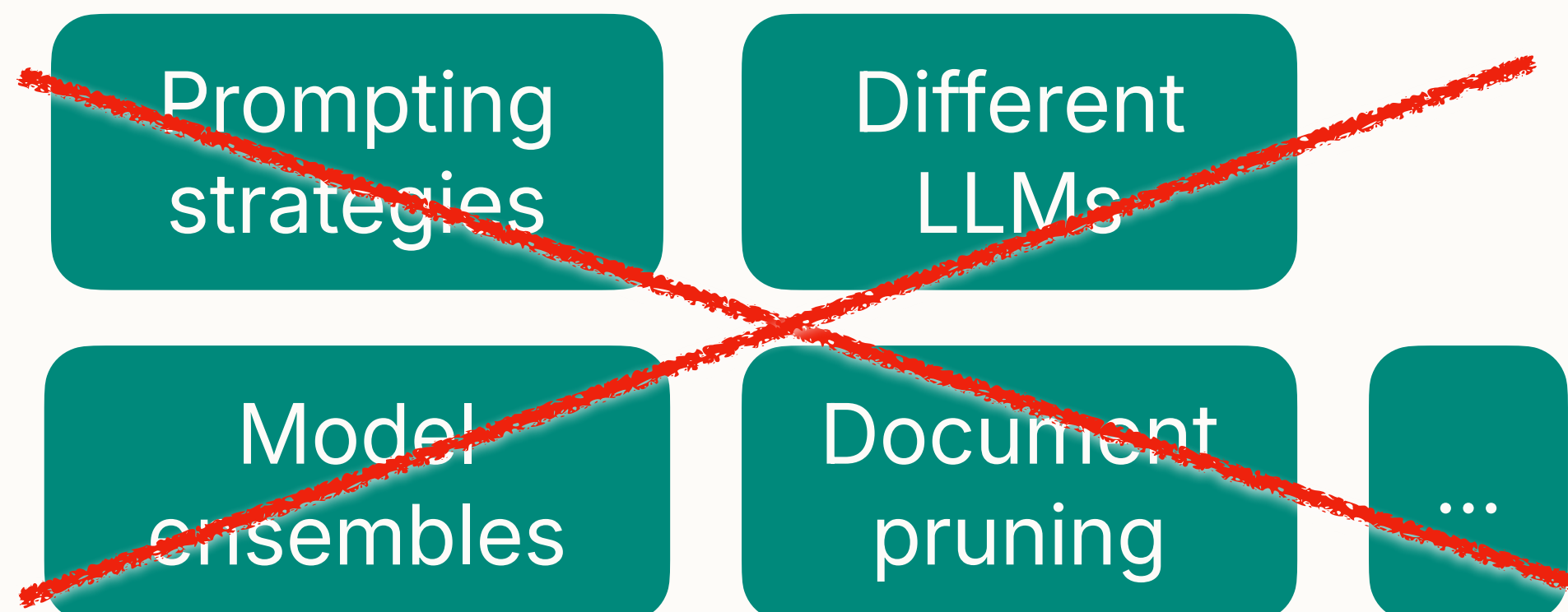
Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline



Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

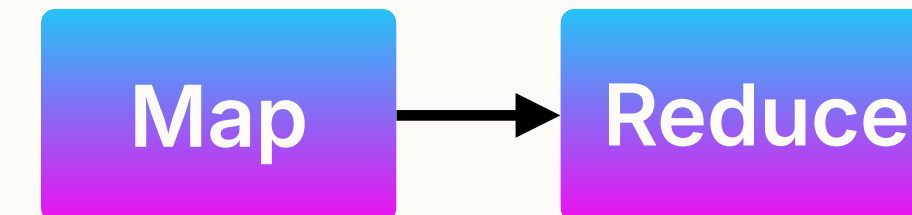


Errors:

- * Outlier errors
 - ◆ *Fix: validation & retries* ✓

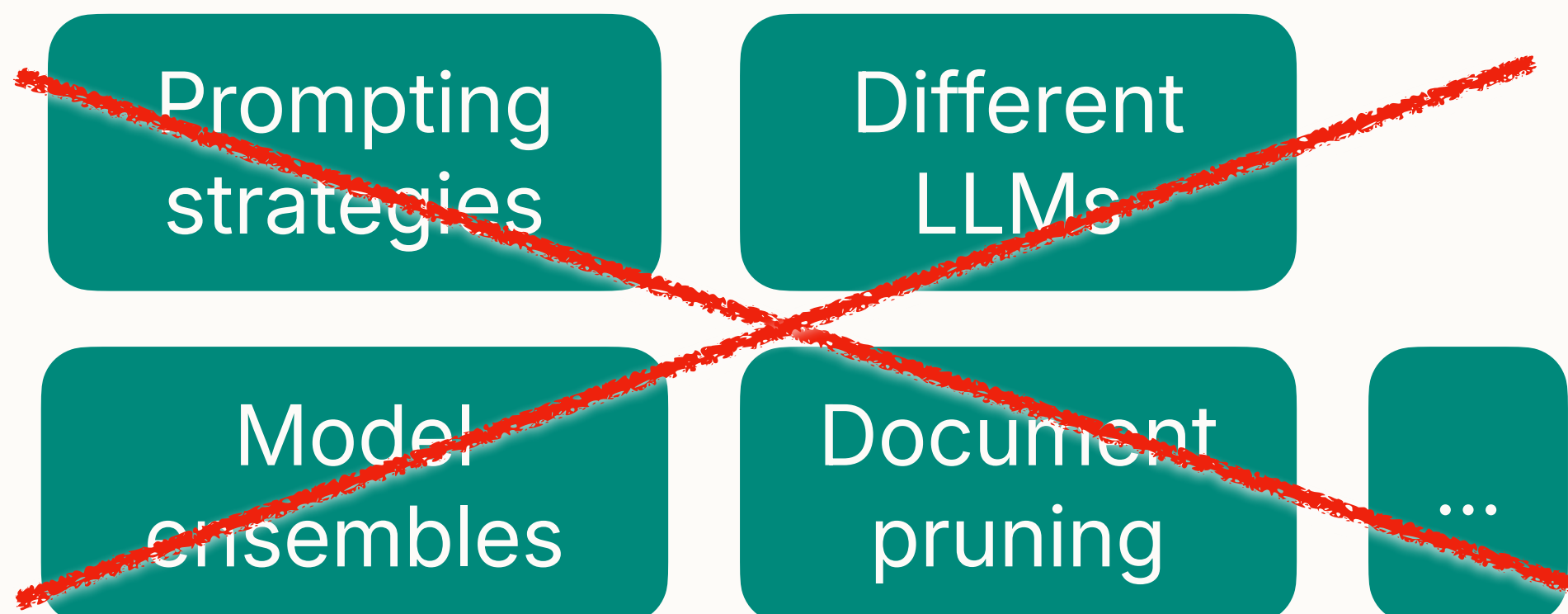
Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline



Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



Errors:

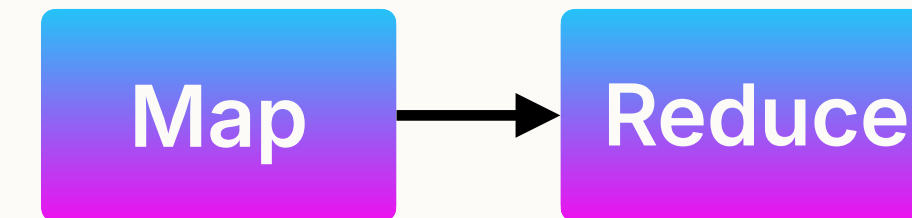
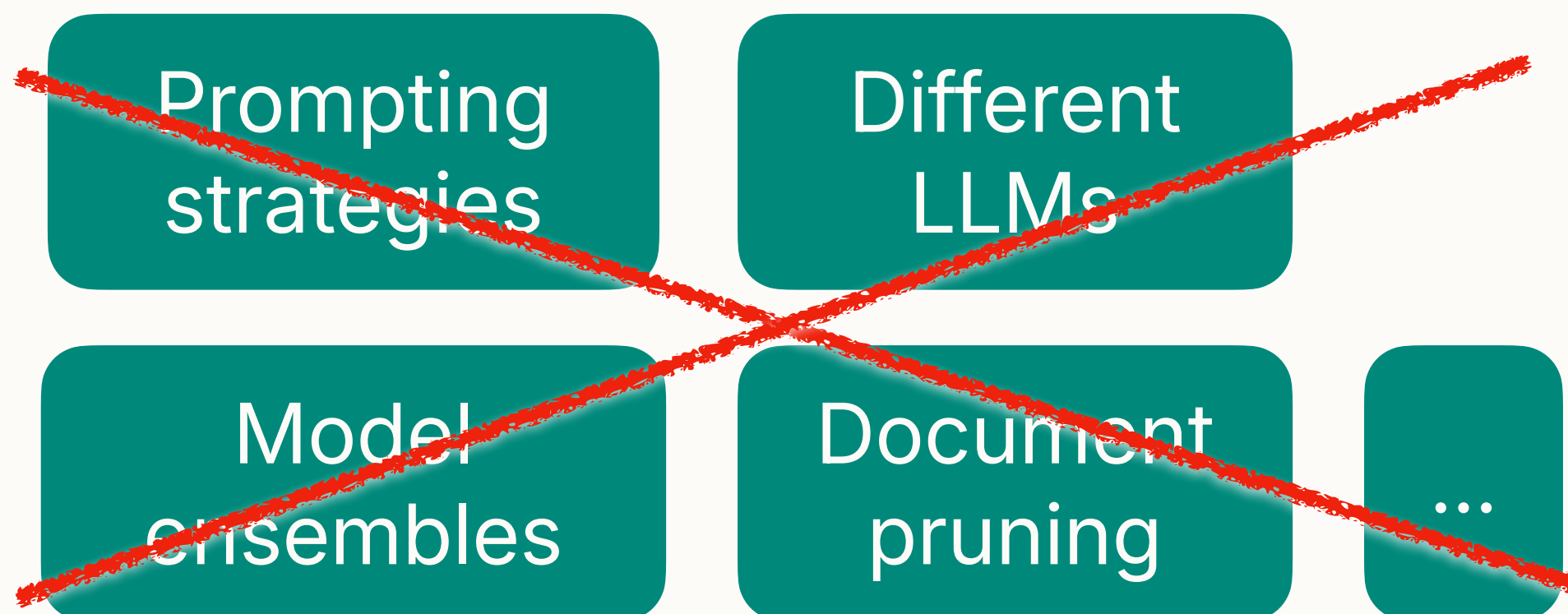
- * Outlier errors
 - ◆ *Fix: validation & retries* ✓
- * Poorly-scoped semantic operators

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



Summarize this, based on A and B and C and D and...

Errors:

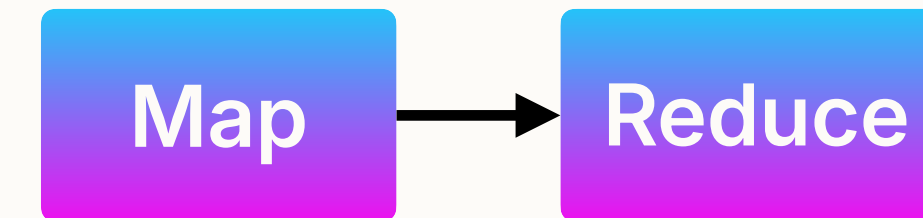
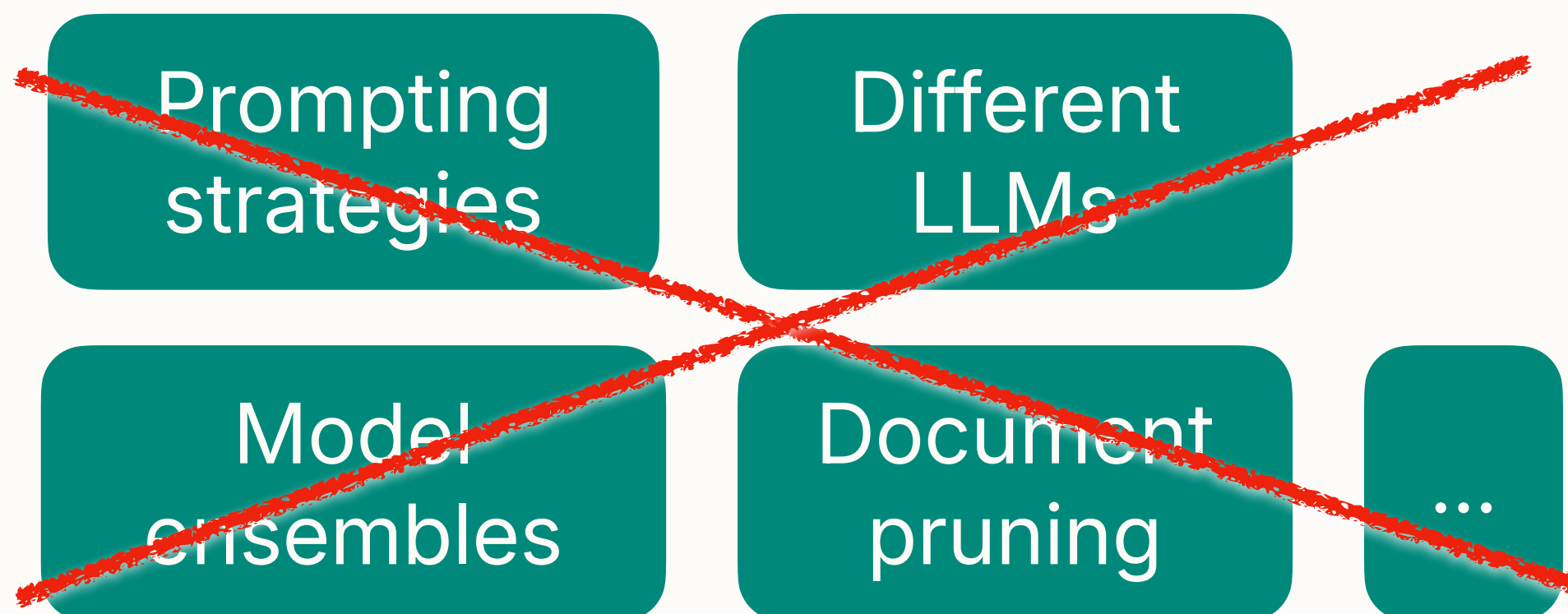
- * Outlier errors
 - ◆ *Fix: validation & retries* ✓
- * Poorly-scoped semantic operators

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



LLM can't reason about all things!

Summarize this, based on A and B and C and D and...

Errors:

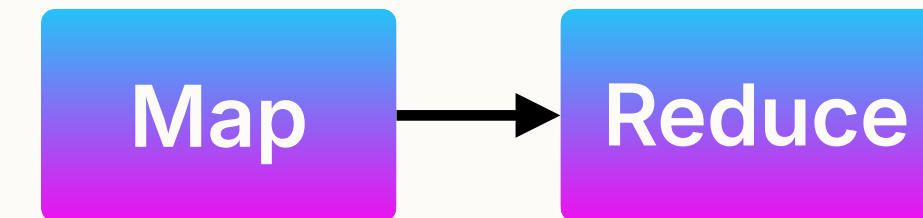
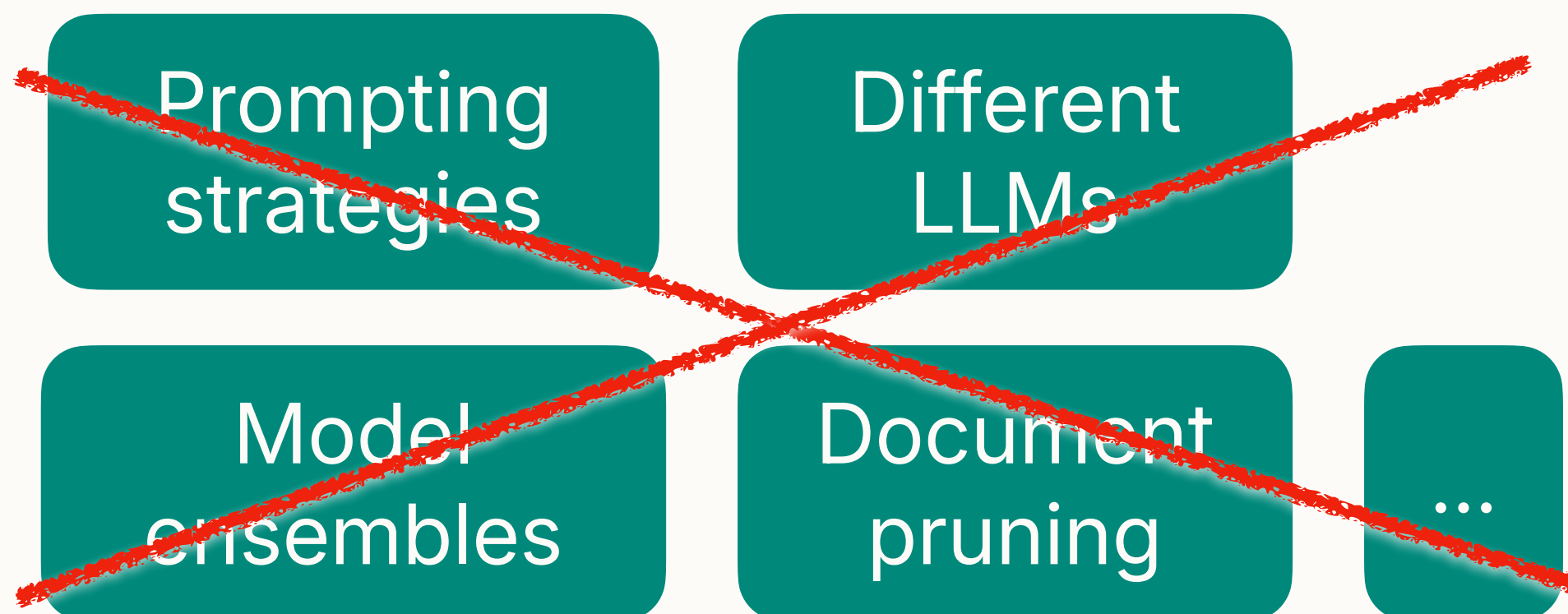
- * Outlier errors
 - ◆ *Fix: validation & retries* ✓
- * Poorly-scoped semantic operators

Optimizing Semantic Data Pipelines

Interface: Semantic operator pipeline

Query Optimizers:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



LLM can't reason about all things!

Summarize this, based on A and B and C and D and...

Errors:

- * Outlier errors
 - ◆ *Fix: validation & retries* ✓
- * Poorly-scoped semantic operators
 - ◆ *Fix: ??*

Inspiration: Query Rewrites

Inspiration: Query Rewrites

SQL query rewrite engines apply transformation rules. E.g.,

- ◆ Selection/projection pushdown
- ◆ Fold constants (e.g., `salary > 10000 + 5000` → `salary > 15000`)
- ◆ Join associativity/commutativity
- ◆ Rewrite subqueries into semi-joins

Inspiration: Query Rewrites

SQL query rewrite engines apply transformation rules. E.g.,

- ◆ Selection/projection pushdown
- ◆ Fold constants (e.g., `salary > 10000 + 5000` → `salary > 15000`)
- ◆ Join associativity/commutativity
- ◆ Rewrite subqueries into semi-joins

Rules are rigid; legal or legal, no ambiguity.

Inspiration: Query Rewrites

SQL query rewrite engines apply transformation rules. E.g.,

- ◆ Selection/projection pushdown
- ◆ Fold constants (e.g., `salary > 10000 + 5000` → `salary > 15000`)
- ◆ Join associativity/commutativity
- ◆ Rewrite subqueries into semi-joins

Rules are rigid; legal or legal, no ambiguity.

Can a rule-based engine optimize semantic data processing?

Inspiration: Query Rewrites

SQL query rewrite engines apply transformation rules. E.g.,

- ♦ Selection/projection pushdown
- ♦ Fold constants (e.g., `salary > 10000 + 5000` → `salary > 15000`)
- ♦ Join associativity/commutativity
- ♦ Rewrite subqueries into semi-joins

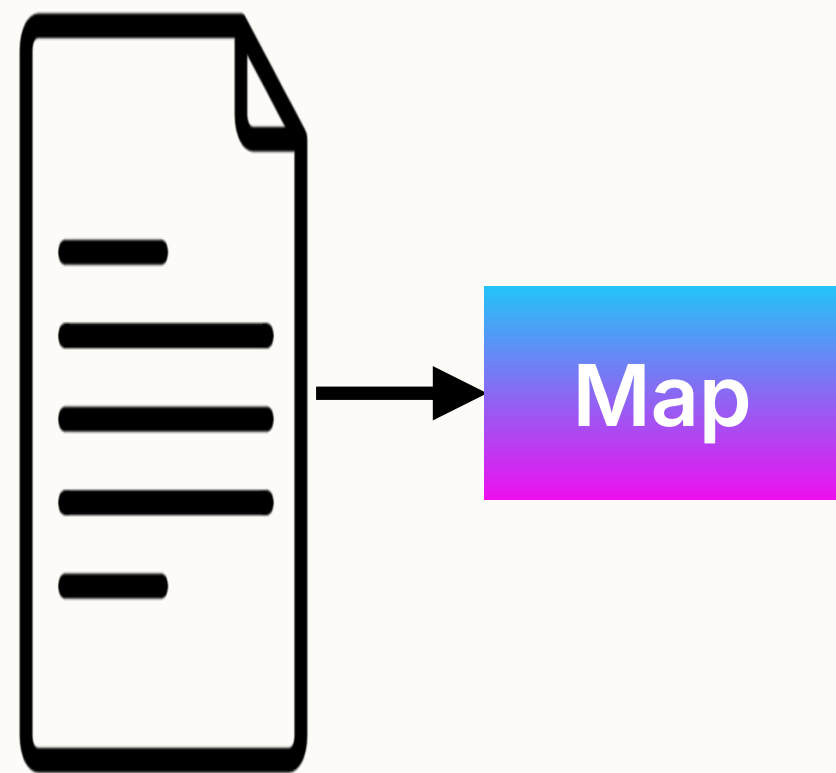
Rules are rigid; legal or legal, no ambiguity.

Can a rule-based engine optimize semantic data processing?

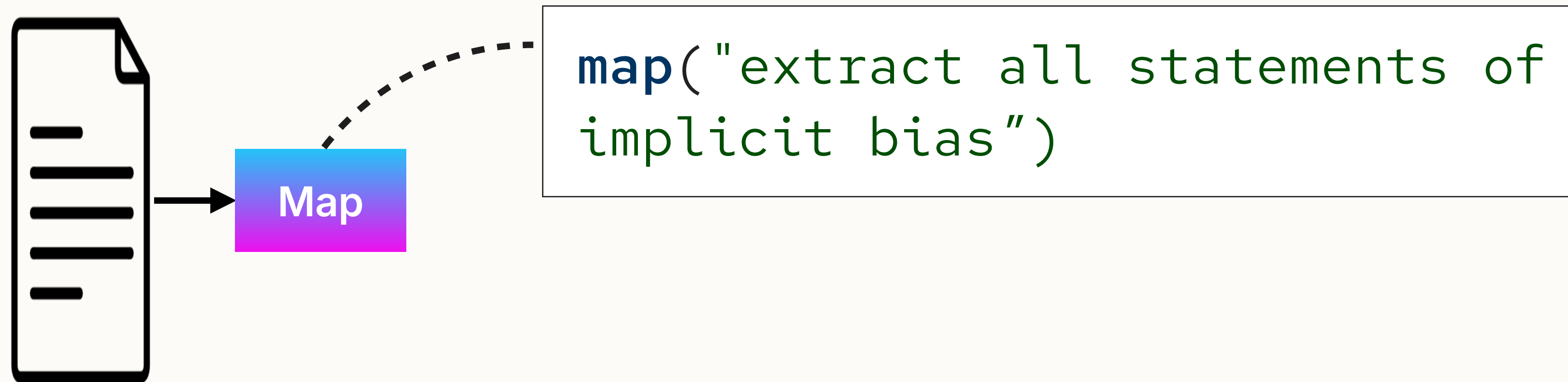
Challenge: Semantic operators are fuzzy!

Key Insight: Systematic Decomposition

Key Insight: Systematic Decomposition



Key Insight: Systematic Decomposition

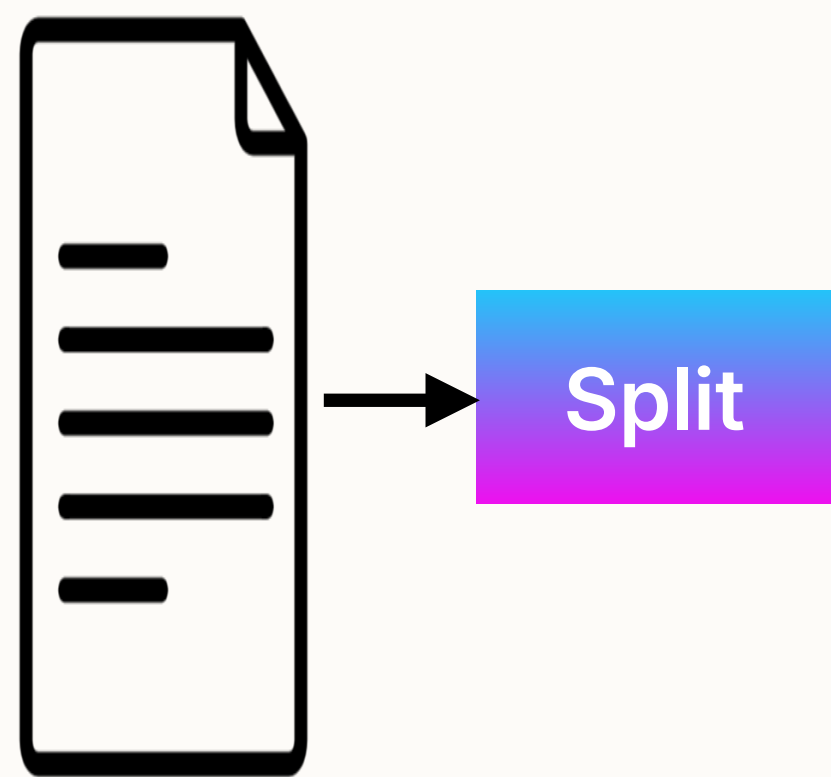


Key Insight: Systematic Decomposition

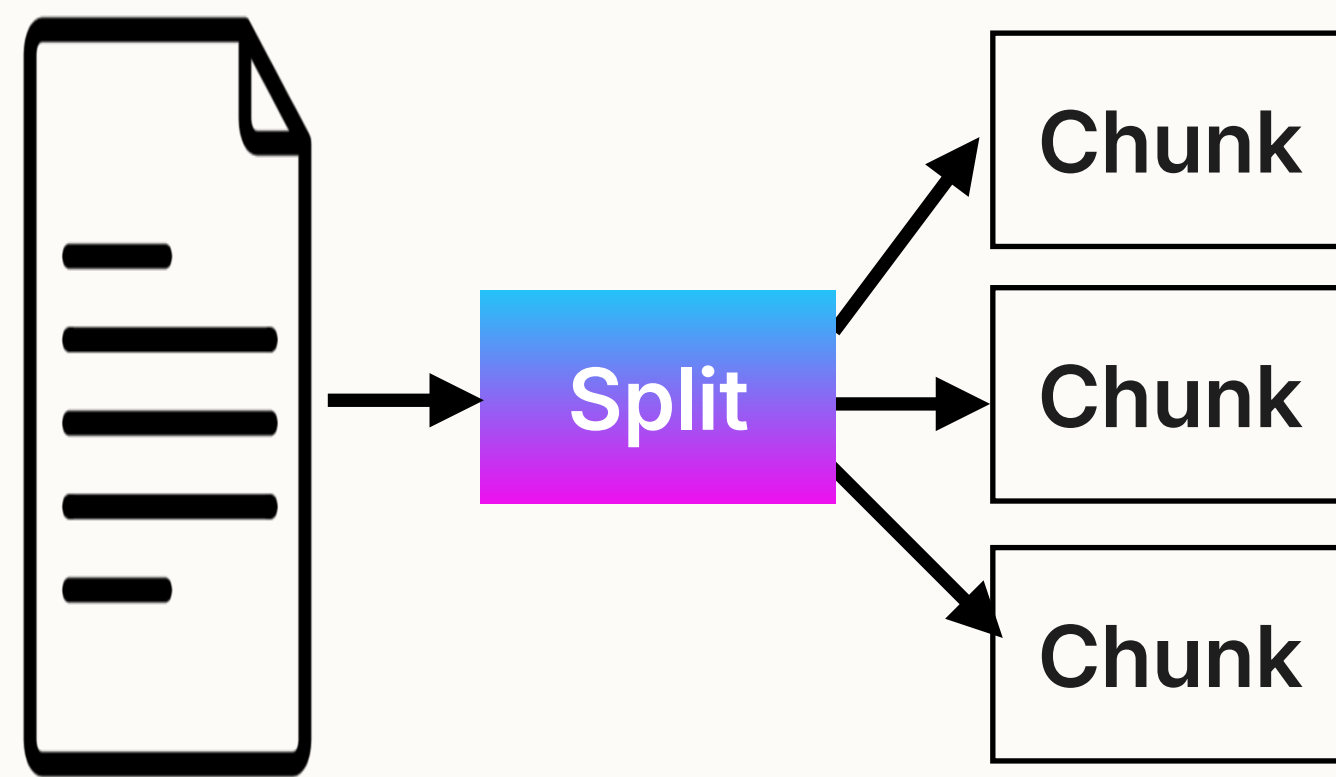
Key Insight: Systematic Decomposition



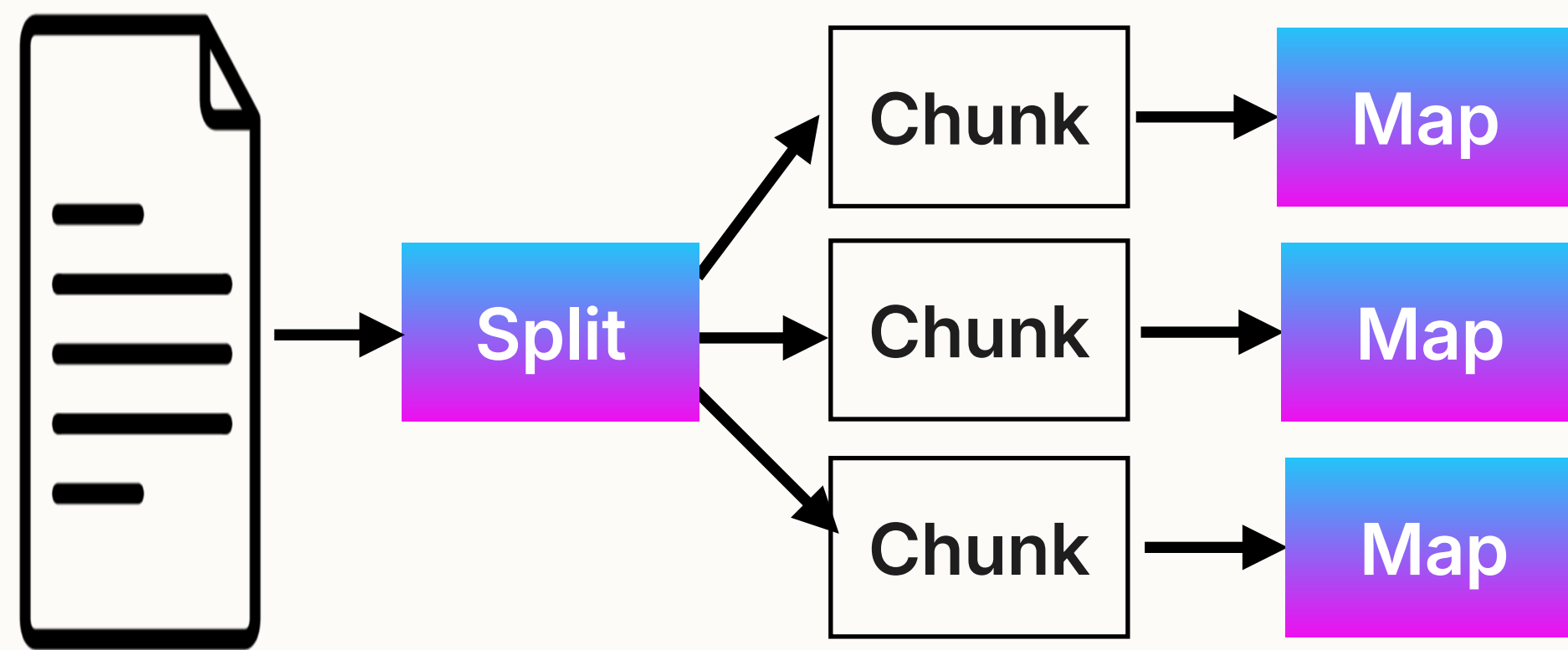
Key Insight: Systematic Decomposition



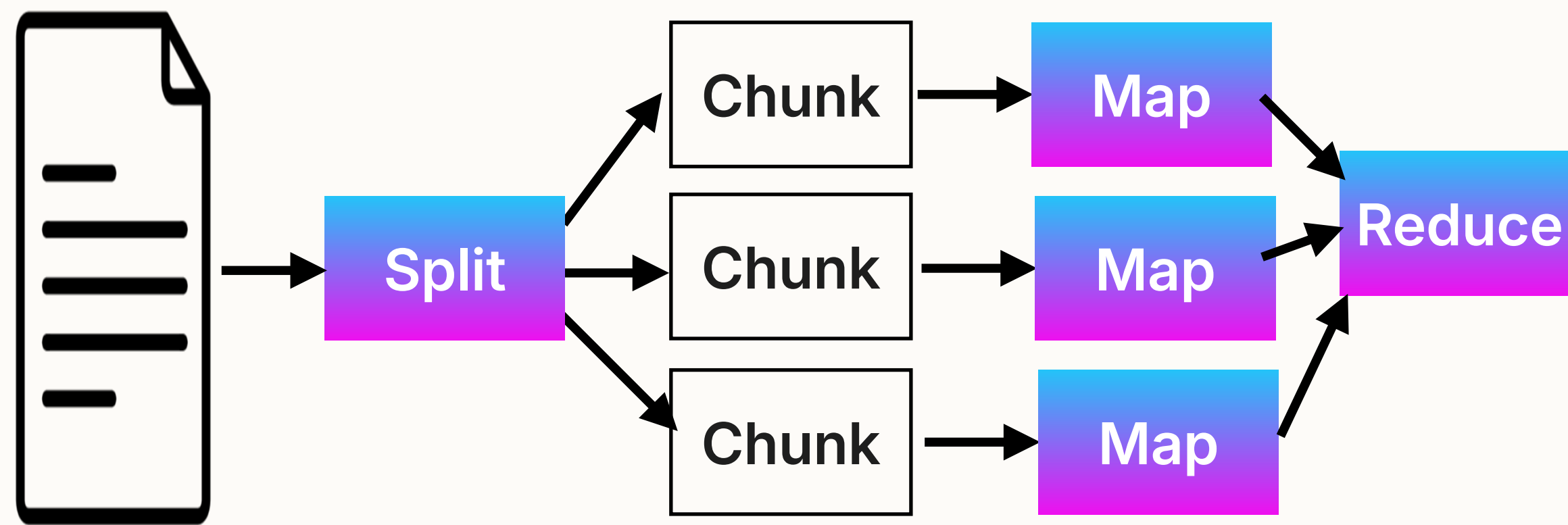
Key Insight: Systematic Decomposition



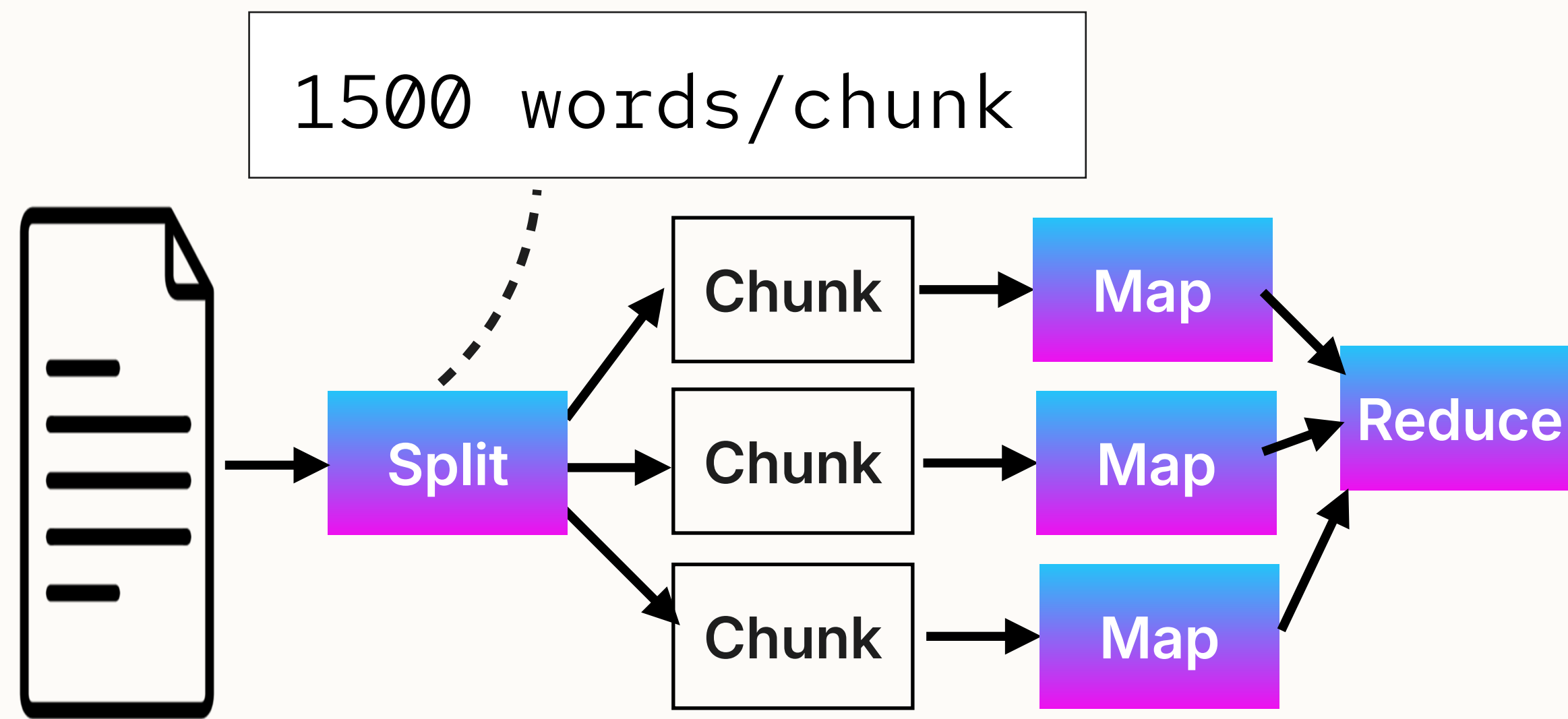
Key Insight: Systematic Decomposition



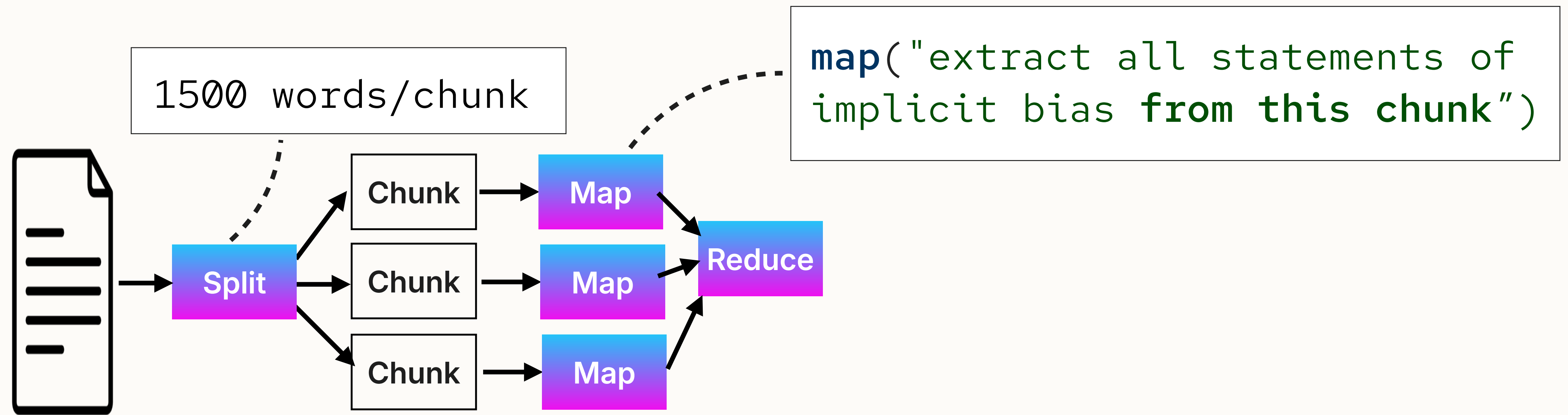
Key Insight: Systematic Decomposition



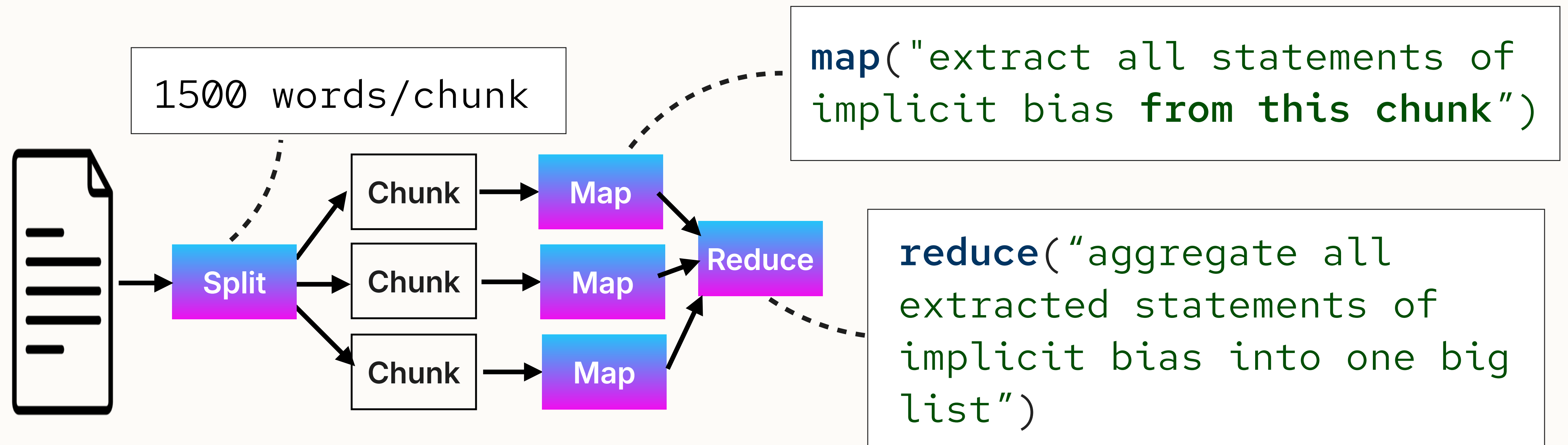
Key Insight: Systematic Decomposition



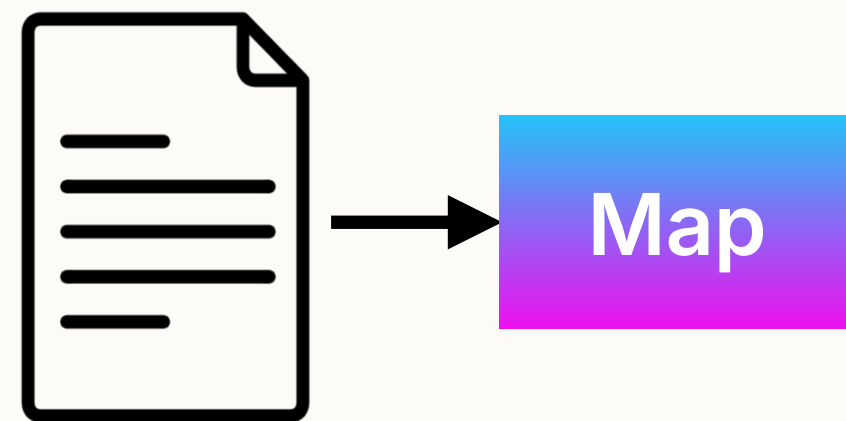
Key Insight: Systematic Decomposition



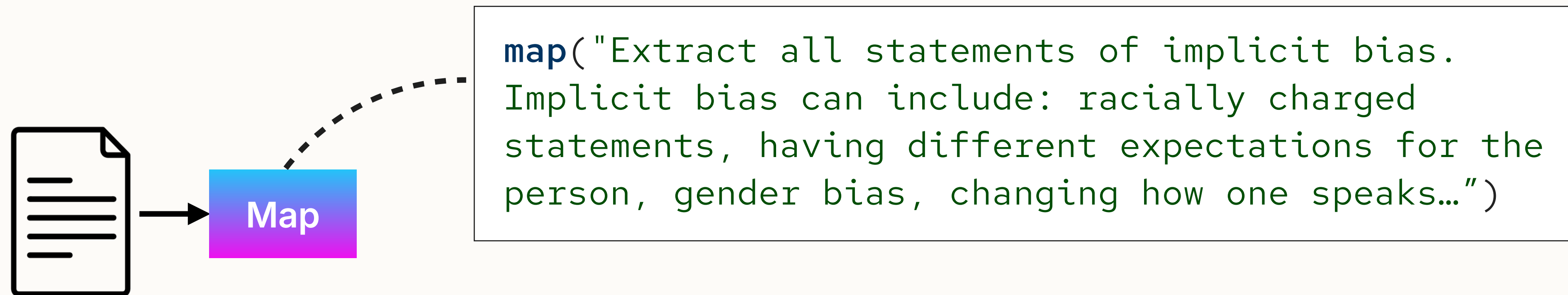
Key Insight: Systematic Decomposition



More Decomposition



More Decomposition

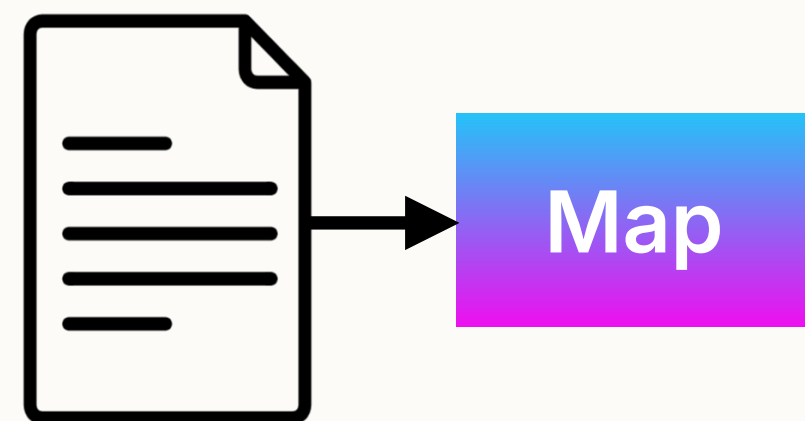


More Decomposition

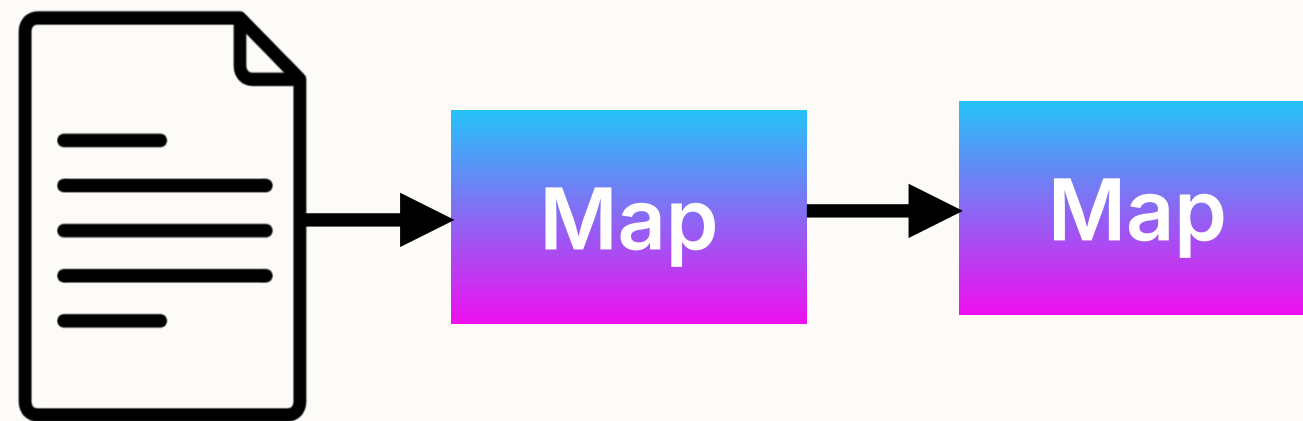
More Decomposition



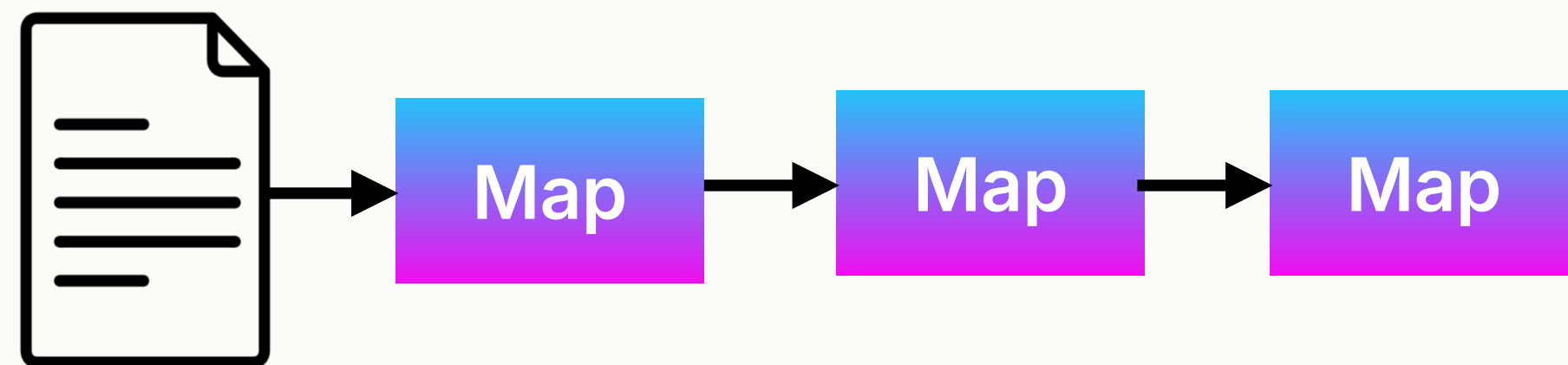
More Decomposition



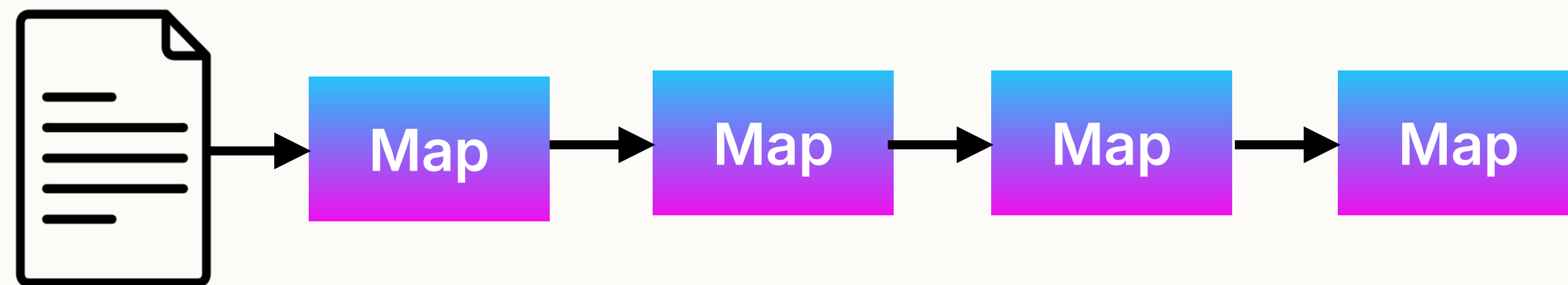
More Decomposition



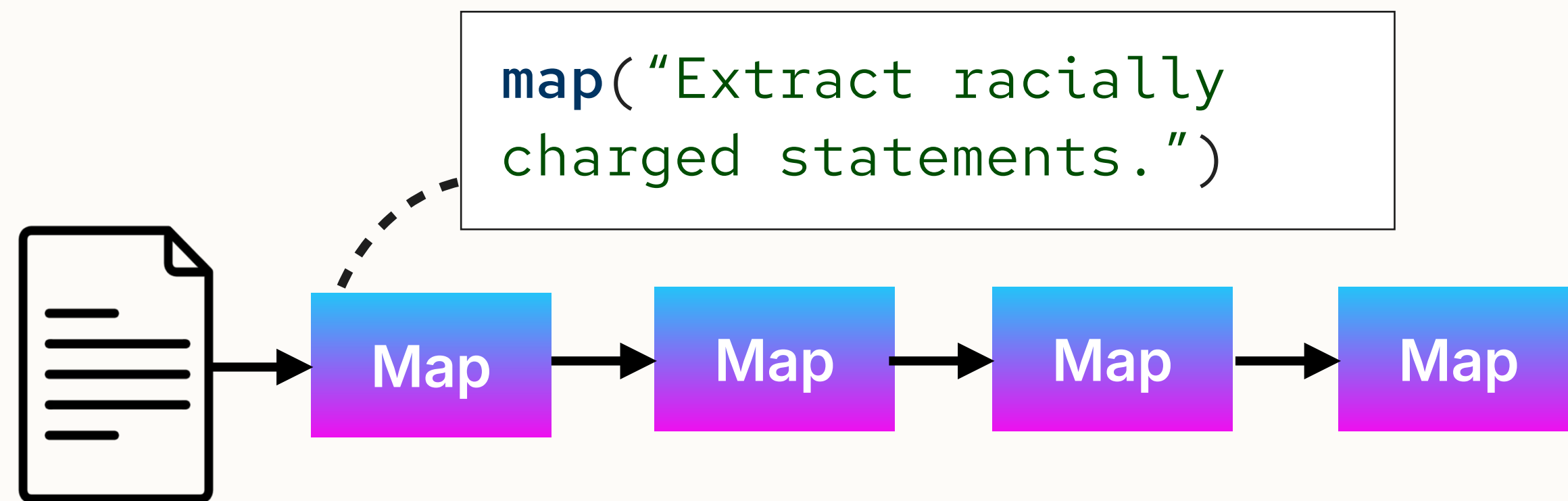
More Decomposition



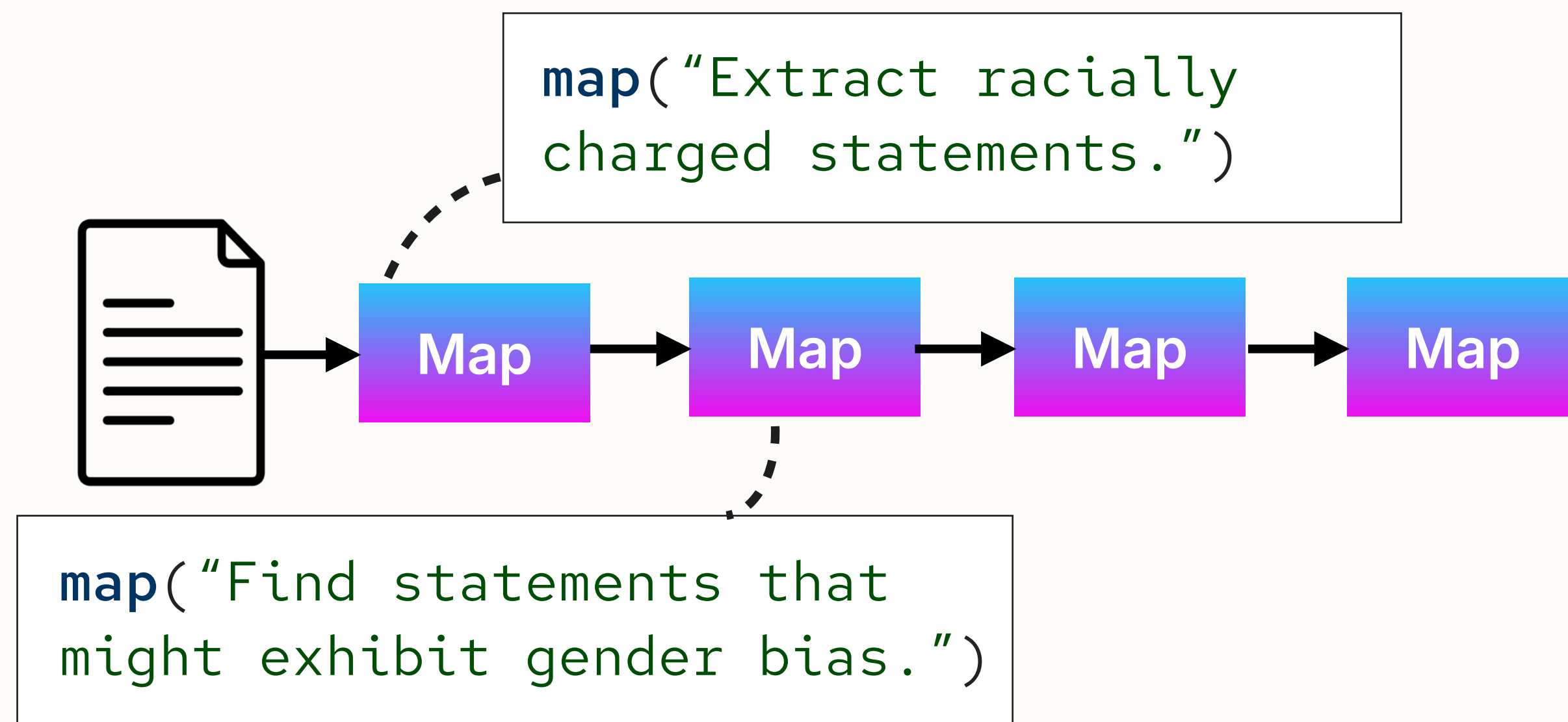
More Decomposition



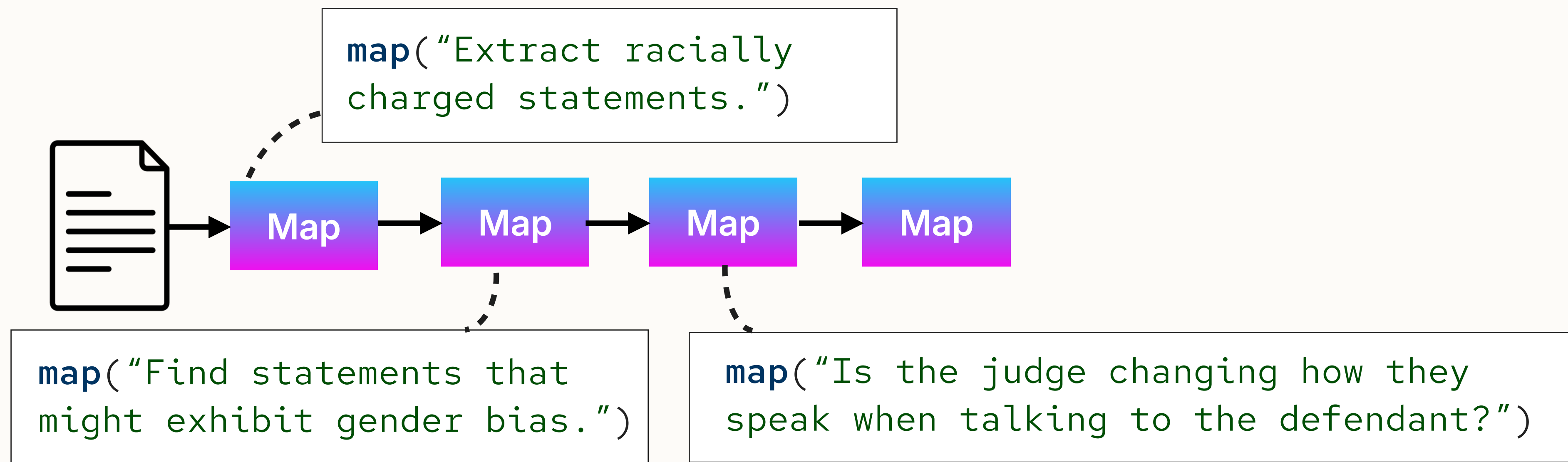
More Decomposition



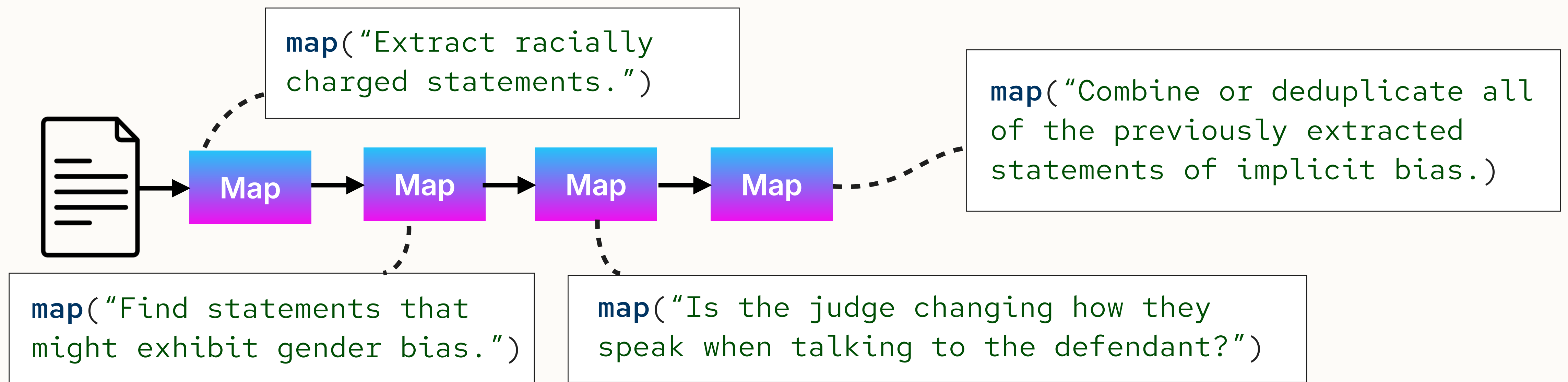
More Decomposition



More Decomposition



More Decomposition



Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. [VLDB '25](#).

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. [VLDB '25](#).

Idea: Templates that describe how to rewrite a subsequence of operators.

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

♦ `map` \Rightarrow `split` \rightarrow `map` \rightarrow `reduce`

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

◆ `map` \Rightarrow `split` \rightarrow `map` \rightarrow `reduce`

◆ `map` \Rightarrow `map+`

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

◆ `map` \Rightarrow `split` \rightarrow `map` \rightarrow `reduce`

◆ `map` \Rightarrow `map+`

◆ `op` \Rightarrow `map` \rightarrow `op`

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

◆ $\text{map} \Rightarrow \text{split} \rightarrow \text{map} \rightarrow \text{reduce}$

◆ $\text{map} \Rightarrow \text{map}^+$

◆ $\text{op} \Rightarrow \text{map} \rightarrow \text{op}$

What they do:

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

- ◆ $\text{map} \Rightarrow \text{split} \rightarrow \text{map} \rightarrow \text{reduce}$
- ◆ $\text{map} \Rightarrow \text{map}^+$
- ◆ $\text{op} \Rightarrow \text{map} \rightarrow \text{op}$

What they do:

- ◆ Encode reusable transformation patterns

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

- ◆ $\text{map} \Rightarrow \text{split} \rightarrow \text{map} \rightarrow \text{reduce}$
- ◆ $\text{map} \Rightarrow \text{map}^+$
- ◆ $\text{op} \Rightarrow \text{map} \rightarrow \text{op}$

What they do:

- ◆ Encode reusable transformation patterns
- ◆ Instantiated by LLM agents

Solution: Rewrite Directives

DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. **Shankar** et al. *VLDB '25*.

Idea: Templates that describe how to rewrite a subsequence of operators.

Examples:

- ◆ `map` \Rightarrow `split` \rightarrow `map` \rightarrow `reduce`
- ◆ `map` \Rightarrow `map+`
- ◆ `op` \Rightarrow `map` \rightarrow `op`

What they do:

- ◆ Encode reusable transformation patterns
- ◆ Instantiated by LLM agents
- ◆ Require new operator types (e.g., `split`, `gather`, `resolve`)

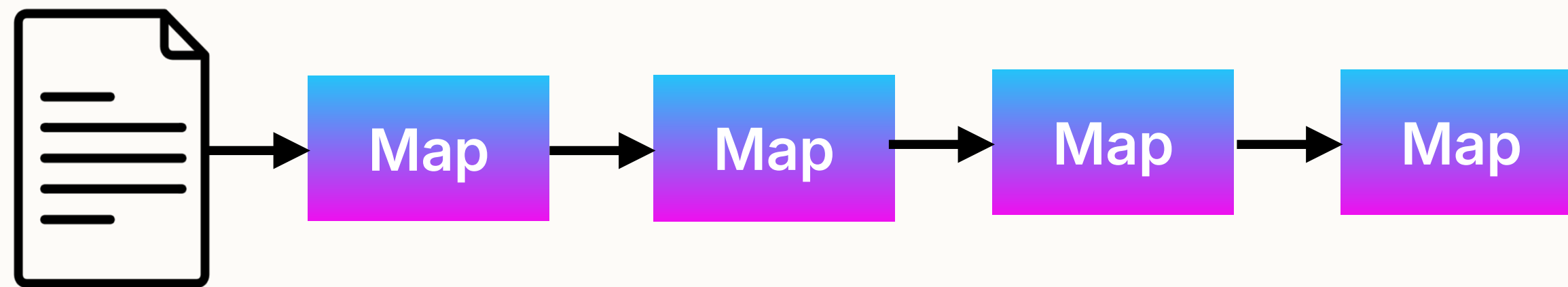
Cost-Oriented Rewrite Directives

Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$

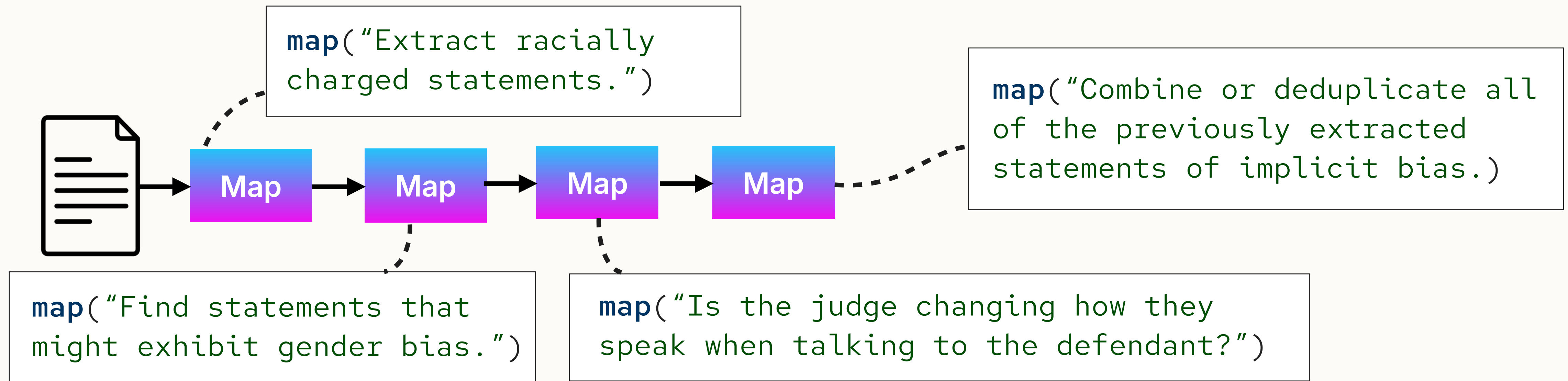
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



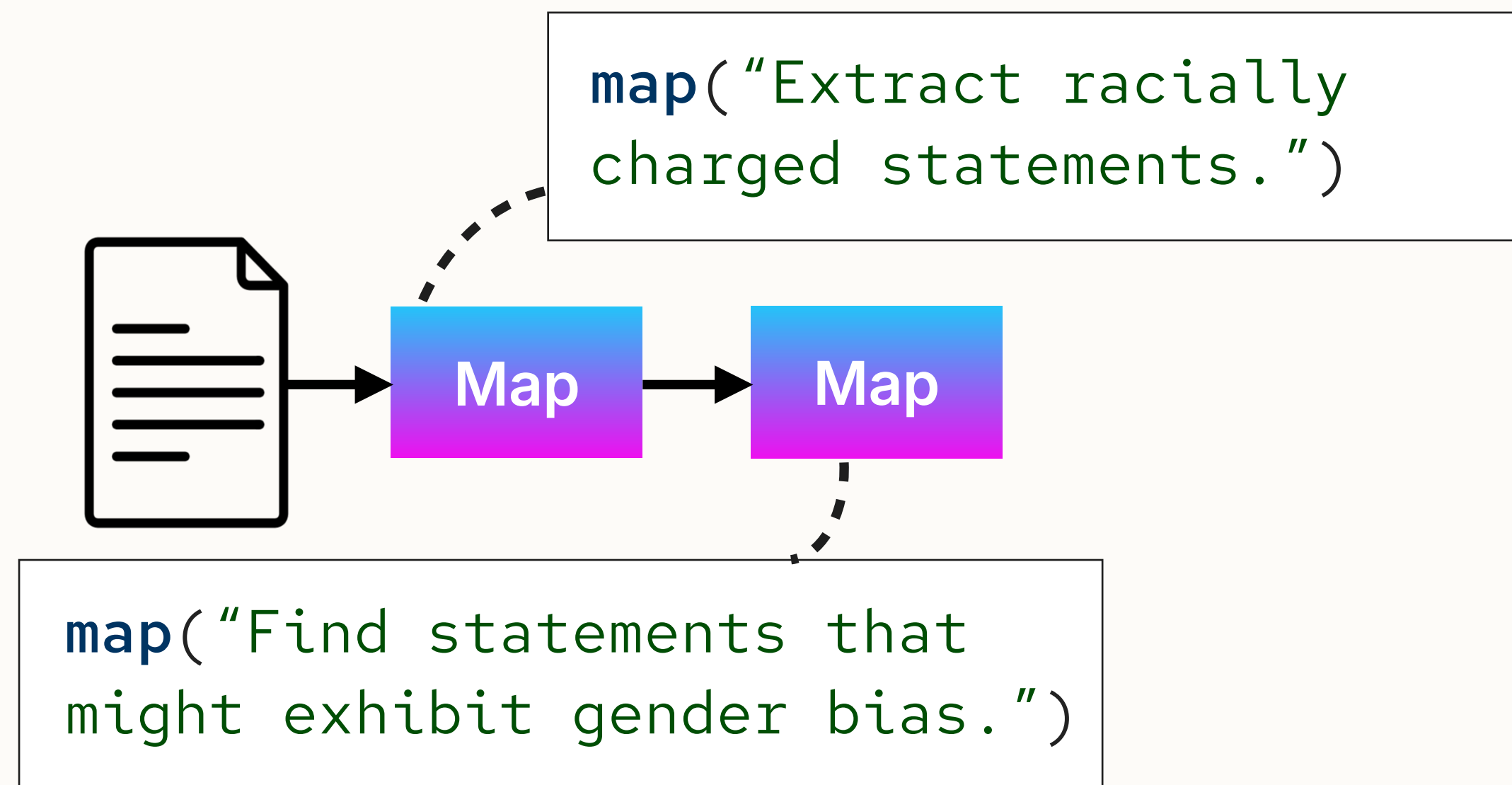
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



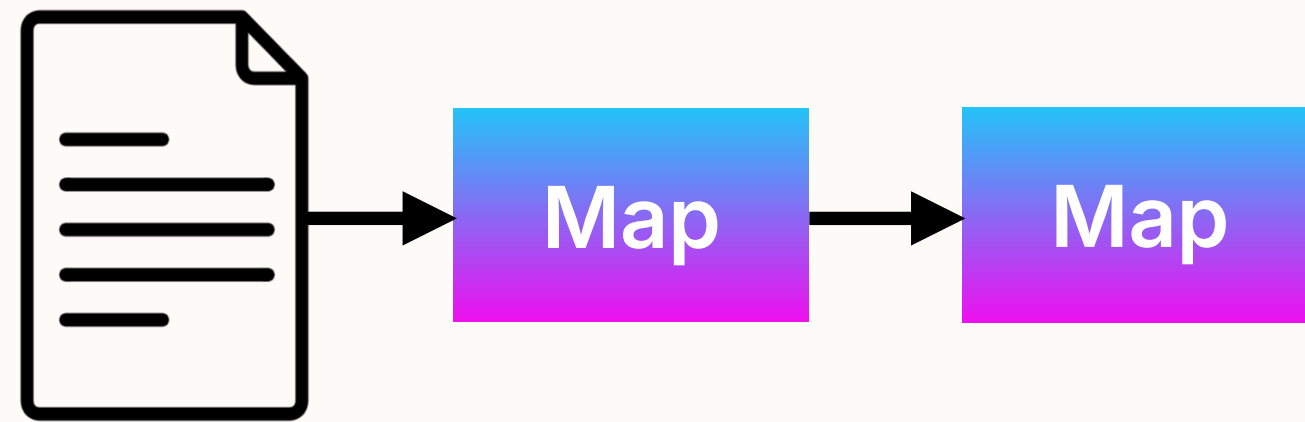
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



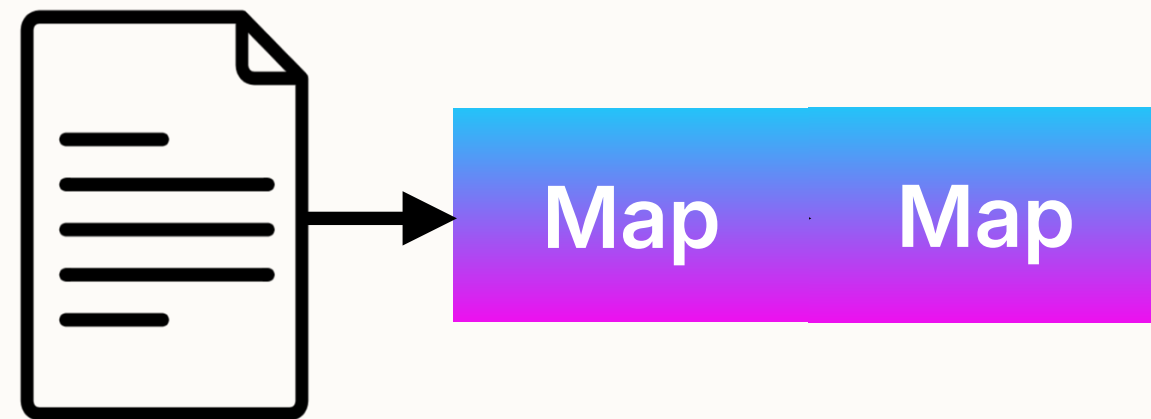
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



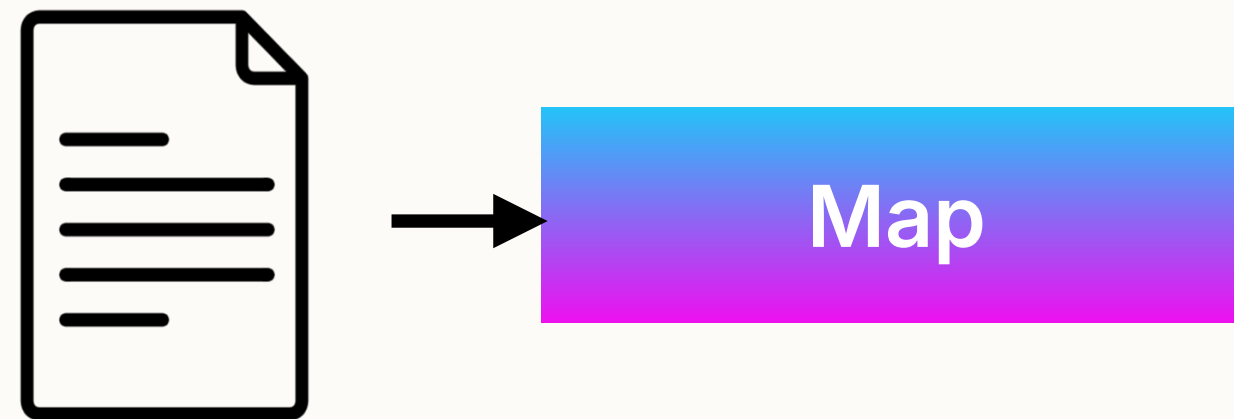
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



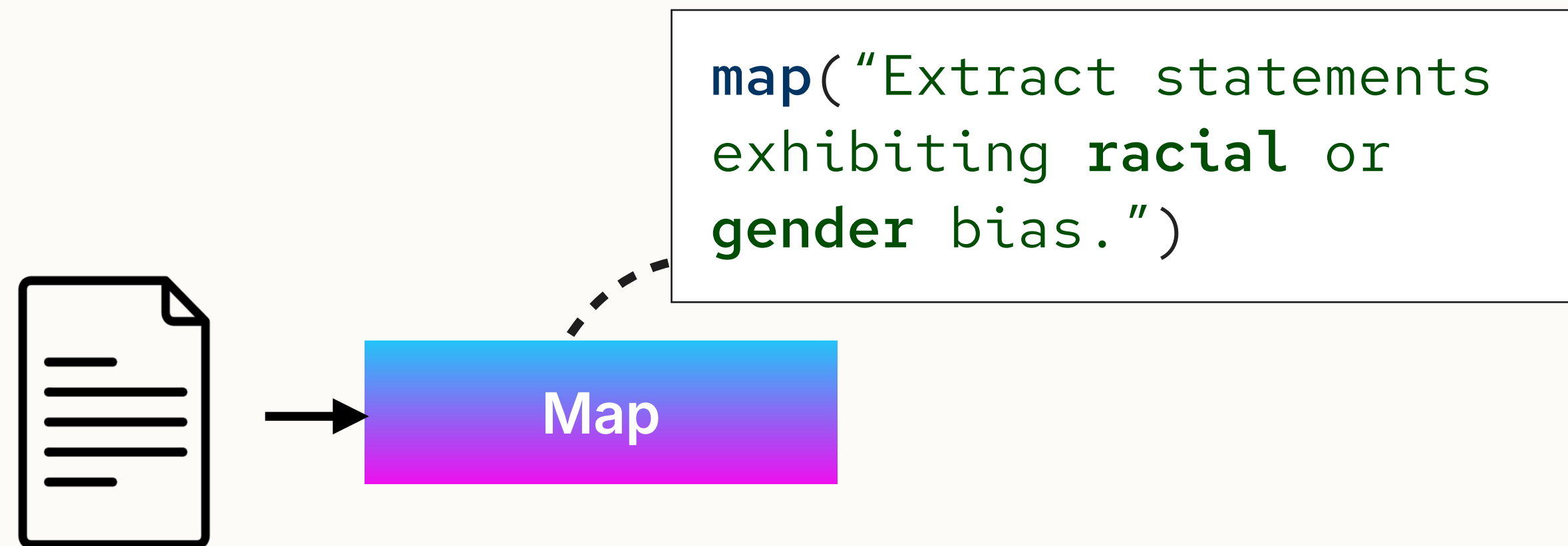
Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



Cost-Oriented Rewrite Directives

Operator fusion: $op \rightarrow op \Rightarrow op$



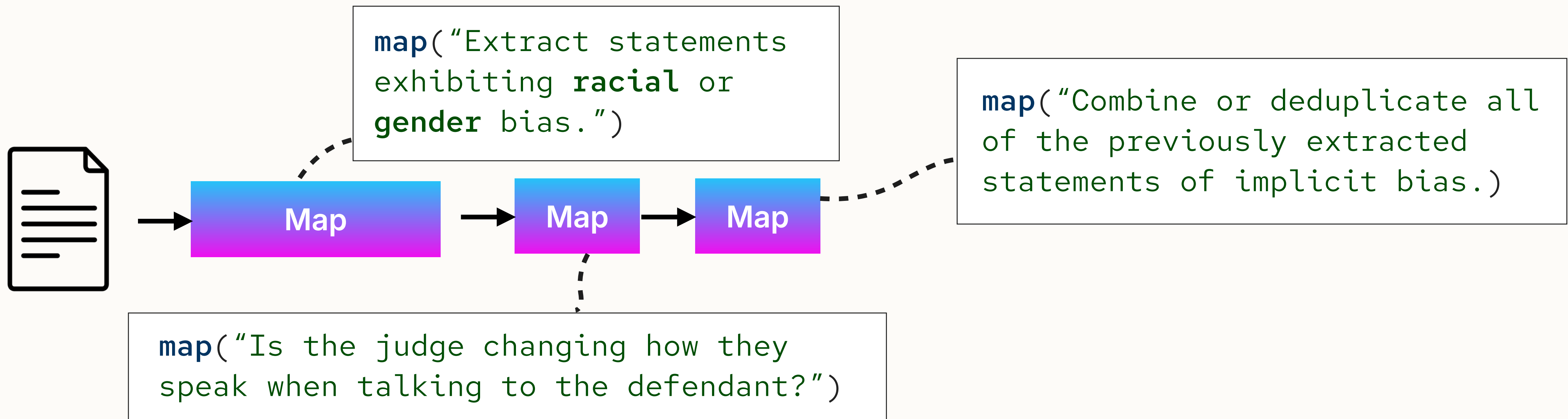
Cost-Oriented Rewrite Directives

Cost-Oriented Rewrite Directives

Replace with Code: `op` \Rightarrow `code_op`

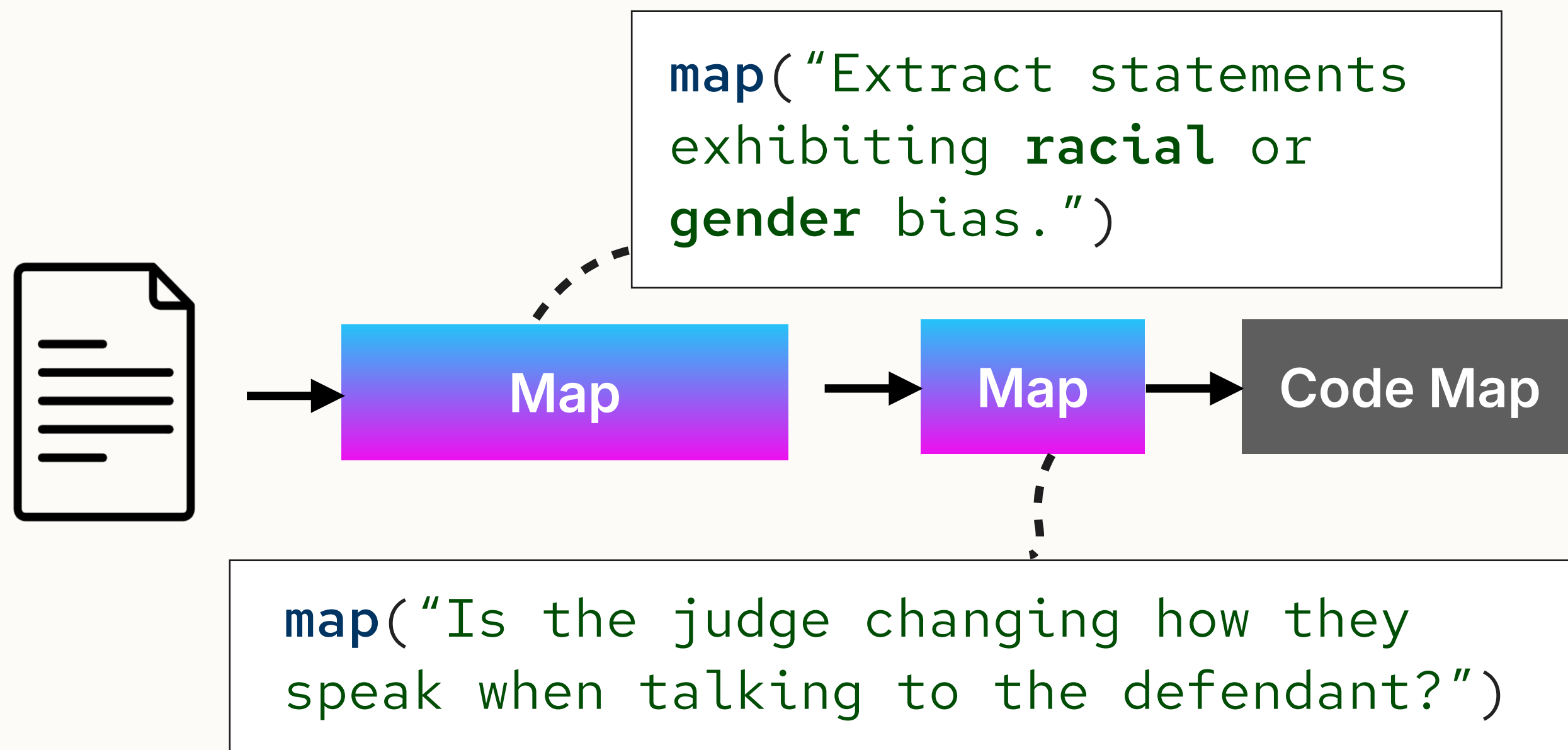
Cost-Oriented Rewrite Directives

Replace with Code: $op \Rightarrow code_op$



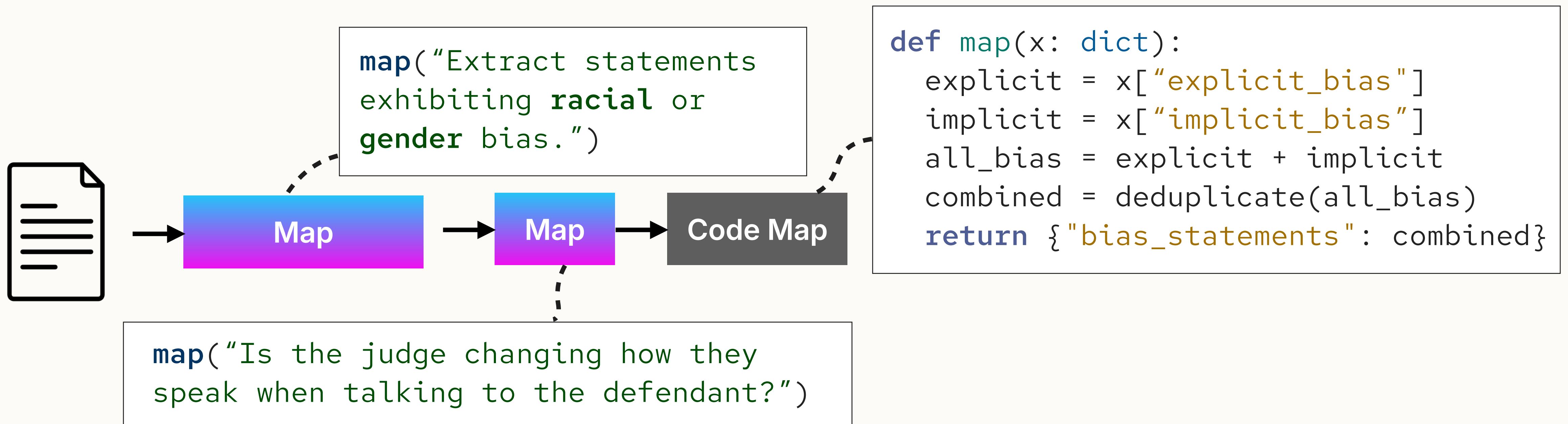
Cost-Oriented Rewrite Directives

Replace with Code: $op \Rightarrow \text{code_op}$

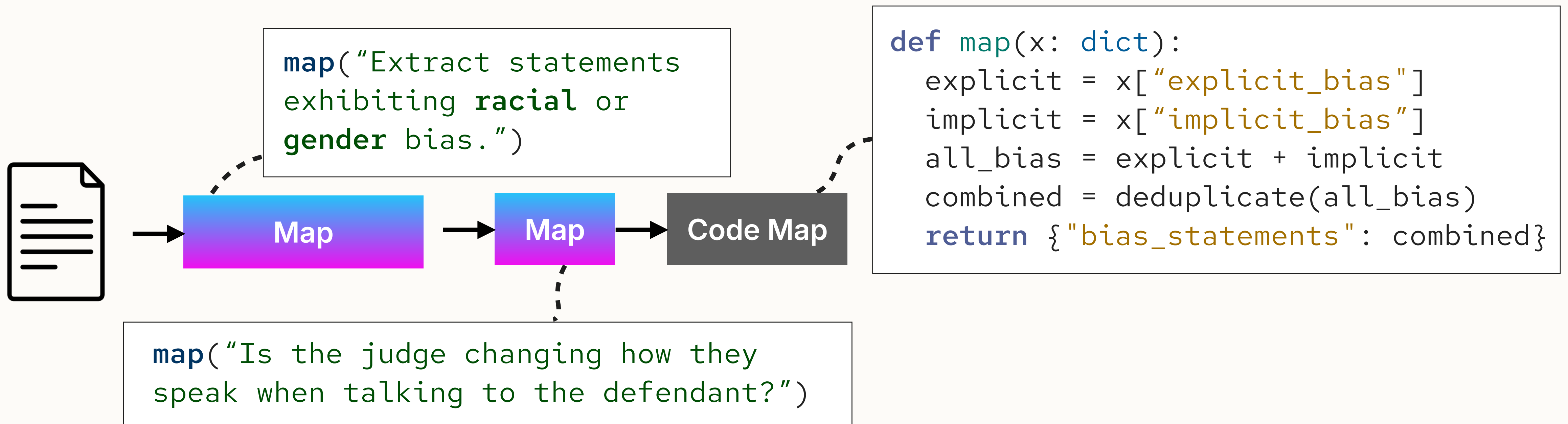


Cost-Oriented Rewrite Directives

Replace with Code: $op \Rightarrow \text{code_op}$

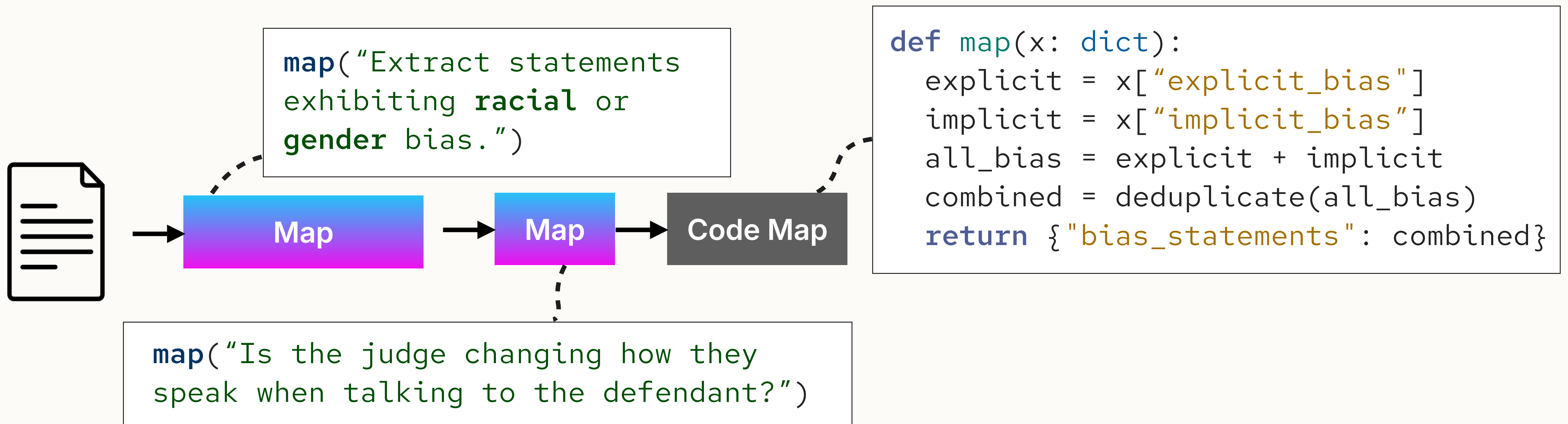


Cost-Oriented Rewrite Directives



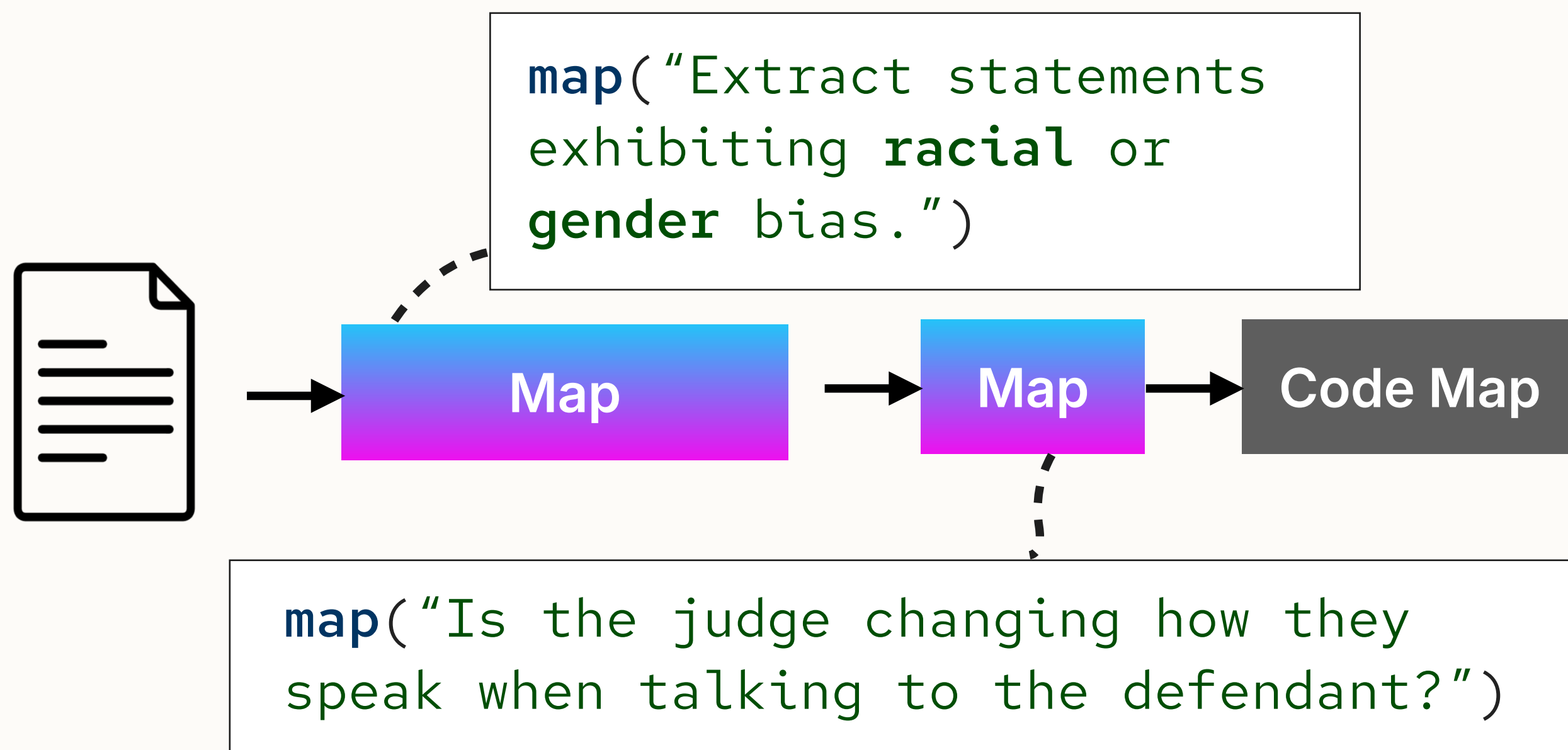
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



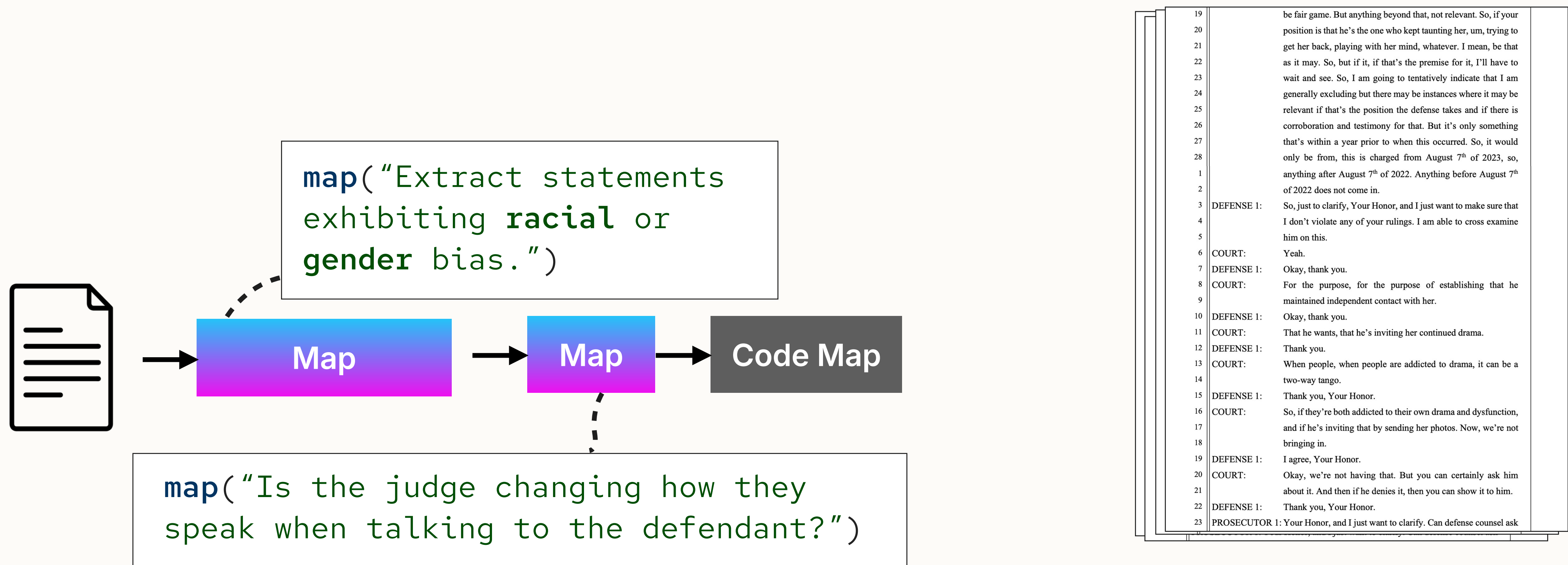
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



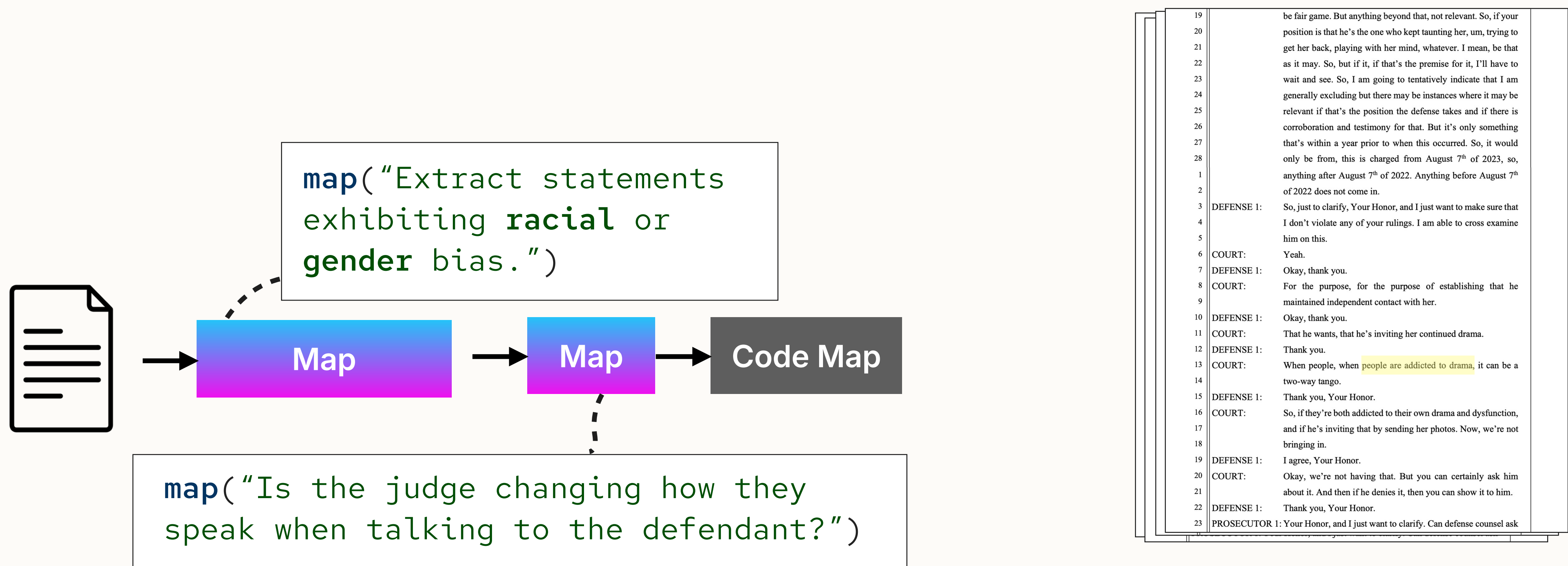
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



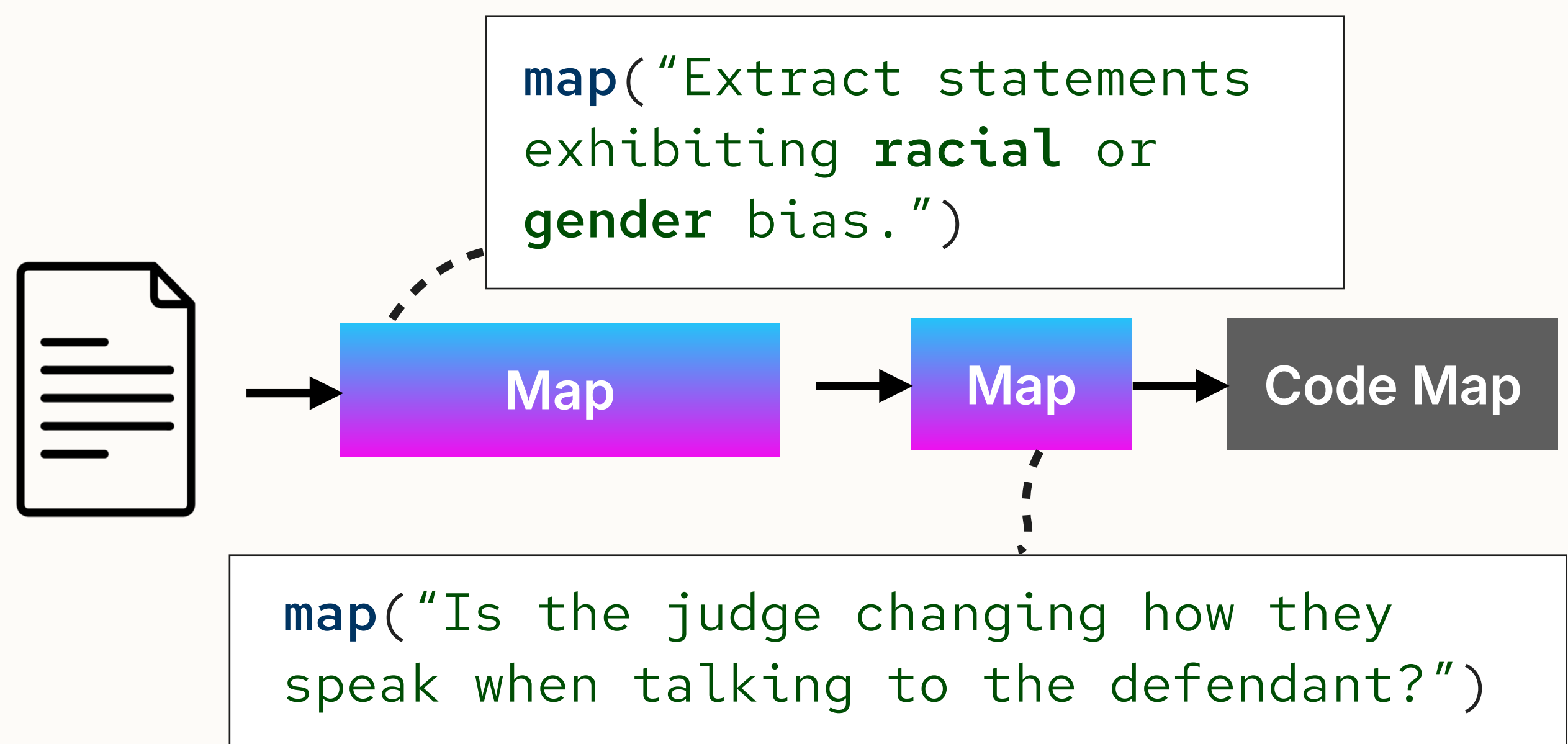
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



Cost-Oriented Rewrite Directives

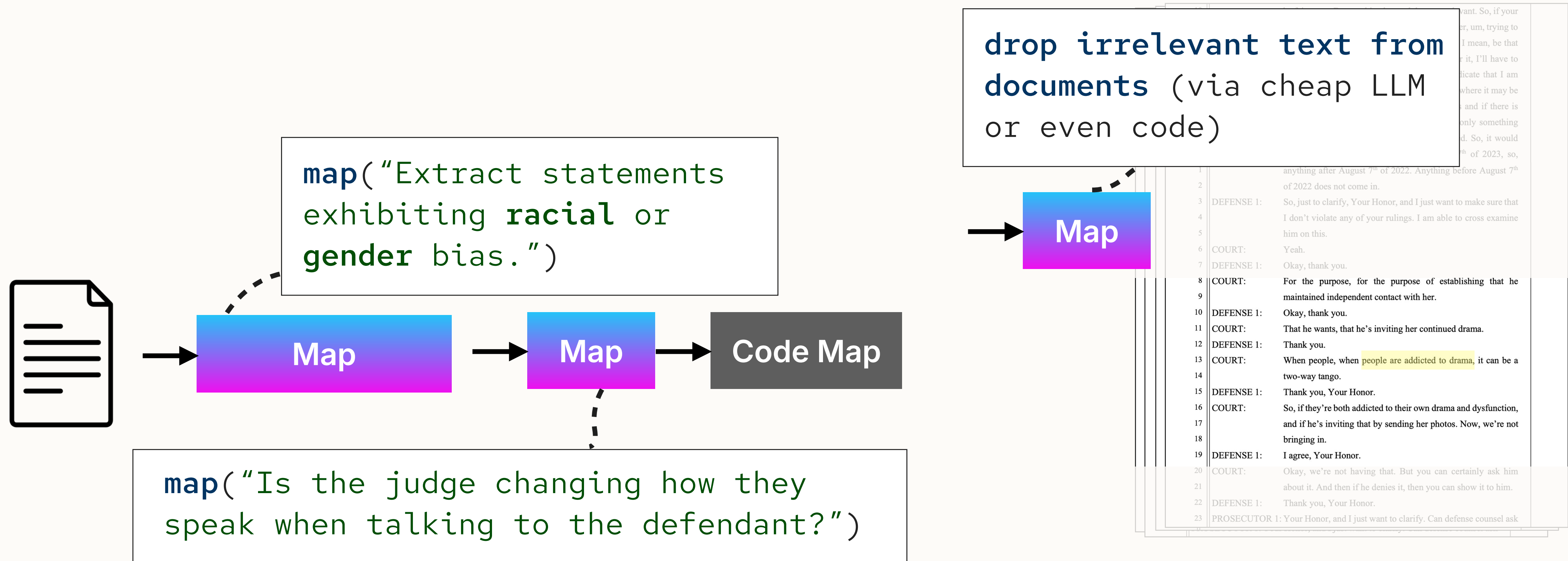
Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



19		be fair game. But anything beyond that, not relevant. So, if your	
20		position is that he's the one who kept taunting her, um, trying to	
21		get her back, playing with her mind, whatever. I mean, be that	
22		as it may. So, but if it, if that's the premise for it, I'll have to	
23		wait and see. So, I am going to tentatively indicate that I am	
24		generally excluding but there may be instances where it may be	
25		relevant if that's the position the defense takes and if there is	
26		corroboration and testimony for that. But it's only something	
27		that's within a year prior to when this occurred. So, it would	
28		only be from, this is charged from August 7 th of 2023, so,	
1		anything after August 7 th of 2022. Anything before August 7 th	
2		of 2022 does not come in.	
3	DEFENSE 1:	So, just to clarify, Your Honor, and I just want to make sure that	
4		I don't violate any of your rulings. I am able to cross examine	
5		him on this.	
6	COURT:	Yeah.	
7	DEFENSE 1:	Okay, thank you.	
8	COURT:	For the purpose, for the purpose of establishing that he	
9		maintained independent contact with her.	
10	DEFENSE 1:	Okay, thank you.	
11	COURT:	That he wants, that he's inviting her continued drama.	
12	DEFENSE 1:	Thank you.	
13	COURT:	When people, when people are addicted to drama, it can be a	
14		two-way tango.	
15	DEFENSE 1:	Thank you, Your Honor.	
16	COURT:	So, if they're both addicted to their own drama and dysfunction,	
17		and if he's inviting that by sending her photos. Now, we're not	
18		bringing in.	
19	DEFENSE 1:	I agree, Your Honor.	
20	COURT:	Okay, we're not having that. But you can certainly ask him	
21		about it. And then if he denies it, then you can show it to him.	
22	DEFENSE 1:	Thank you, Your Honor.	
23	PROSECUTOR 1:	Your Honor, and I just want to clarify. Can defense counsel ask	

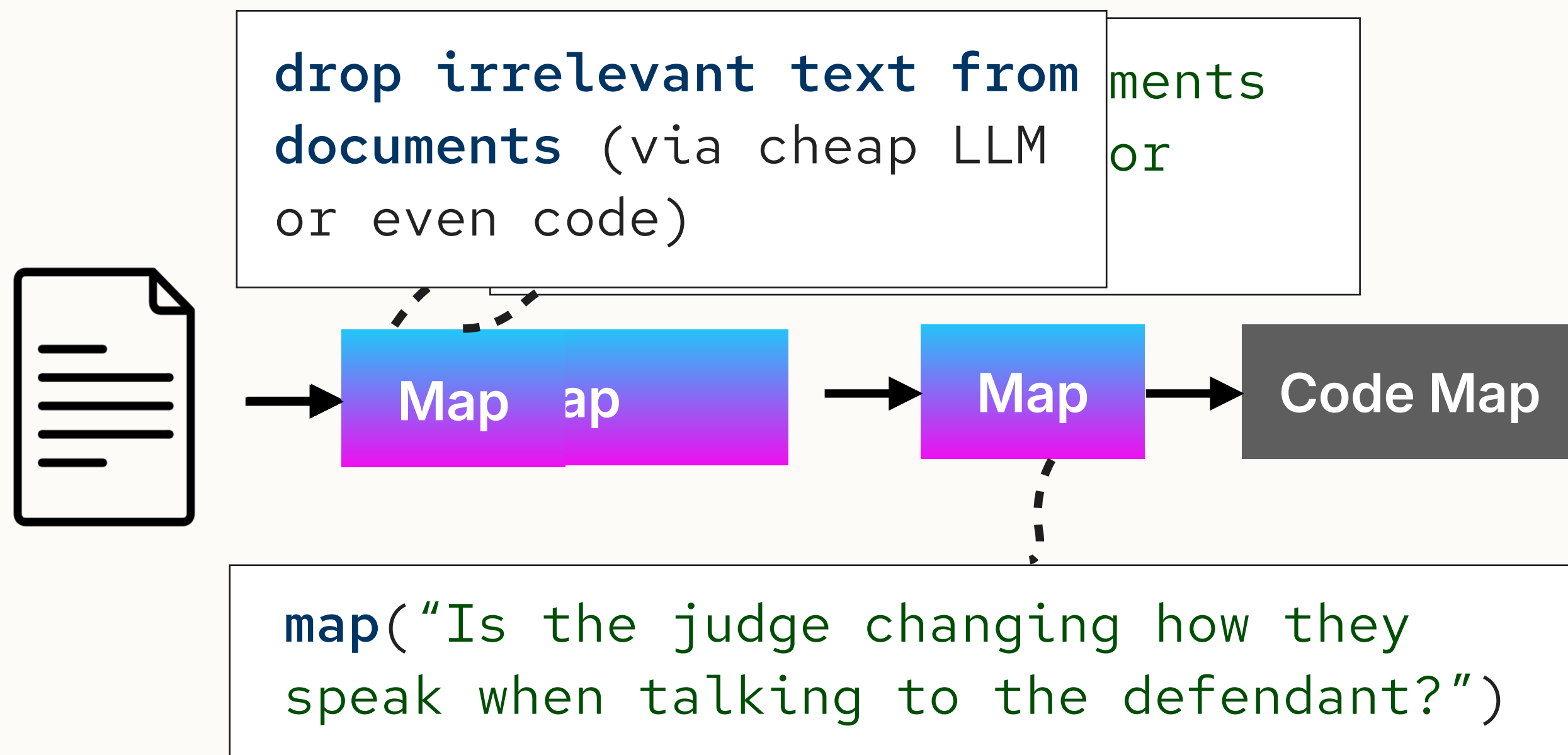
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



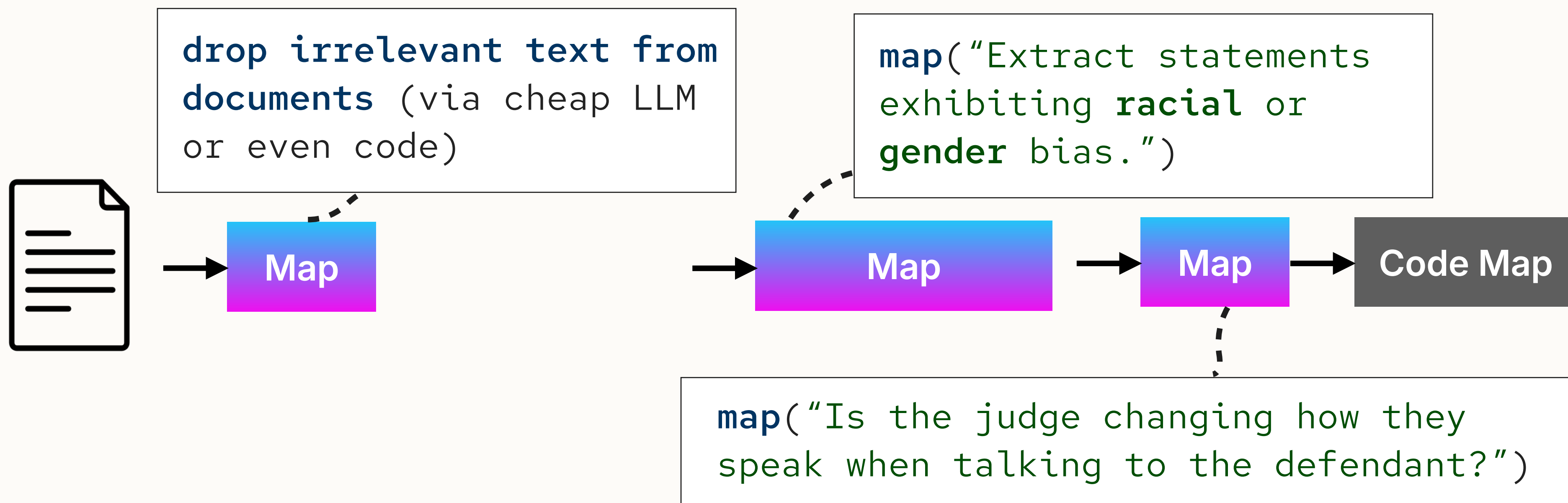
Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



Cost-Oriented Rewrite Directives

Add a compression step & push it down: $op^+ \Rightarrow \text{map} \rightarrow op^+$



New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ Resolve
- ◆ Extract
- ◆ Sample

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1

 Officer J. Smith...

Chunk 2

He then proceeded to...

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1

 Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1

 Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

Context Types

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1

👮 Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

Context Types

⬇ Previous/Next
Chunks

See Figure 2 on the
next page for...

[Fig 2] Architecture
diagram...

Need next chunk for
referenced context

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1

 Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

Context Types

Previous/Next Chunks

See Figure 2 on the next page for...

[Fig 2] Architecture diagram...

Need next chunk for referenced context

Transformed Content

Previous 200 pages:
"Suspect was last seen in Paris..."

"He boarded a train to..."

Summary of a long prefix

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1



Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

Context Types



Previous/Next
Chunks

See Figure 2 on the
next page for...

[Fig 2] Architecture
diagram...

Need next chunk for
referenced context



Transformed
Content

Previous 200 pages:
"Suspect was last
seen in Paris..."

"He boarded a train
to..."

Summary of a long prefix



Document
Metadata

1. Contract Terms
1.1 Licensing
1.1.2 Usage Rights

The licensee shall...

Section hierarchy context

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ **Gather**
- ◆ Resolve
- ◆ Extract
- ◆ Sample

Challenge

Chunk 1



Officer J. Smith...

Chunk 2

He then proceeded to...



Who is "he"?

What happened before?

Context Types



Previous/Next
Chunks

See Figure 2 on the
next page for...

[Fig 2] Architecture
diagram...

Need next chunk for
referenced context



Transformed
Content

Previous 200 pages:
"Suspect was last
seen in Paris..."

"He boarded a train
to..."

Summary of a long prefix



Document
Metadata

1. Contract Terms
1.1 Licensing
1.1.2 Usage Rights

The licensee shall...

Section hierarchy context

split → map ⇒ split → **gather** → map

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

New Operators for DocETL

Rewrite directives require new operators:


- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

Challenge

Document 1

 Officer X. Quinnsworth...

Document 2

 Sgt. Xander Quinnsworth was...

Document 3

 Officer Quinnswrth, badge #...

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

Challenge

Document 1



Officer X. Quinnsworth...

Quinnsworth

Document 2



Sgt. Xander Quinnsworth was...

Document 3



Officer Quinnswrth, badge #...

Officer
Quinnsworth

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

Challenge

Document 1

👮 Officer X. Quinnsworth...

Document 2

👮 Sgt. Xander Quinnsworth was...

Document 3

👮 Officer Quinnswrth, badge #...

Quinnsworth

Officer
Quinnsworth

Officer names are
inconsistently referred to
across documents!

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

Challenge

Document 1

👮 Officer X. Quinnsworth...

Document 2

👮 Sgt. Xander Quinnsworth was...

Document 3

👮 Officer Quinnswrth, badge #...

Quinnsworth

Officer
Quinnsworth

Officer names are
inconsistently referred to
across documents!

Even if they are
consistently represented,
**an LLM might
inconsistently extract
them.**

New Operators for DocETL

Rewrite directives require new operators:

- ◆ Split
- ◆ Gather
- ◆ **Resolve**
- ◆ Extract
- ◆ Sample

Challenge

Document 1

👮 Officer X. Quinnsworth...

Document 2

👮 Sgt. Xander Quinnsworth was...

Document 3

👮 Officer Quinnswrth, badge #...

Quinnsworth

Officer
Quinnsworth

Officer names are
inconsistently referred to
across documents!

Even if they are
consistently represented,
**an LLM might
inconsistently extract
them.**

map → reduce ⇒ map → **resolve** → reduce

Rewrite Directives at Scale

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:


1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan



Challenge: multi-objective (\$\$ and accuracy)

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

(New) Goal: multiple plans spanning the accuracy—cost frontier

Rewrite Directives at Scale

We have 30+ directives that “reshape” pipelines — changing document size, number of docs, and the scope of each LLM call in a semantic operator.

Revisiting the optimizer:

1. Define the plan space
2. Estimate costs
3. Search for the optimal plan

(New) Goal: multiple plans spanning the accuracy—cost frontier

Challenges:

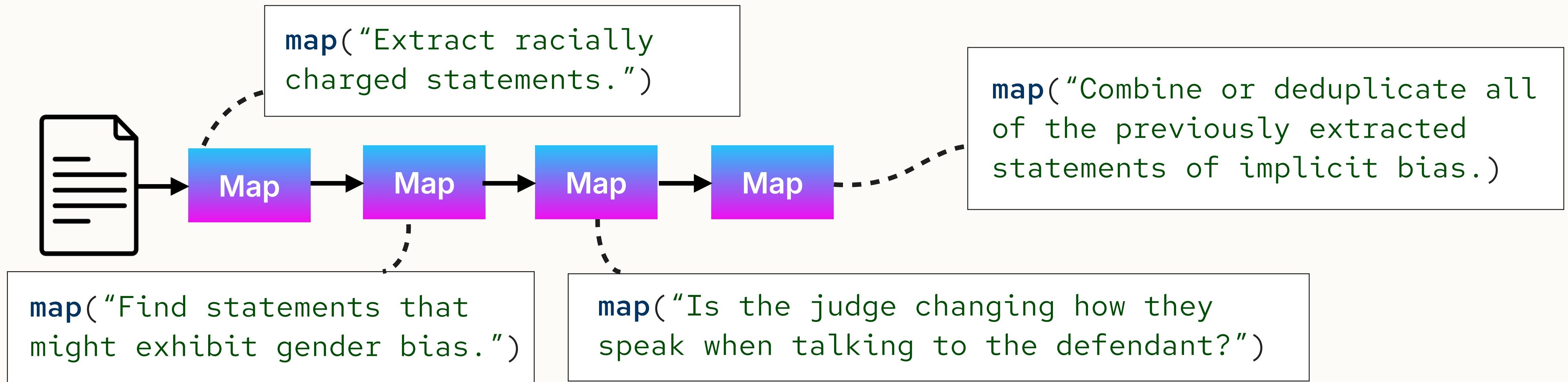
- * Accuracy estimation is expensive (can't explore too many alternatives)
- * Accuracy doesn't cleanly compose

Problem: Optimal Substructure

Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.

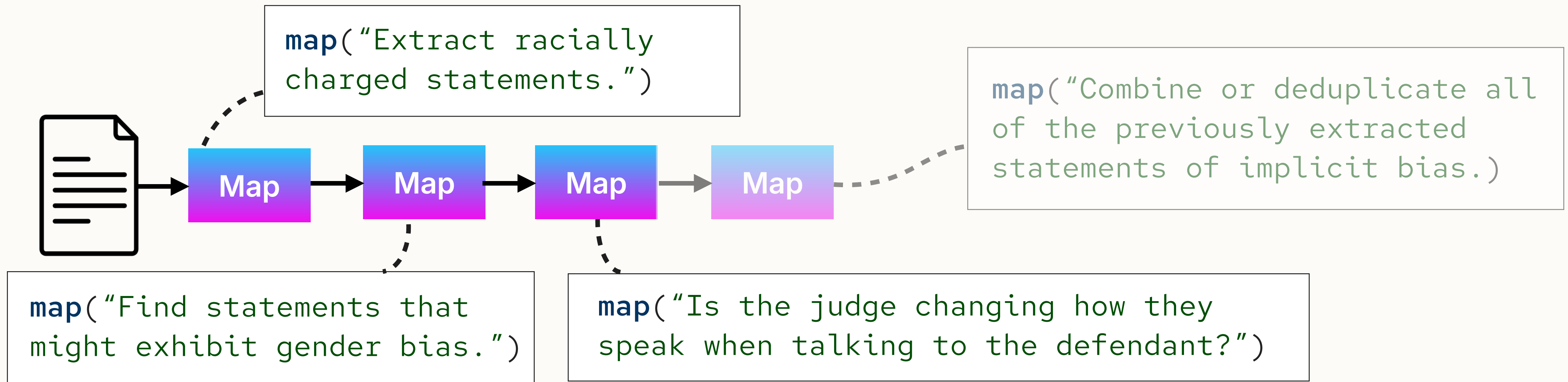
Problem: Optimal Substructure

Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.



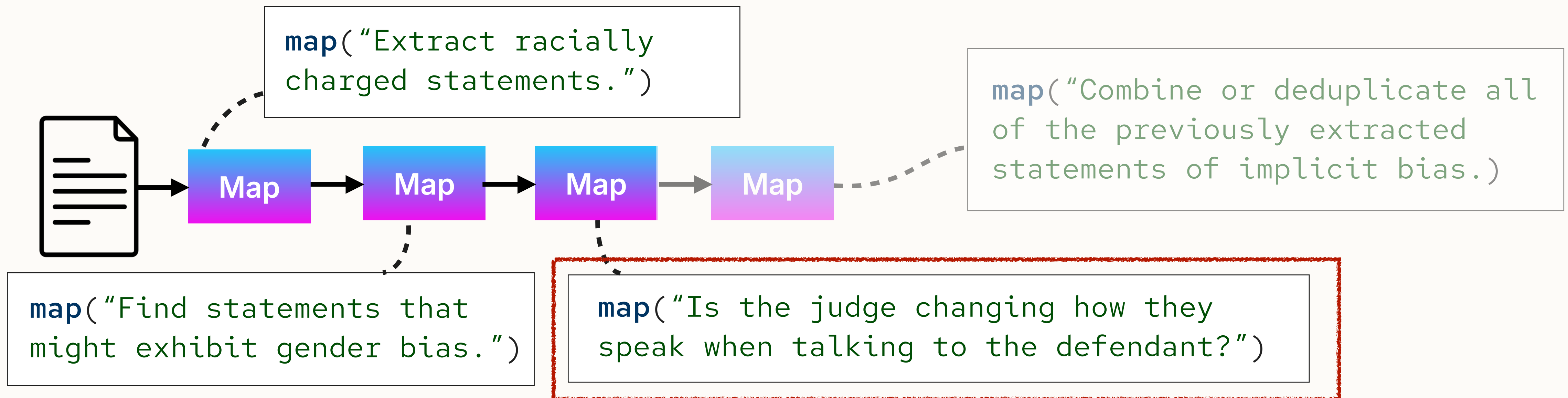
Problem: Optimal Substructure

Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.



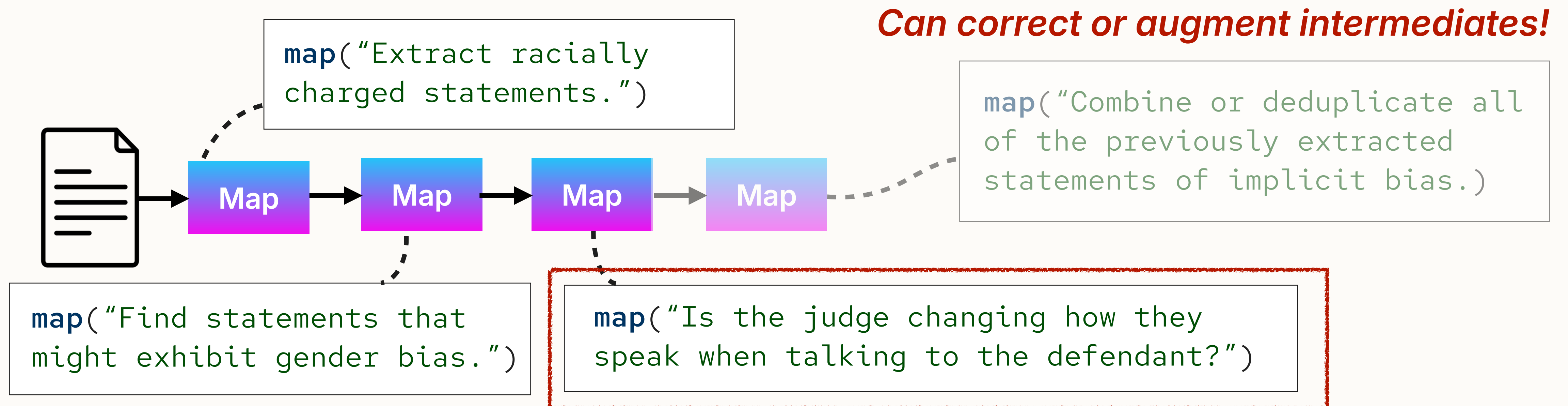
Problem: Optimal Substructure

Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.



Problem: Optimal Substructure

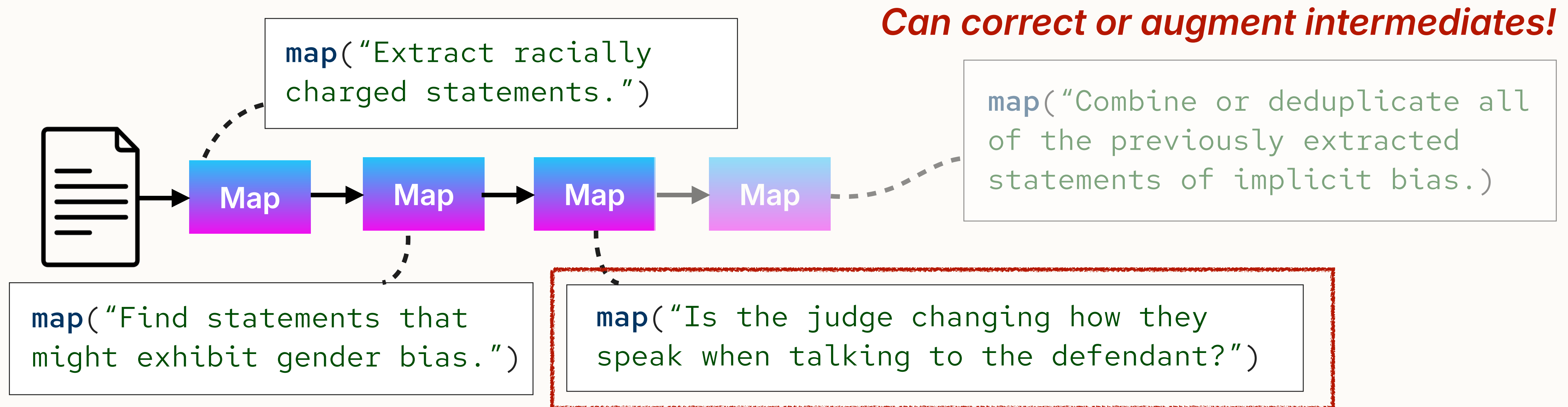
Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.



Problem: Optimal Substructure

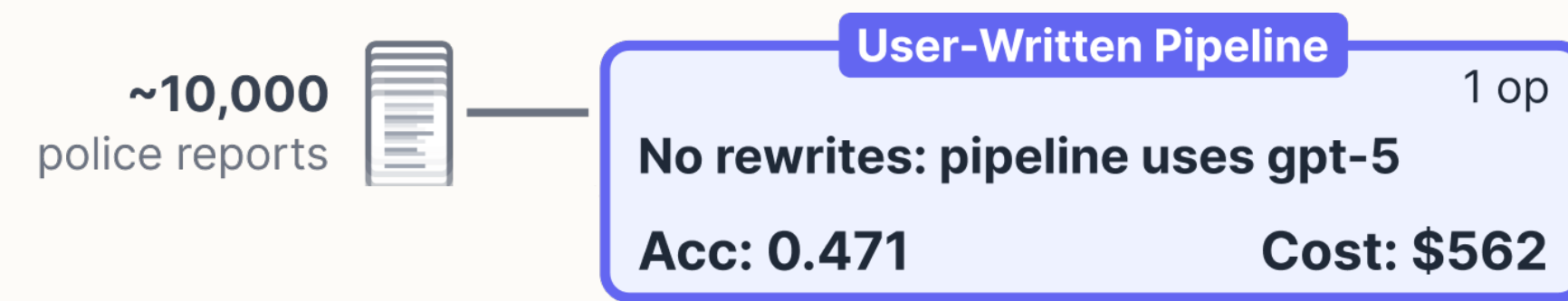
Classical query optimizers rely on the idea that the optimal plan is composed of optimal sub-plans.

Not true for semantic operators.
LLM behavior is context-sensitive and fuzzy! Doesn't cleanly compose.



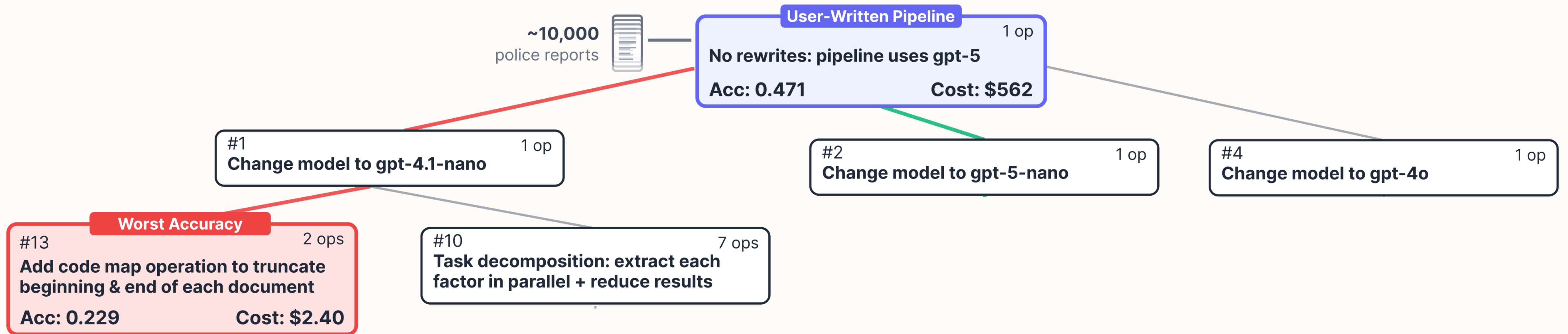
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



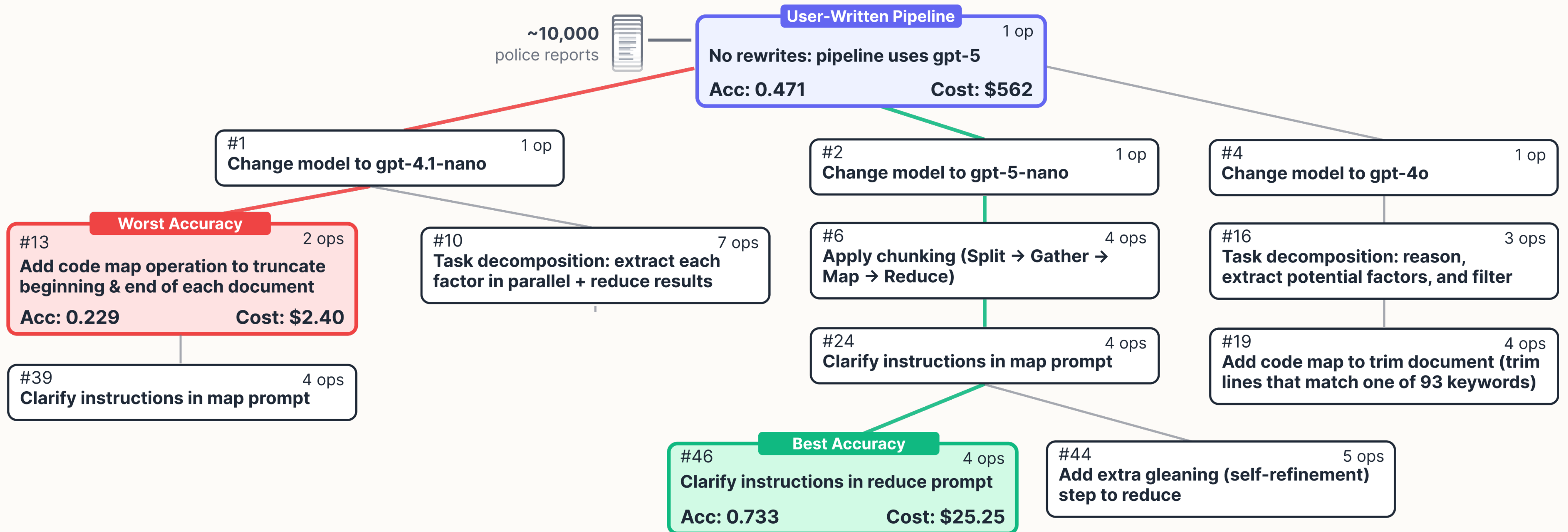
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



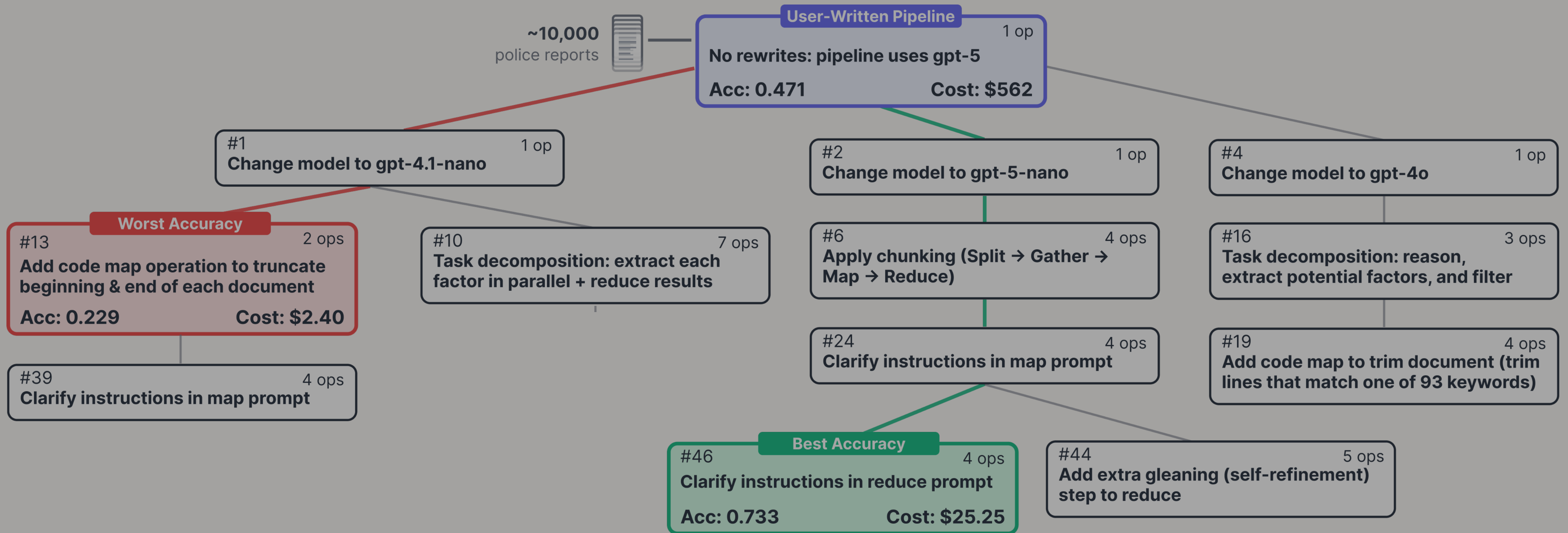
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



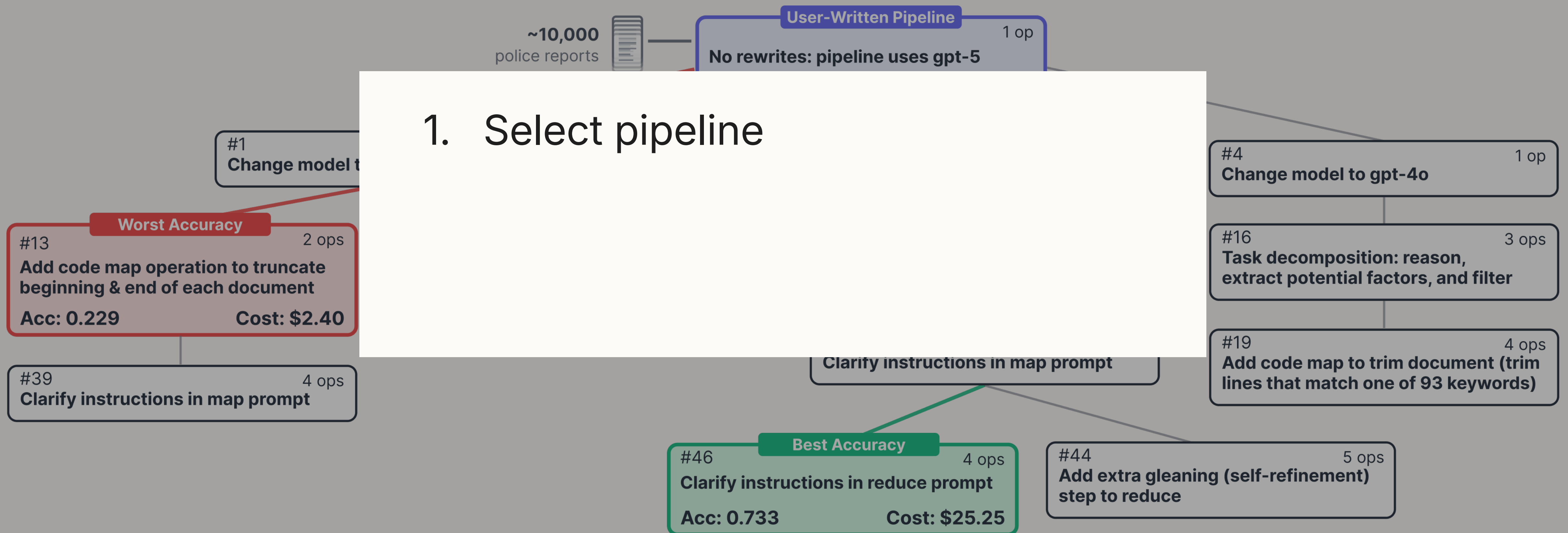
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



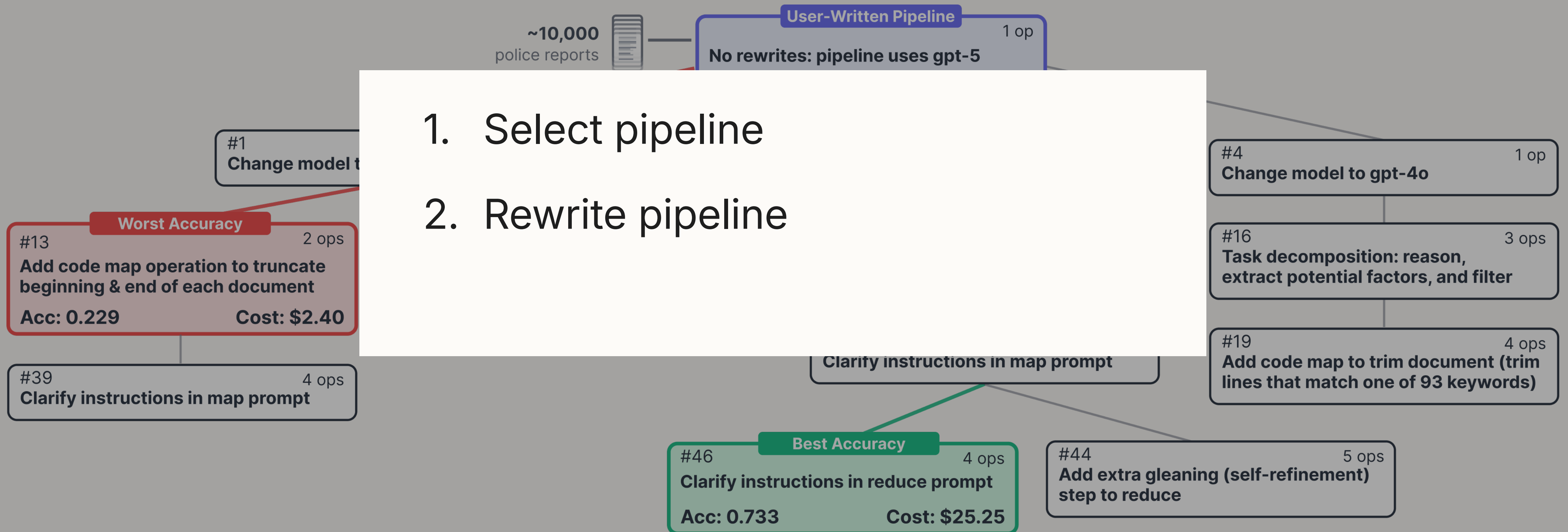
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



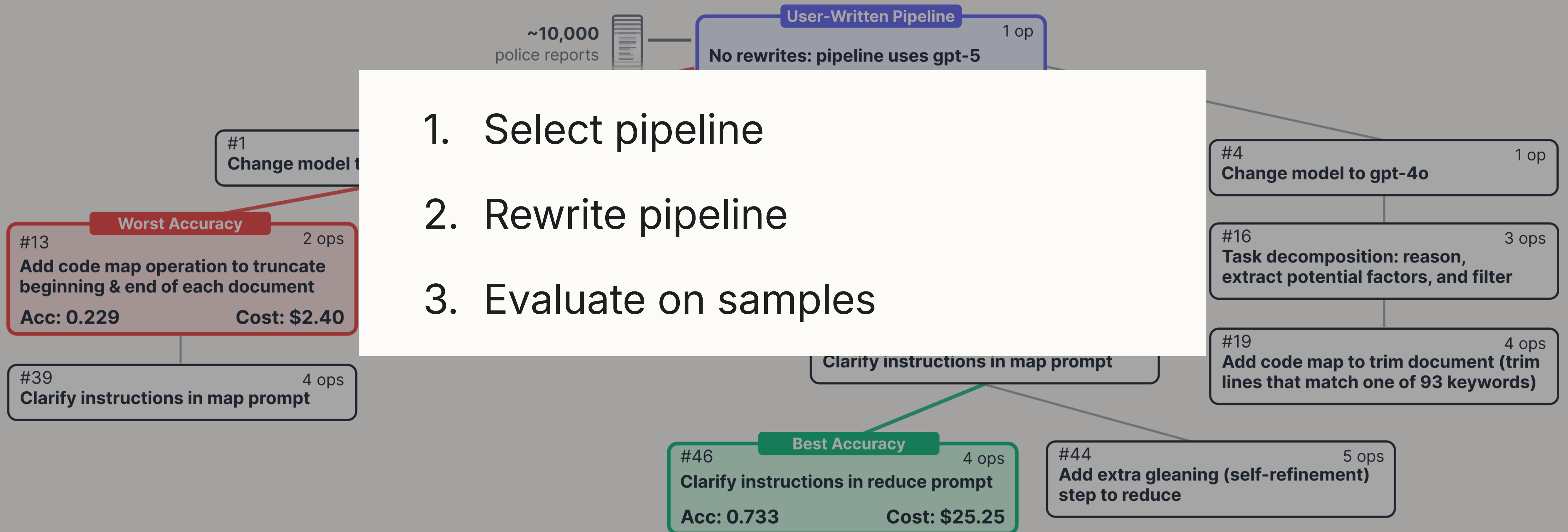
Search over Rewrite Directives

Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*

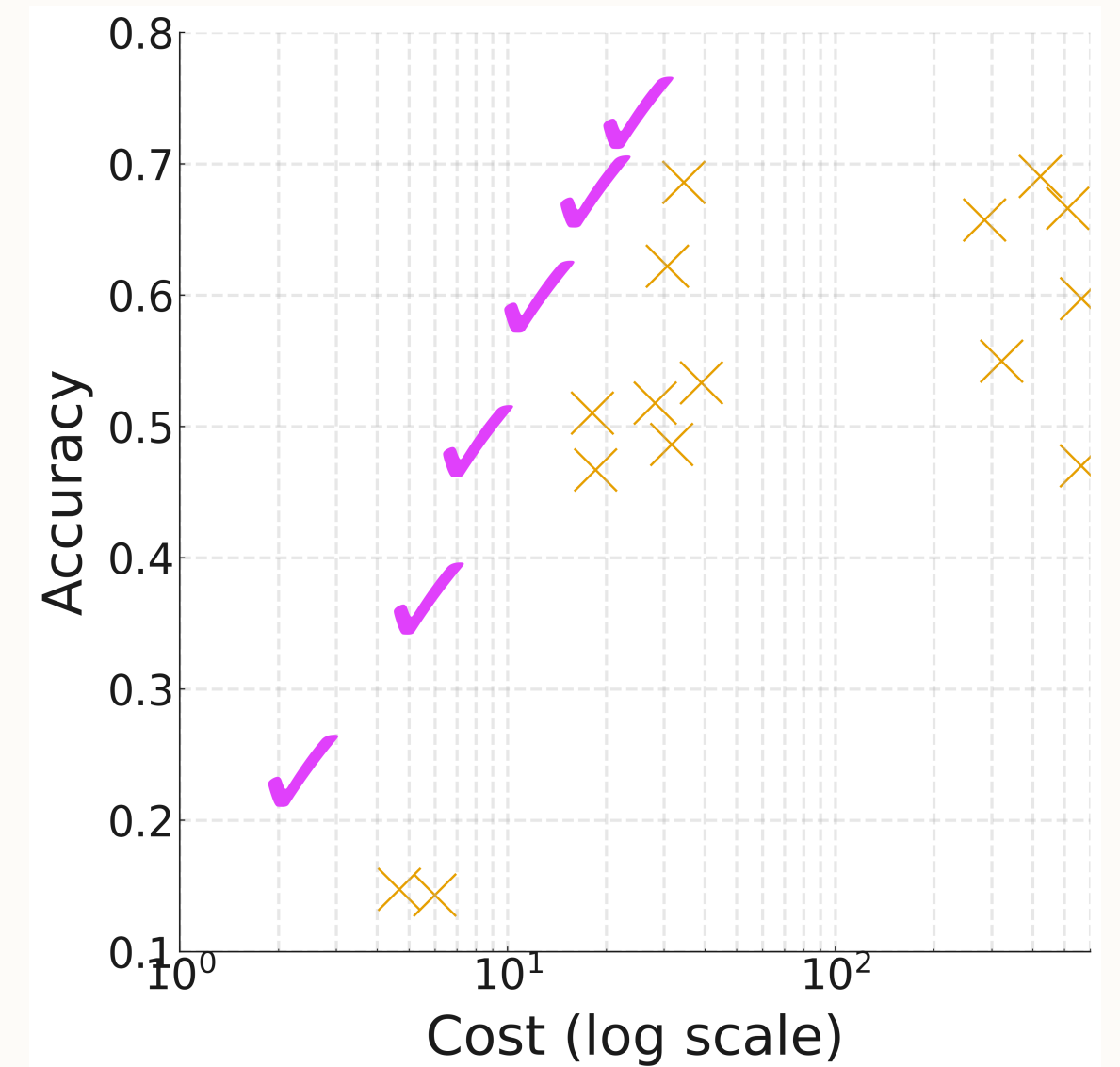
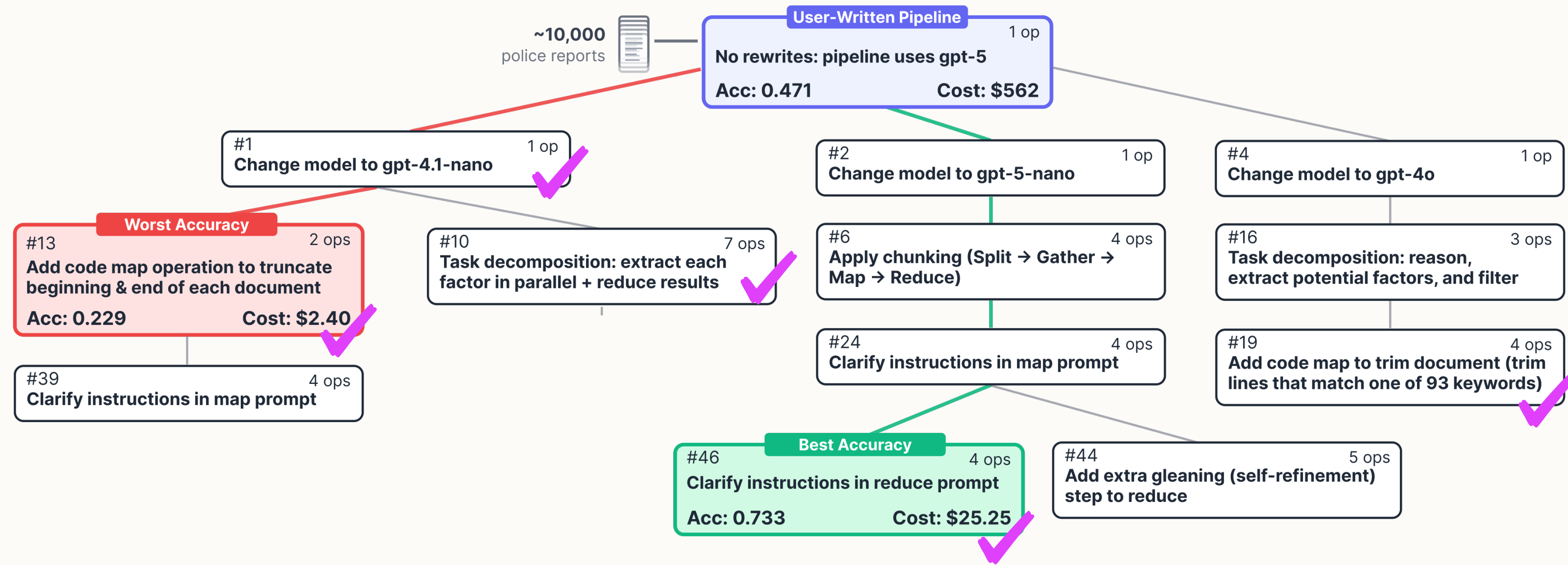


Search over Rewrite Directives

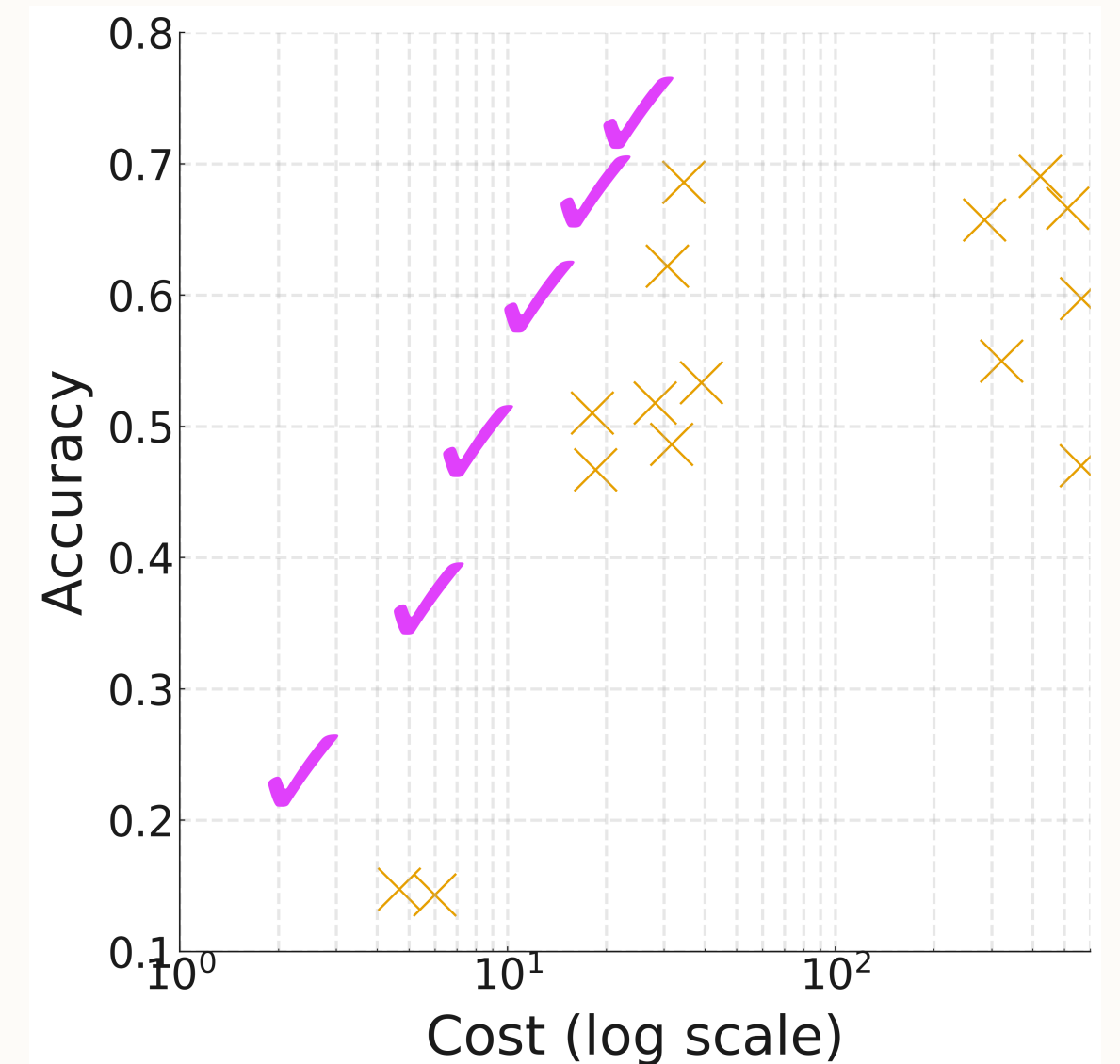
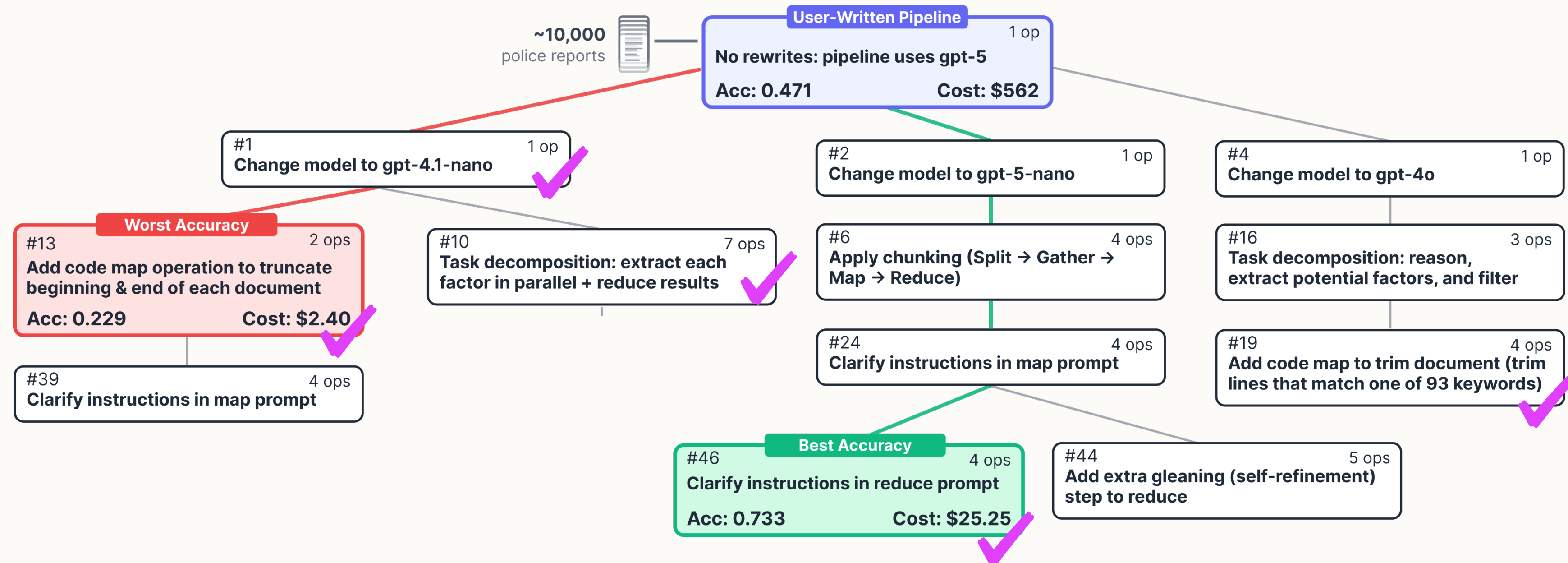
Multi-Objective Agentic Rewrites for Unstructured Data Processing. Wei*, **Shankar*** et al. *Under submission.*



Search over Rewrite Directives

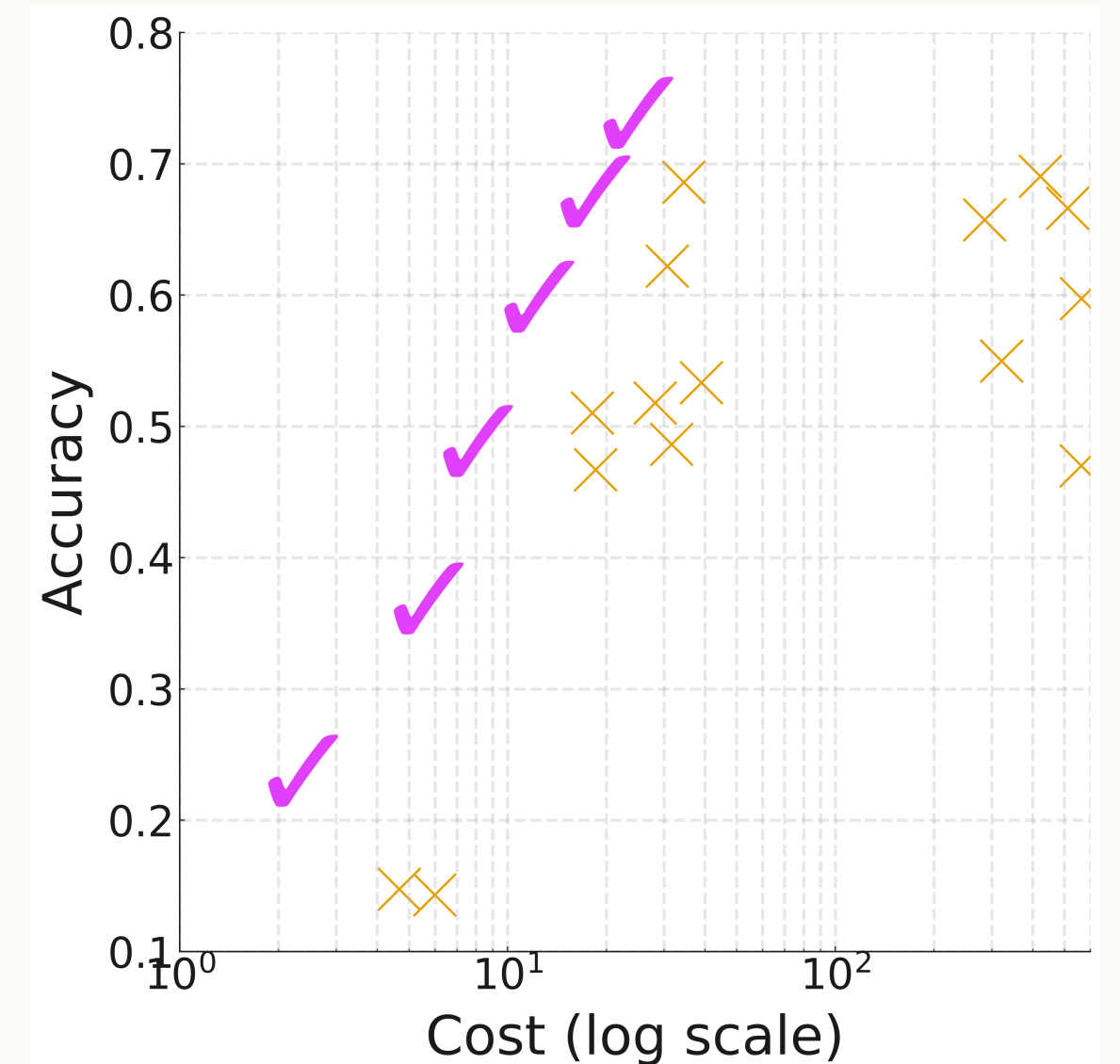
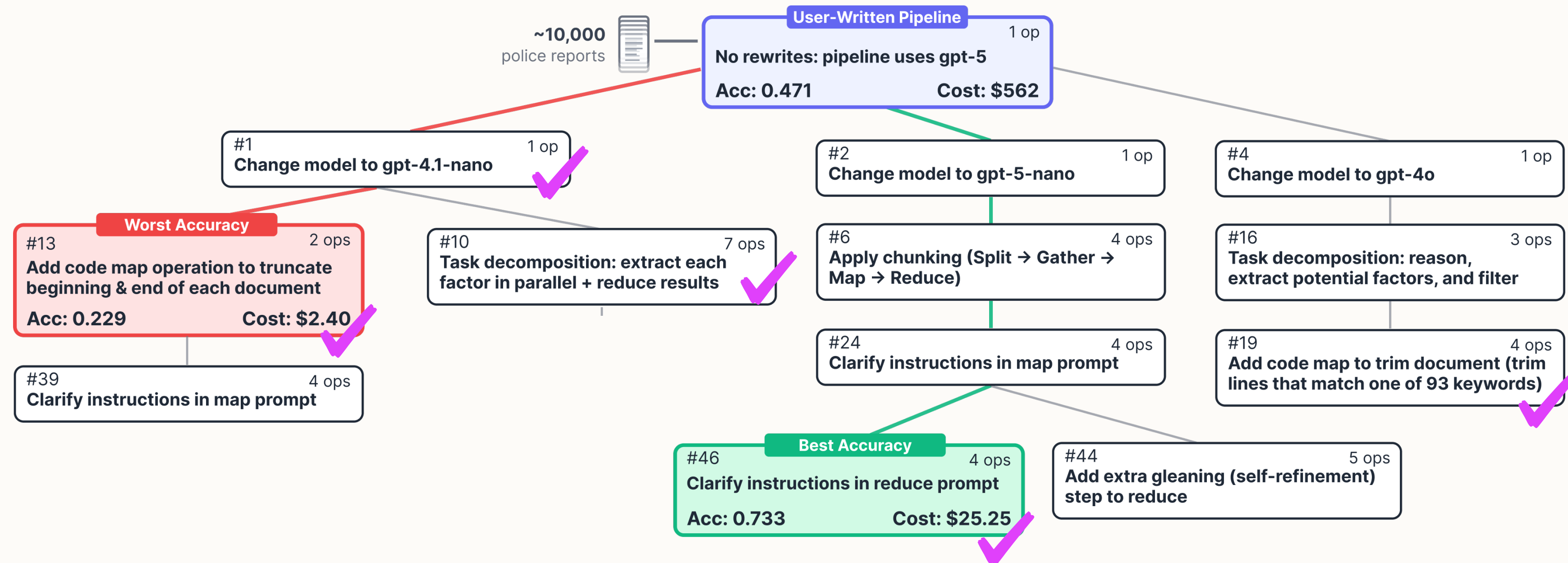


Search over Rewrite Directives



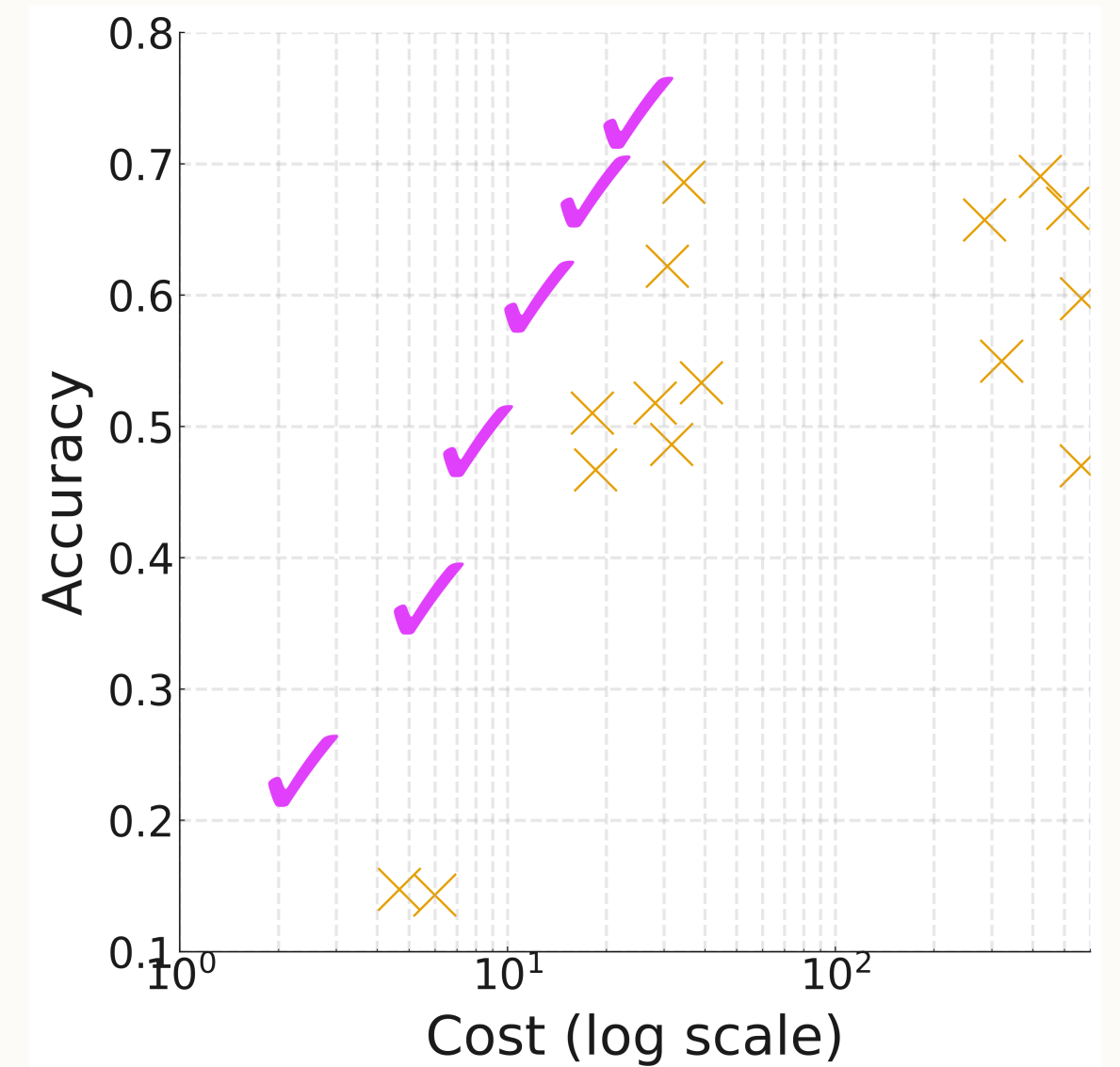
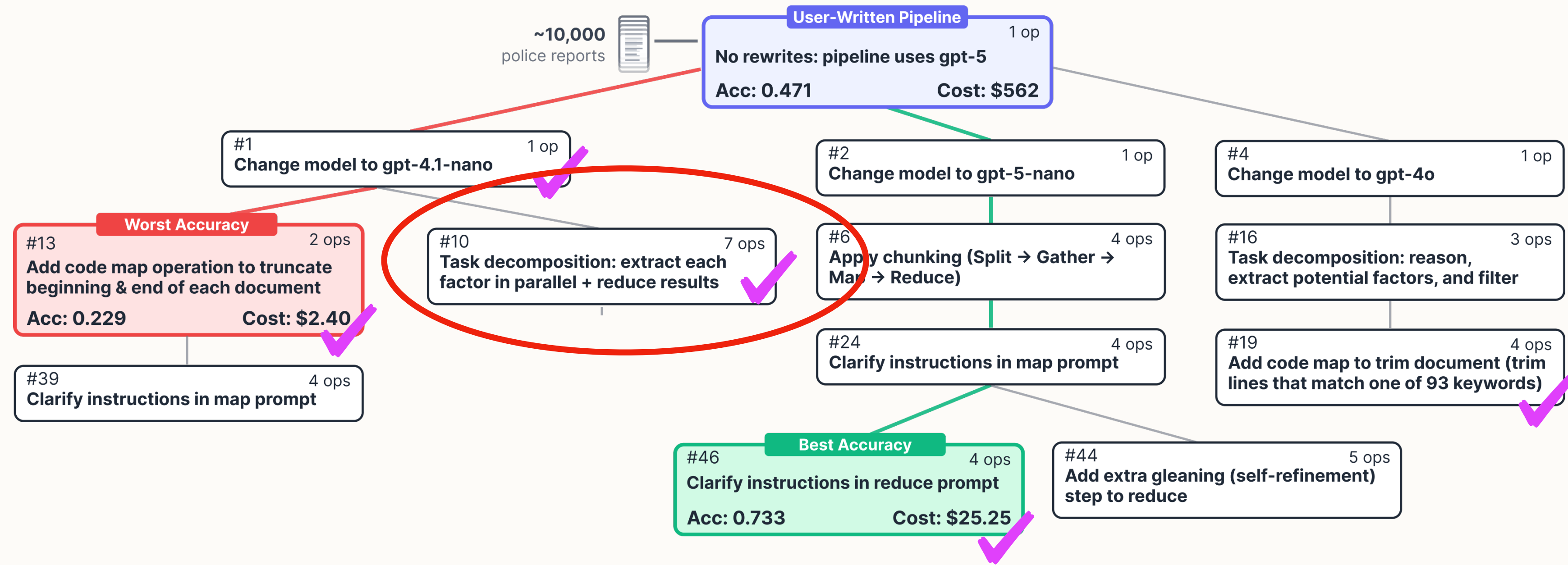
- Selection is deterministic (UCB applied to trees) & not LLM-based.

Search over Rewrite Directives



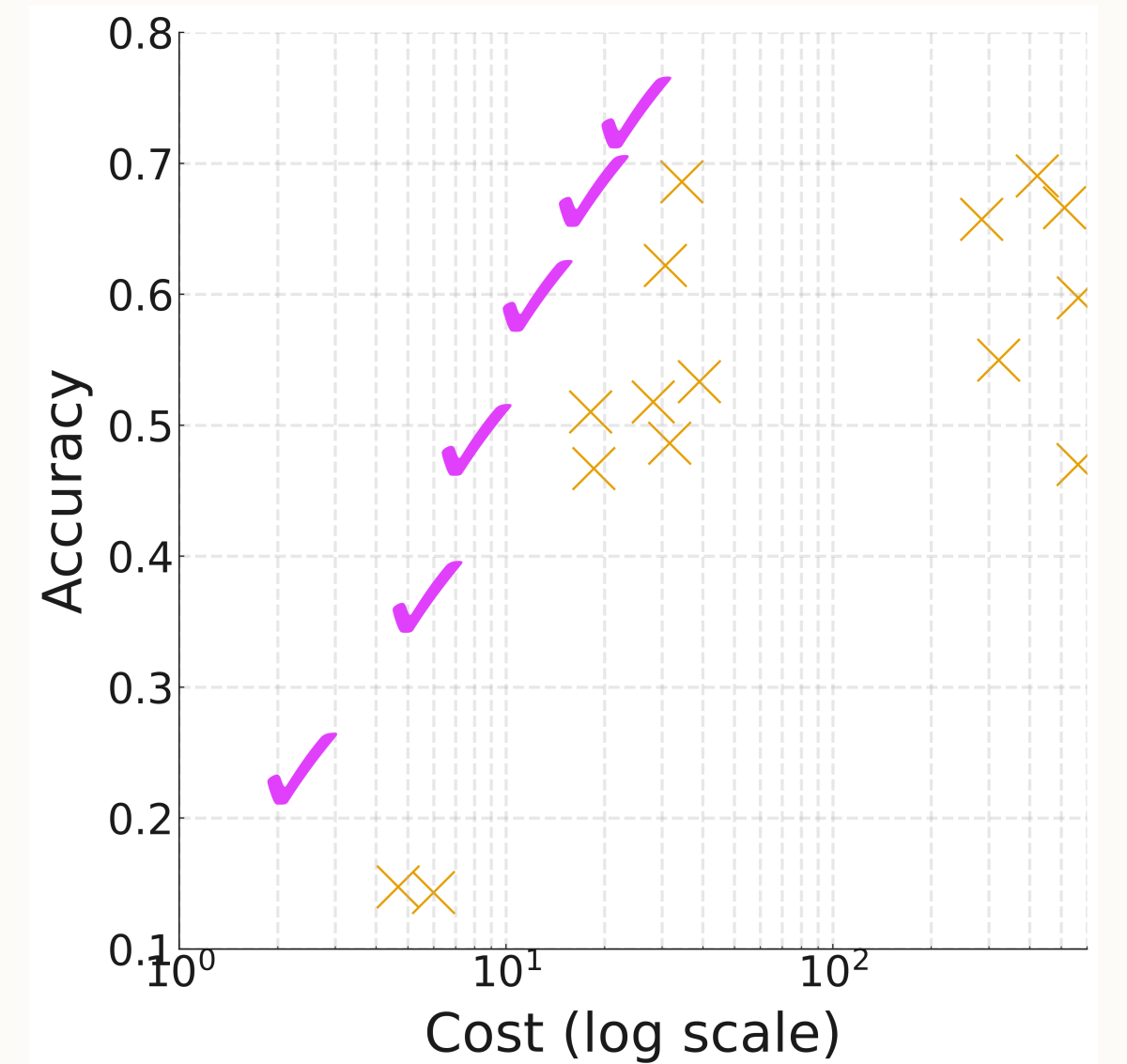
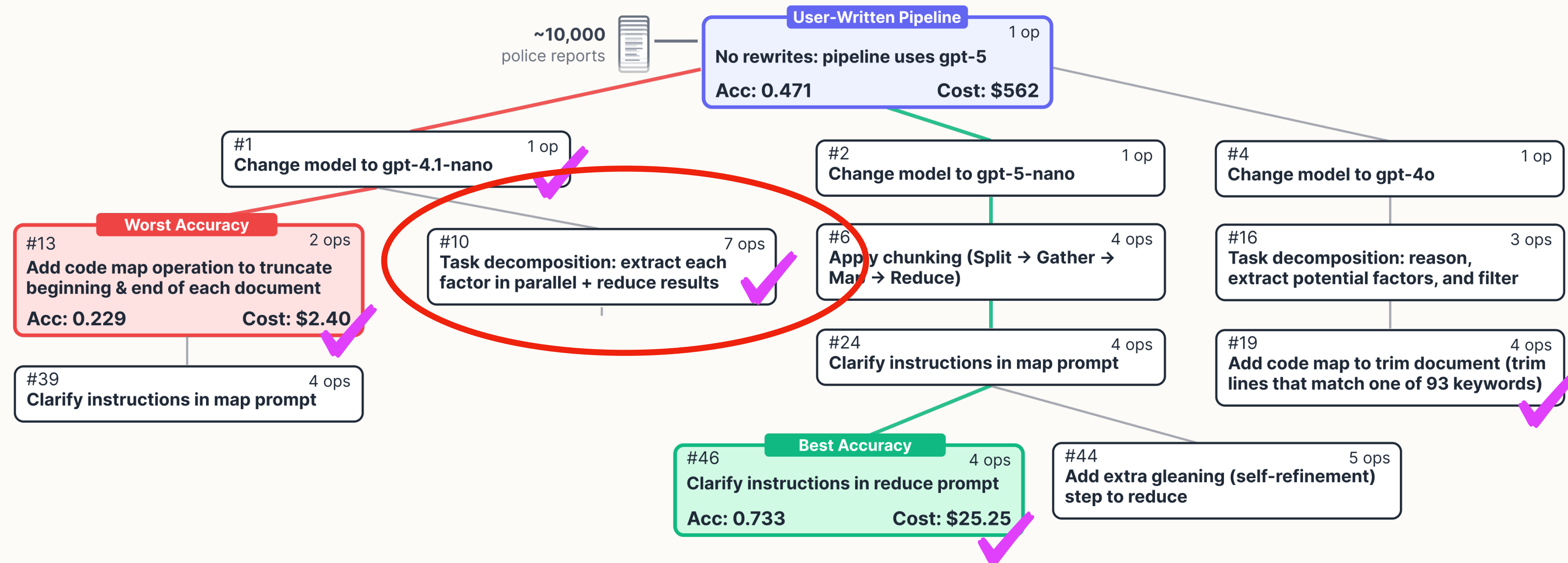
- Selection is deterministic (UCB applied to trees) & not LLM-based.
- We select a node to rewrite based on itself and its children's contributions to the frontier.

Search over Rewrite Directives

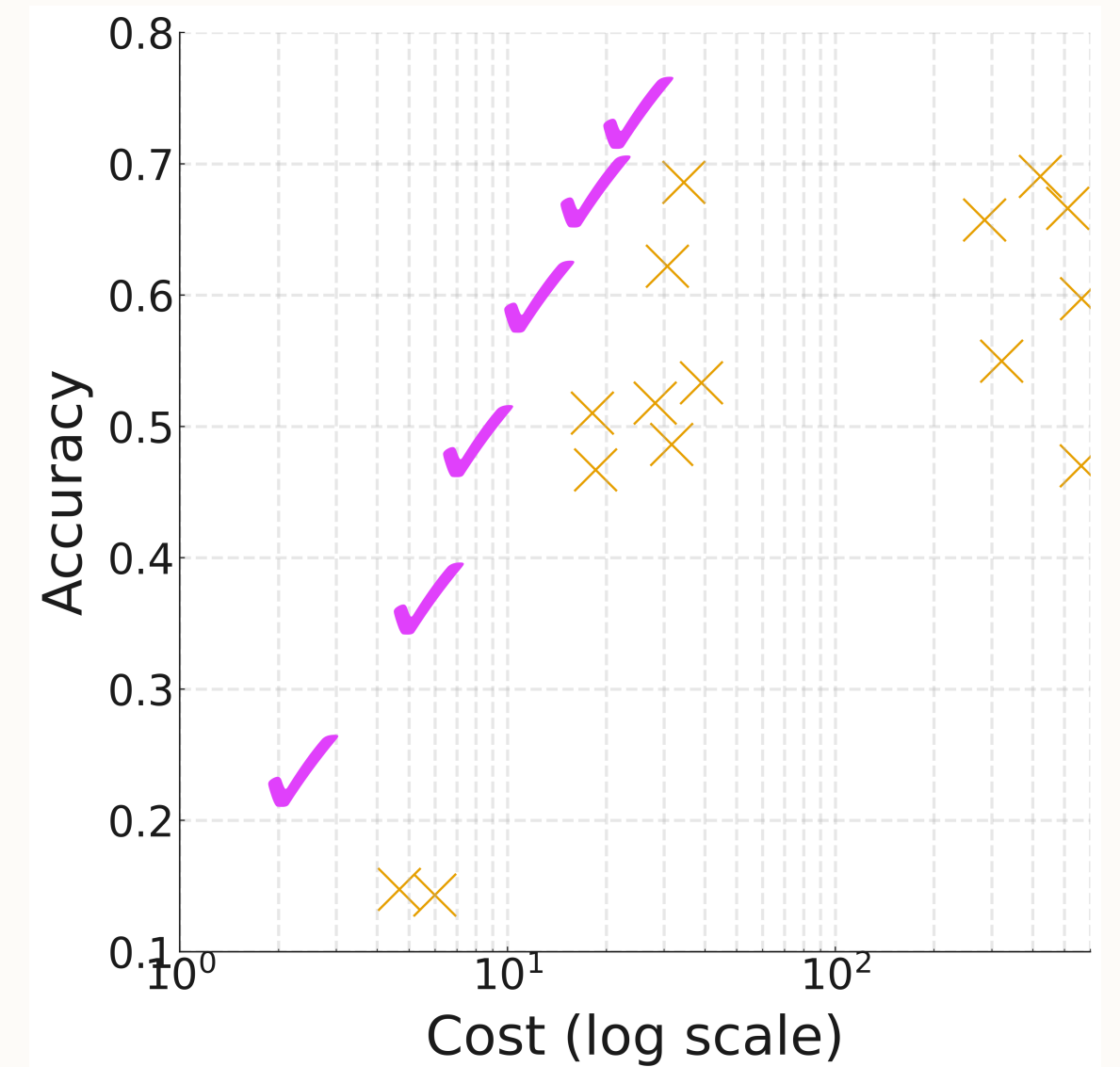
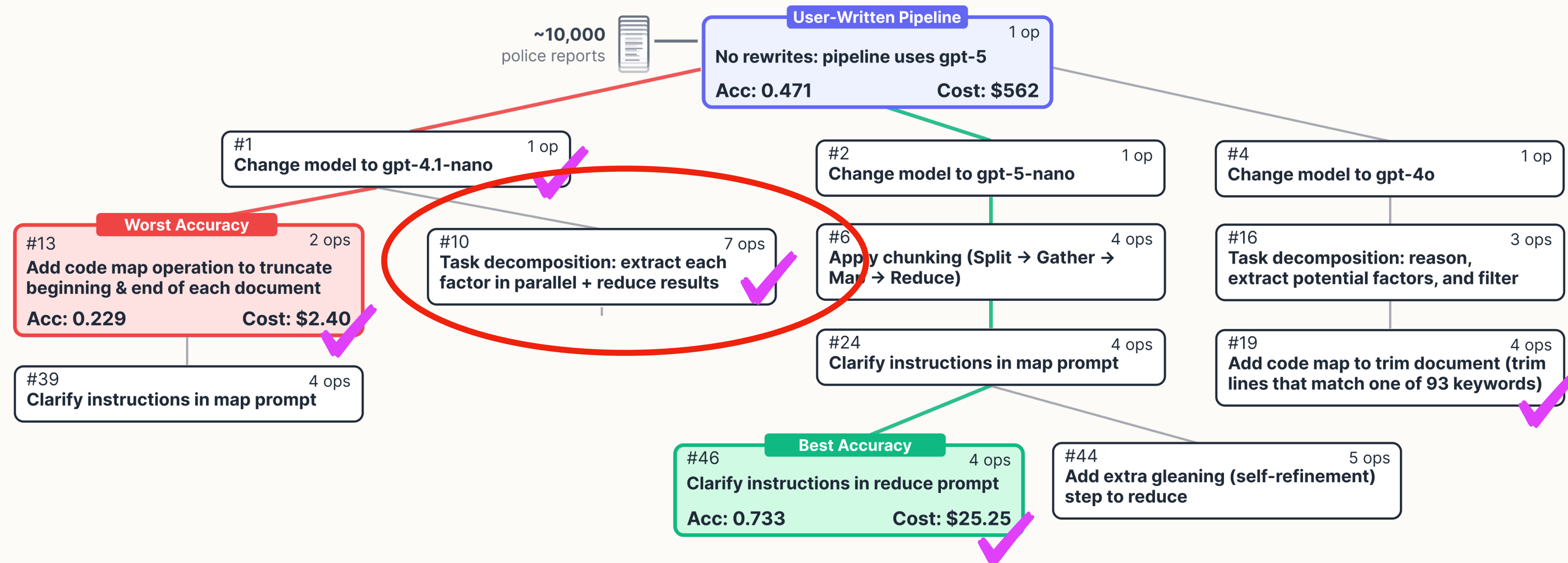


- Selection is deterministic (UCB applied to trees) & not LLM-based.
- We select a node to rewrite based on itself and its children's contributions to the frontier.

Search over Rewrite Directives

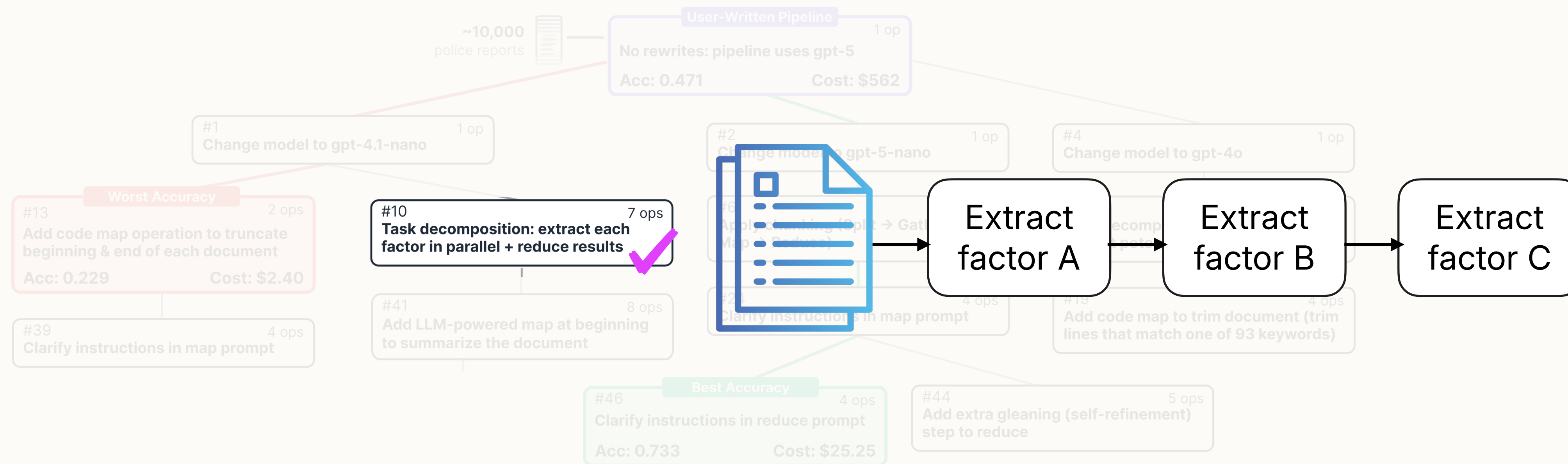


Search over Rewrite Directives



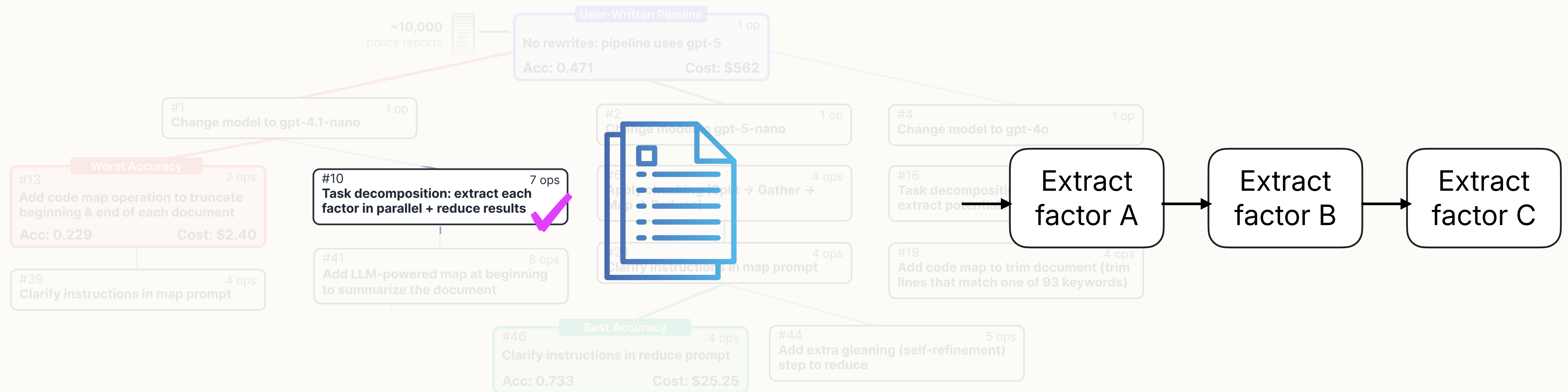
- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).

Search over Rewrite Directives



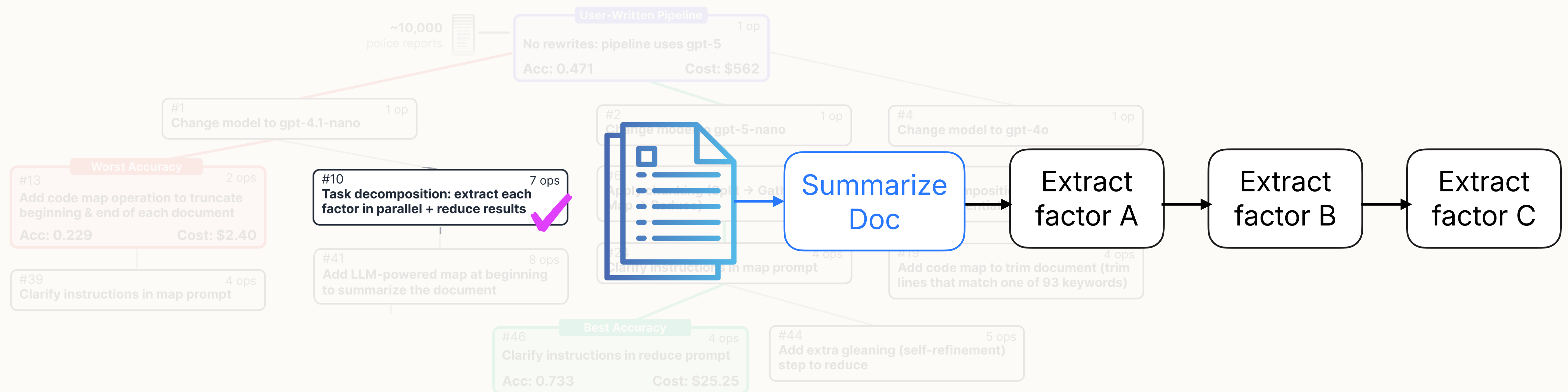
- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).

Search over Rewrite Directives



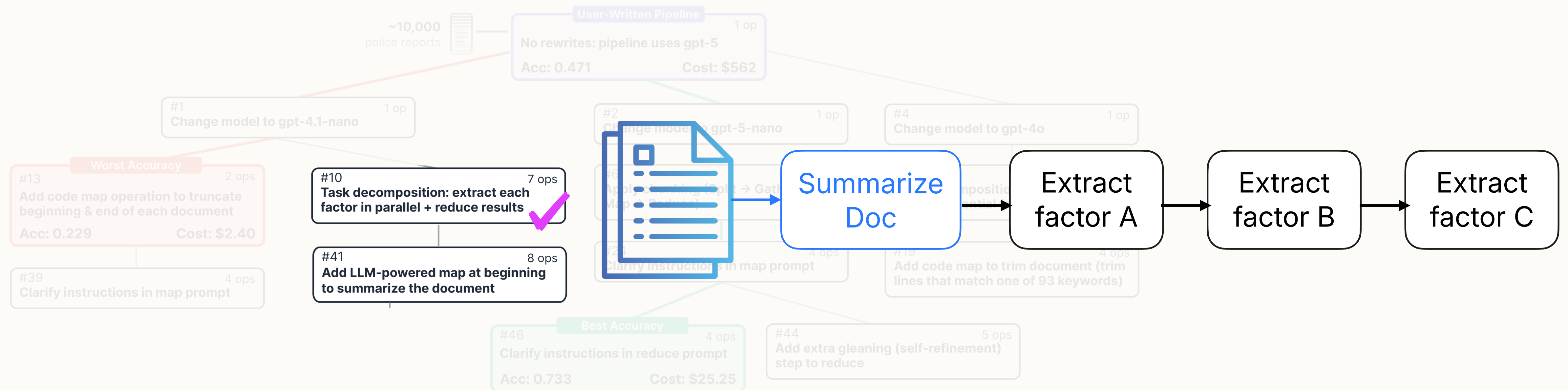
- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).

Search over Rewrite Directives

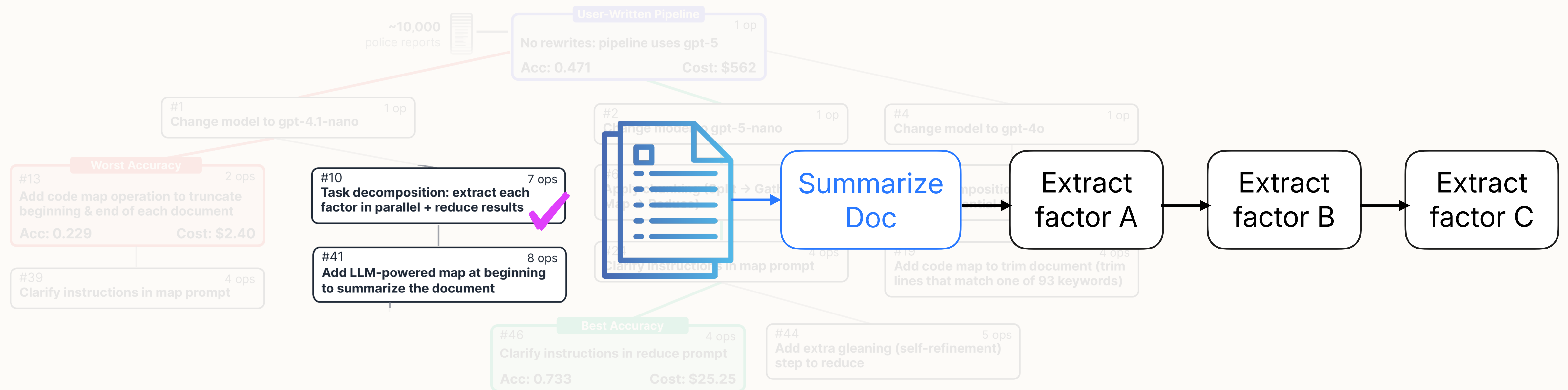


- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).

Search over Rewrite Directives

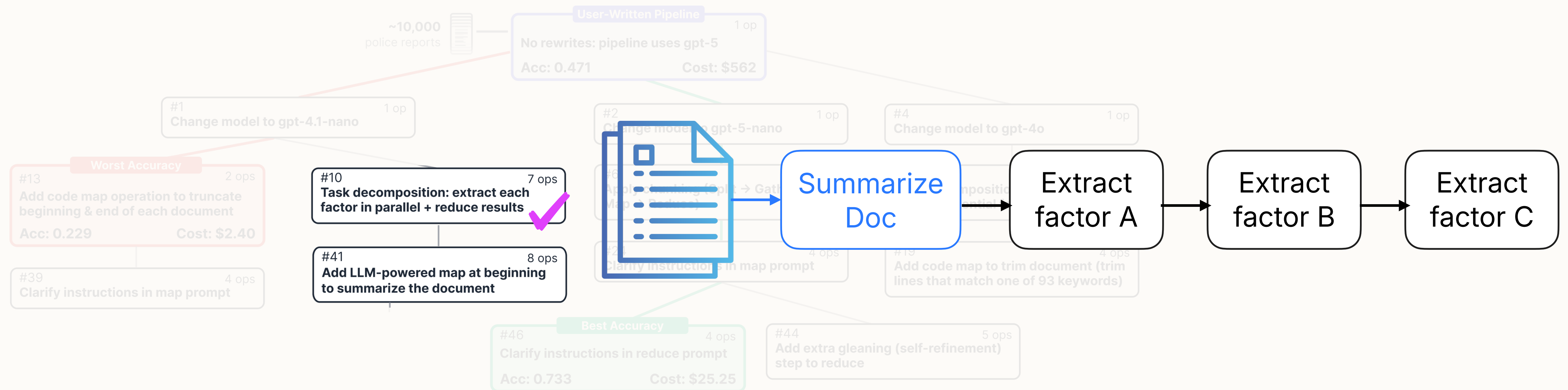


Search over Rewrite Directives



- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).

Search over Rewrite Directives



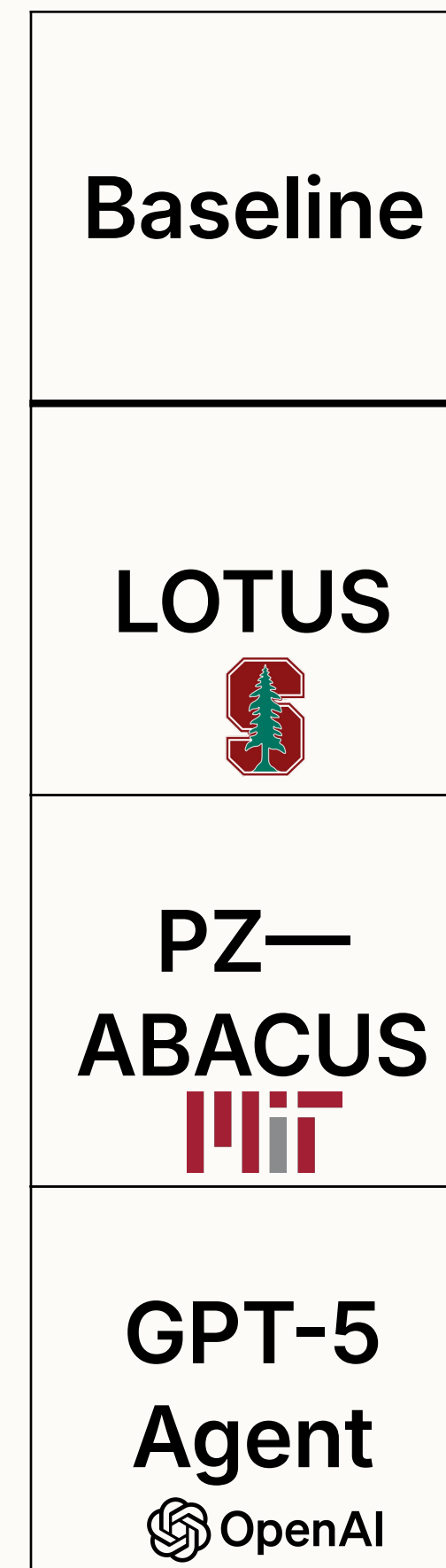
- LLM agent decides what directive to apply and creates the new operators (e.g., prompts, code, etc).
- We evaluate the new plan on a sample, recording cost and accuracy.

Results: Agentic Query Optimization

- ◆ 6 workloads (legal, game reviews, biomedical, gov reports, clinical, enterprise reports)
- ◆ 3 Baselines
- ◆ Systems could choose from 11 models
- ◆ We produce the most accurate plan *every time*




Results: Agentic Query Optimization

- ◆ 6 workloads (legal, game reviews, biomedical, gov reports, clinical, enterprise reports)
- ◆ 3 Baselines
- ◆ Systems could choose from 11 models
- ◆ We produce the most accurate plan *every time*






Results: Agentic Query Optimization

- ◆ 6 workloads (legal, game reviews, biomedical, gov reports, clinical, enterprise reports)
- ◆ 3 Baselines
- ◆ Systems could choose from 11 models
- ◆ We produce the most accurate plan *every time*

Baseline	Our accuracy (@ baseline max accuracy)
LOTUS 	2.10×
PZ— ABACUS 	1.38×
GPT-5 Agent  OpenAI	1.95×

Results: Agentic Query Optimization

- ◆ 6 workloads (legal, game reviews, biomedical, gov reports, clinical, enterprise reports)
- ◆ 3 Baselines
- ◆ Systems could choose from 11 models
- ◆ We produce the most accurate plan *every time*

Baseline	Our accuracy (@ baseline max accuracy)	Our cost (@ baseline max accuracy)
LOTUS 	2.10×	0.487×
PZ— ABACUS 	1.38×	0.575×
GPT-5 Agent  OpenAI	1.95×	0.375×

Challenges in Agentic Query Optimization

Challenges in Agentic Query Optimization

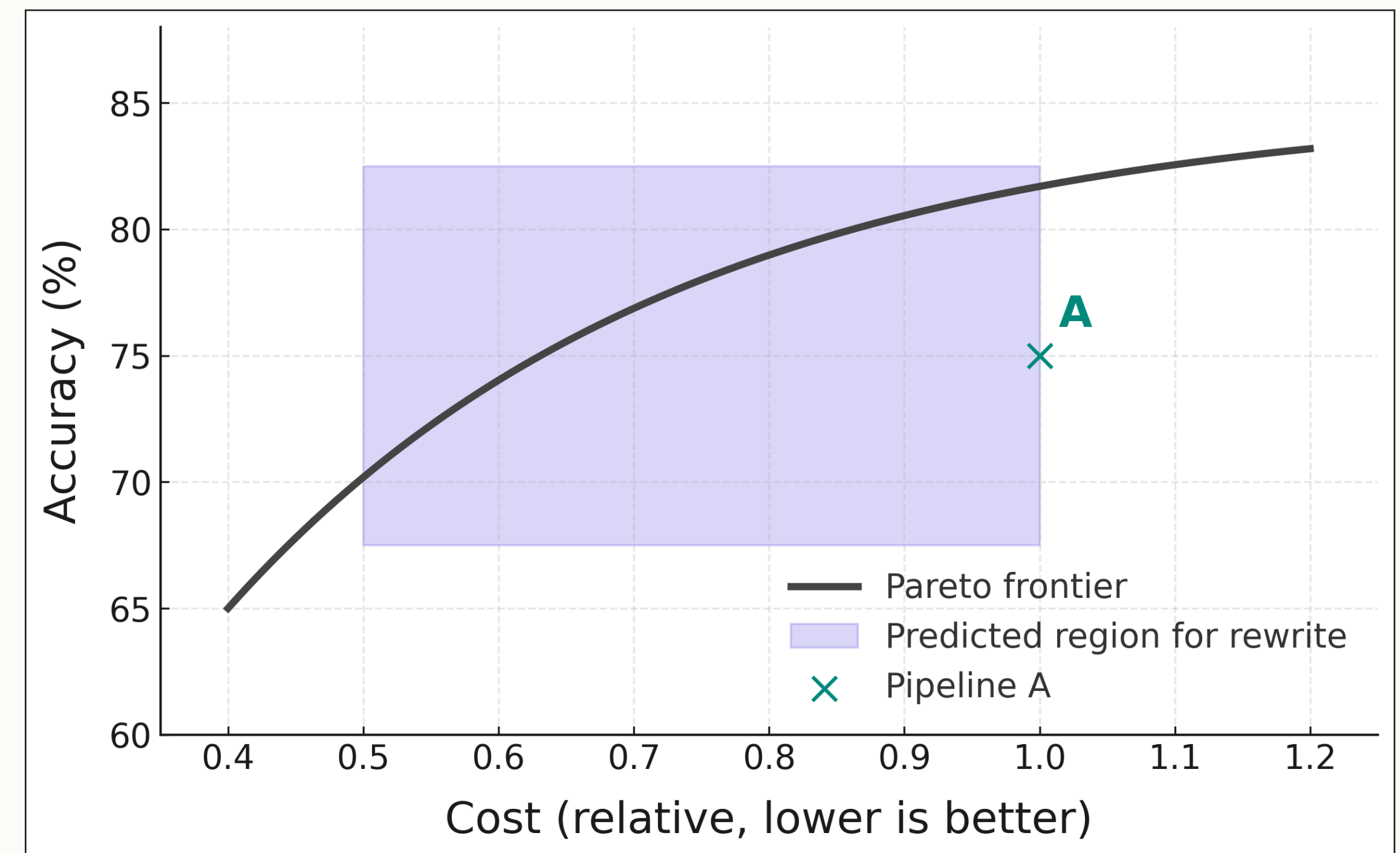
- ◆ Each optimization run takes 1–5 hours!
- ◆ We have to run plans on samples to estimate accuracy

Challenges in Agentic Query Optimization

- ◆ Each optimization run takes 1–5 hours!
- ◆ We have to run plans on samples to estimate accuracy

New question:

Can we design rewrites that guarantee the new plan stays within a target accuracy of the current plan?



Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Key idea: reduce total cost by routing inputs through a sequence of models.

◆ *Used extensively in computer vision; recently applied to LLMs.*

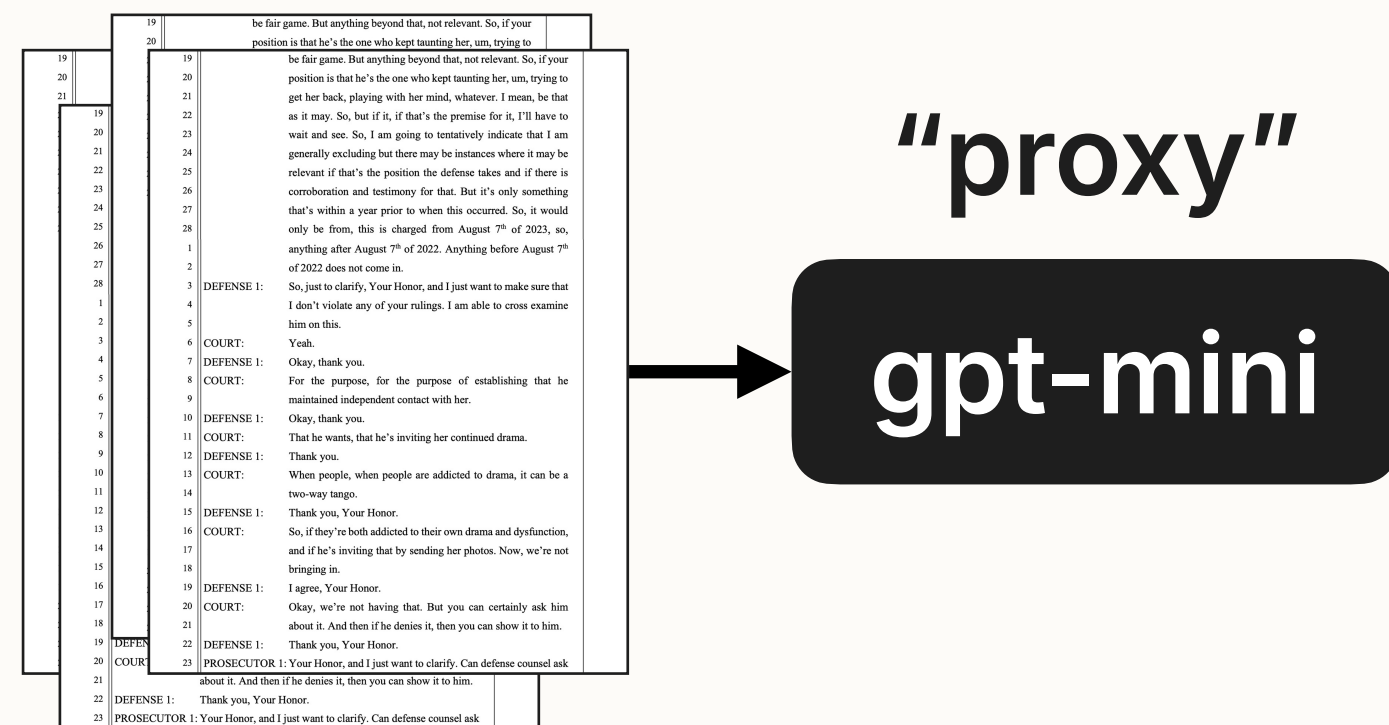
		19	be fair game. But anything beyond that, not relevant. So, if your	
		20	position is that he's the one who kept teasing her, um, trying to	
19		19	be fair game. But anything beyond that, not relevant. So, if your	
20		20	position is that he's the one who kept teasing her, um, trying to	
21		21	get her back, playing with her mind, whatever. I mean, be that	
	19	22	as it may. So, but if it, if that's the premise for it, I'll have to	
	20	23	wait and see. So, I am going to tentatively indicate that I am	
	21	24	generally excluding but there may be instances where it may be	
	22	25	relevant if that's the position the defense takes and if there is	
	23	26	corroboration and testimony for that. But it's only something	
	24	27	that's within a year prior to when this occurred. So, it would	
	25	28	only be then, this is charged from August 7 th of 2022, so,	
	26	1	anything after August 7 th of 2022. Anything before August 7 th	
	27	2	of 2022 does not come in.	
	28	3	DEFENSE 1: So, just to clarify, Your Honor, and I just want to make sure that	
1		4	I don't violate any of your rulings. I am able to cross examine	
2		5	him on this.	
3		6	COURT: Yeah.	
4		7	DEFENSE 1: Okay, thank you.	
5		8	COURT: For the purpose, for the purpose of establishing that he	
6		9	maintained independent contact with her.	
7		10	DEFENSE 1: Okay, thank you.	
8		11	COURT: That he wants, that he's inviting her continued drama.	
9		12	DEFENSE 1: Thank you.	
10		13	COURT: When people, when people are addicted to drama, it can be a	
11		14	two-way thing.	
12		15	DEFENSE 1: Thank you, Your Honor.	
13		16	COURT: So, if they're both addicted to their own drama and dysfunction,	
14		17	and if he's inviting that by sending her photos. Now, we're not	
15		18	bringing in.	
16		19	DEFENSE 1: I agree, Your Honor.	
17		20	COURT: Okay, we're not having that. But you can certainly ask him	
18		21	about it. And then if he denies it, then you can show it to him.	
19	DEFENSE 1:	22	Thank you, Your Honor.	
20	COURT:	23	PROSECUTOR 1: Your Honor, and I just want to clarify. Can defense counsel ask	
21			about it. And then if he denies it, then you can show it to him.	
22	DEFENSE 1:		Thank you, Your Honor.	
23	PROSECUTOR 1:		Your Honor, and I just want to clarify. Can defense counsel ask	

Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

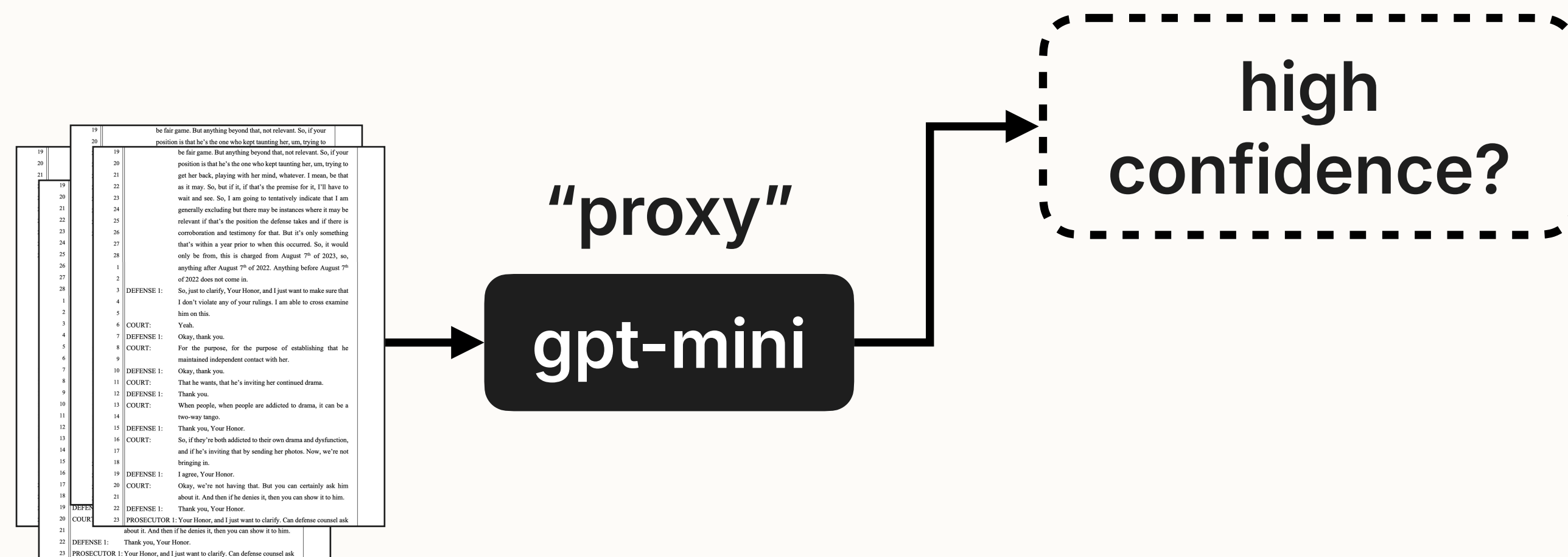
♦ *Used extensively in computer vision; recently applied to LLMs.*



Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.
♦ *Used extensively in computer vision; recently applied to LLMs.*

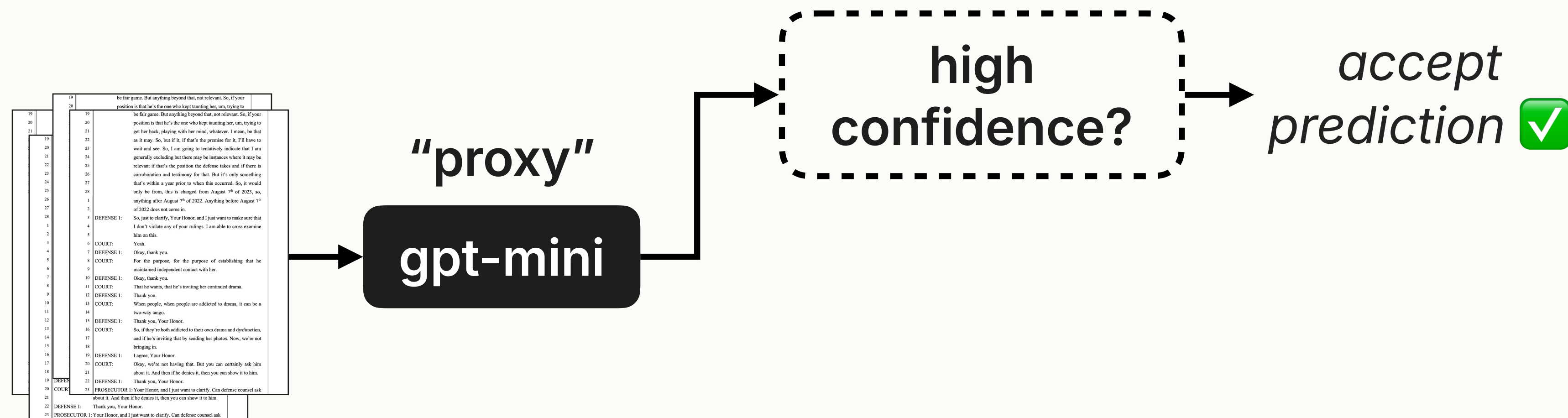


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

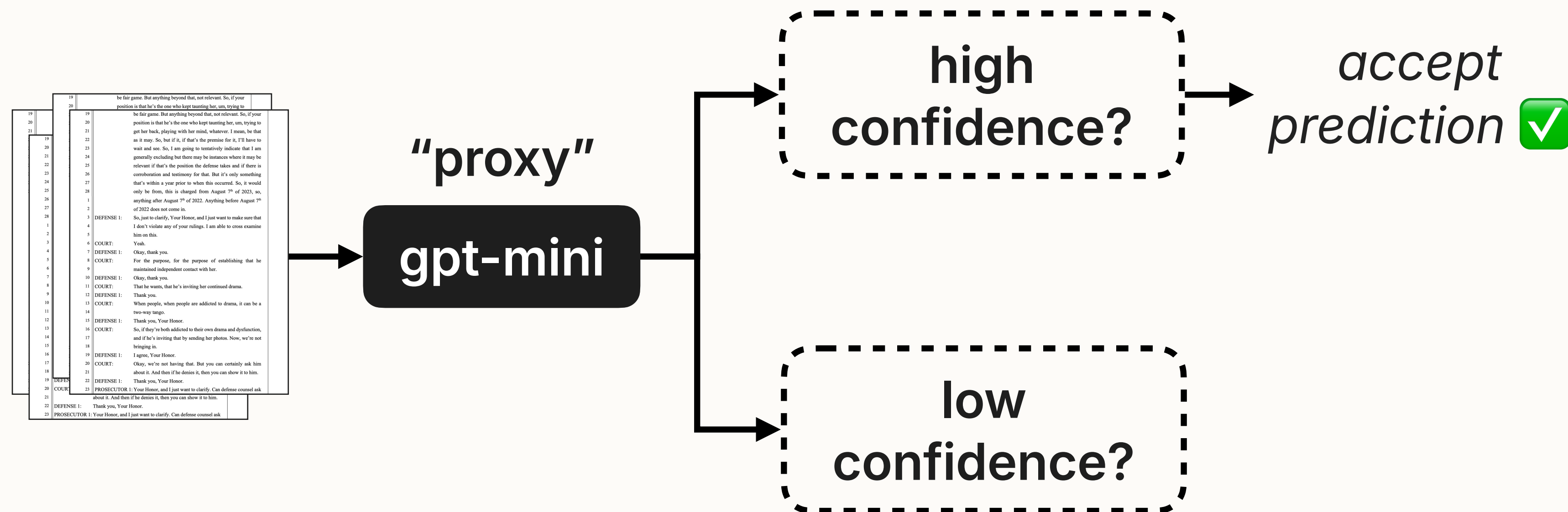


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

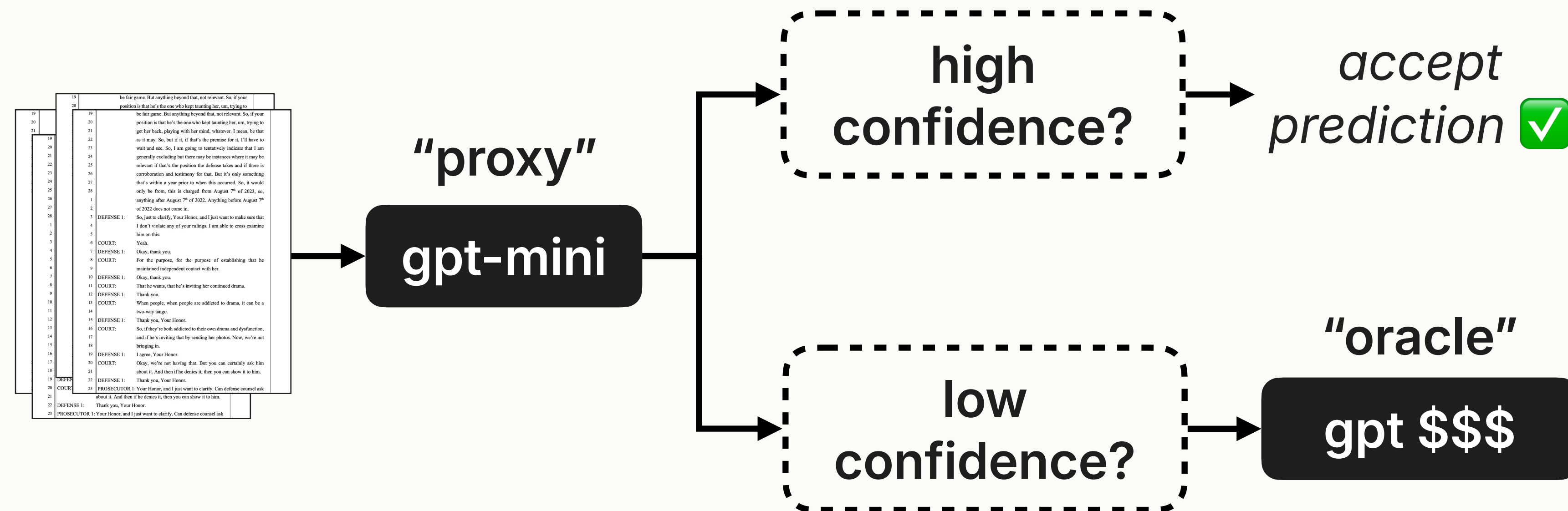


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

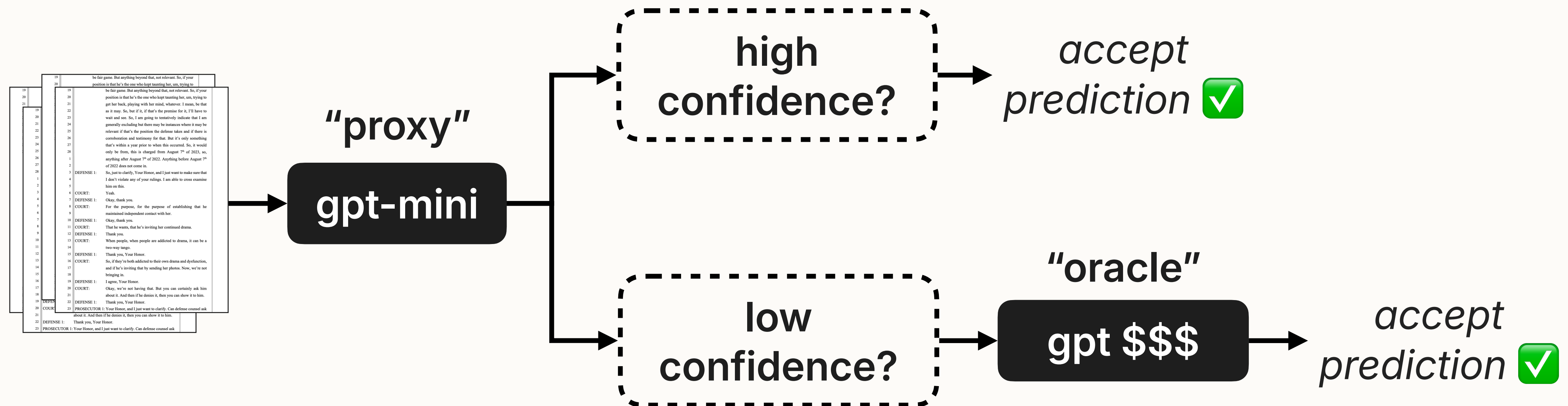


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

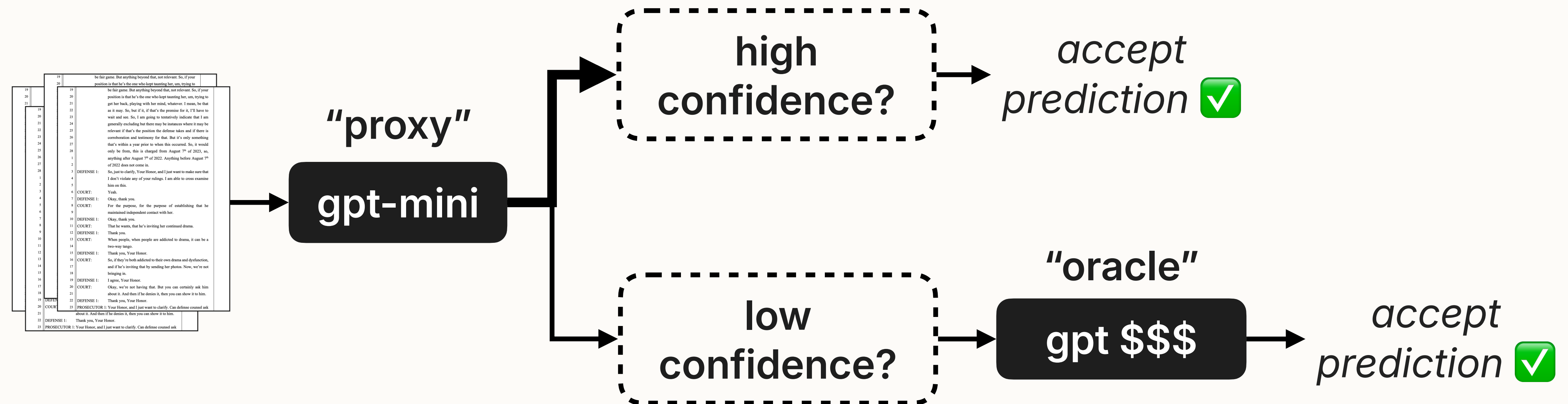


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

♦ *Used extensively in computer vision; recently applied to LLMs.*

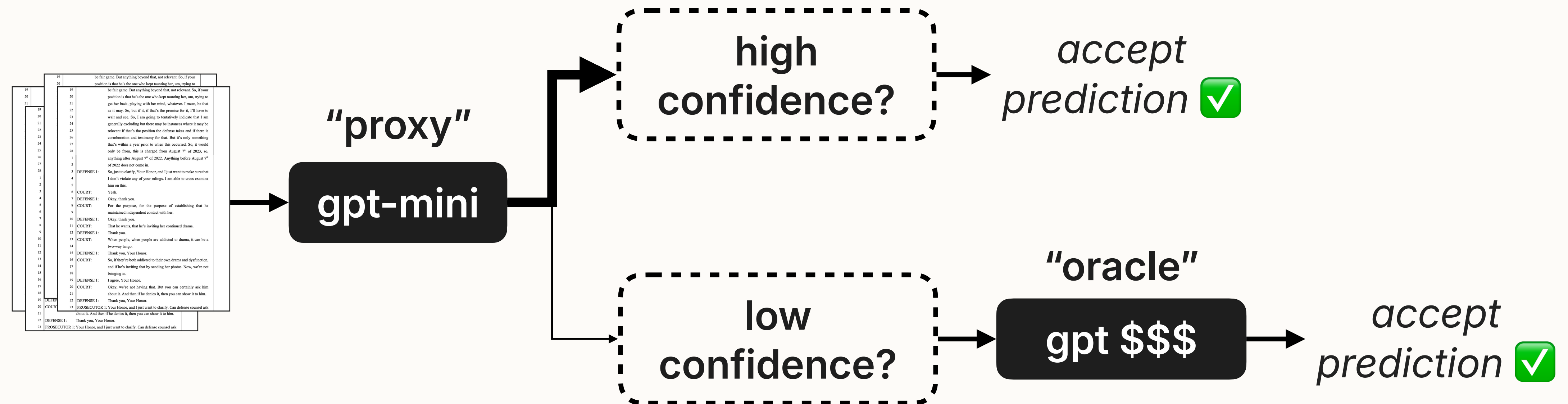


Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.

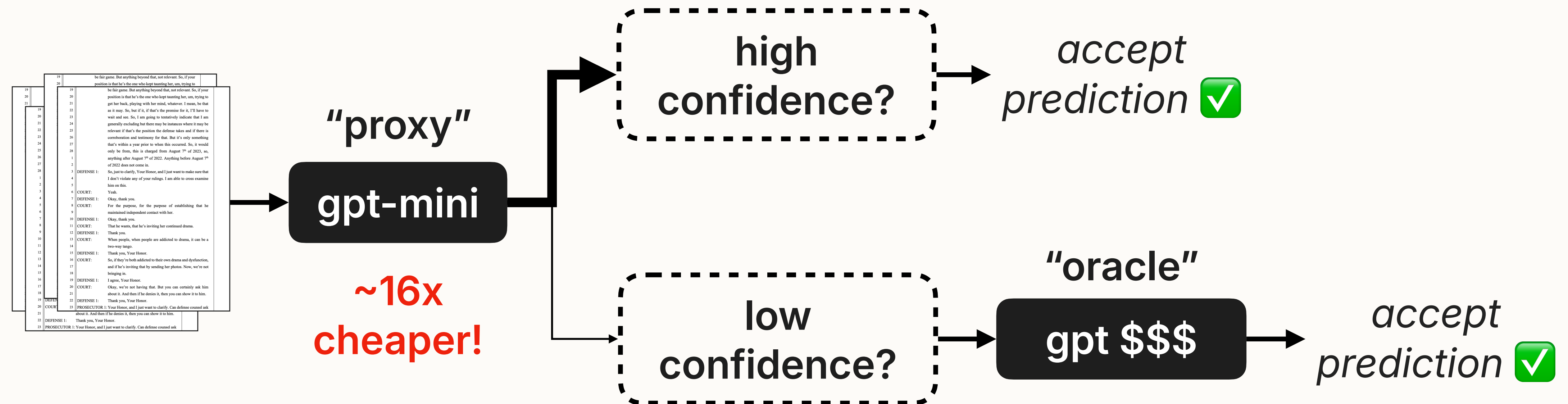
♦ *Used extensively in computer vision; recently applied to LLMs.*



Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

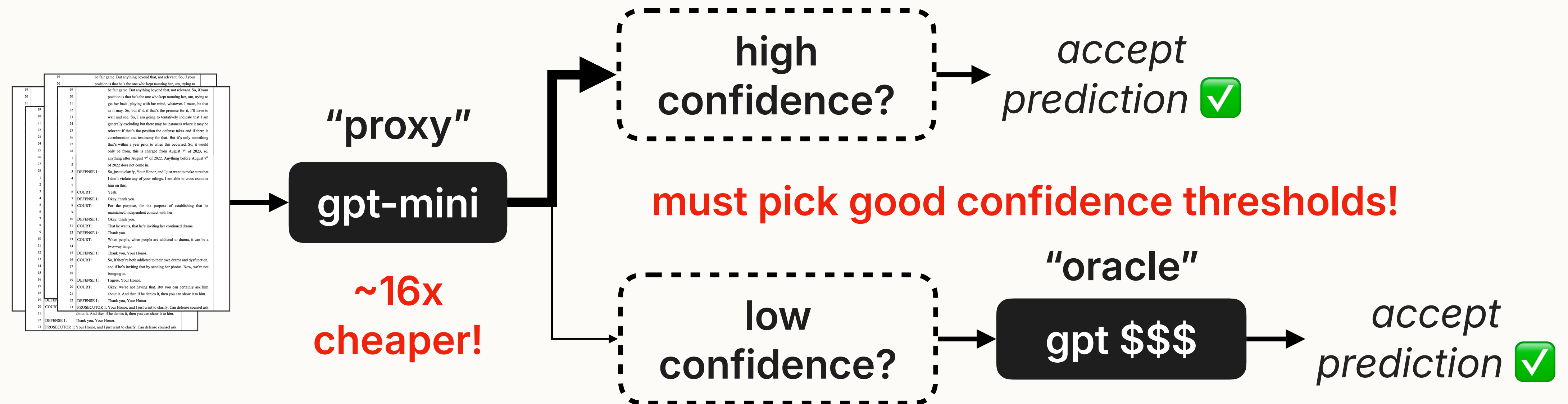
Key idea: reduce total cost by routing inputs through a sequence of models.
♦ *Used extensively in computer vision; recently applied to LLMs.*



Background: Model Cascade

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Key idea: reduce total cost by routing inputs through a sequence of models.
♦ *Used extensively in computer vision; recently applied to LLMs.*



Generalizing to Task Cascades

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Generalizing to Task Cascades

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Three factors of cost:


- ♦ **Model:** e.g., gpt-5-mini vs gpt-5
- ♦ **Data:** how much of the document is stuffed in the LLM call
- ♦ **Operation:** how complex the prompt or reasoning is

Generalizing to Task Cascades

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Three factors of cost:

- ♦ **Model:** e.g., gpt-5-mini vs gpt-5
- ♦ **Data:** how much of the document is stuffed in the LLM call
- ♦ **Operation:** how complex the prompt or reasoning is




A *task*:
(Model,
document
slice,
operation)

Generalizing to Task Cascades

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Three factors of cost:

- ♦ **Model:** e.g., gpt-5-mini vs gpt-5
- ♦ **Data:** how much of the document is stuffed in the LLM call
- ♦ **Operation:** how complex the prompt or reasoning is



A *task*:
(Model,
document
slice,
operation)

Rewrite directive insight: create *simpler* or *cheaper* versions of the task that are still predictive.

Generalizing to Task Cascades

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Three factors of cost:

- ♦ **Model:** e.g., gpt-5-mini vs gpt-5
- ♦ **Data:** how much of the document is stuffed in the LLM call
- ♦ **Operation:** how complex the prompt or reasoning is

A *task*:
(Model,
document
slice,
operation)

Rewrite directive insight: create *simpler* or *cheaper* versions of the task that are still predictive.

MLSys insight: run cheaper tasks on *smaller* pieces of the document and reuse computation as you scale up.

The Task Cascade Idea

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Operation o*:

*Does this
opinion
overturn a
lower court
decision?*

Document



The Task Cascade Idea

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Operation o^* :

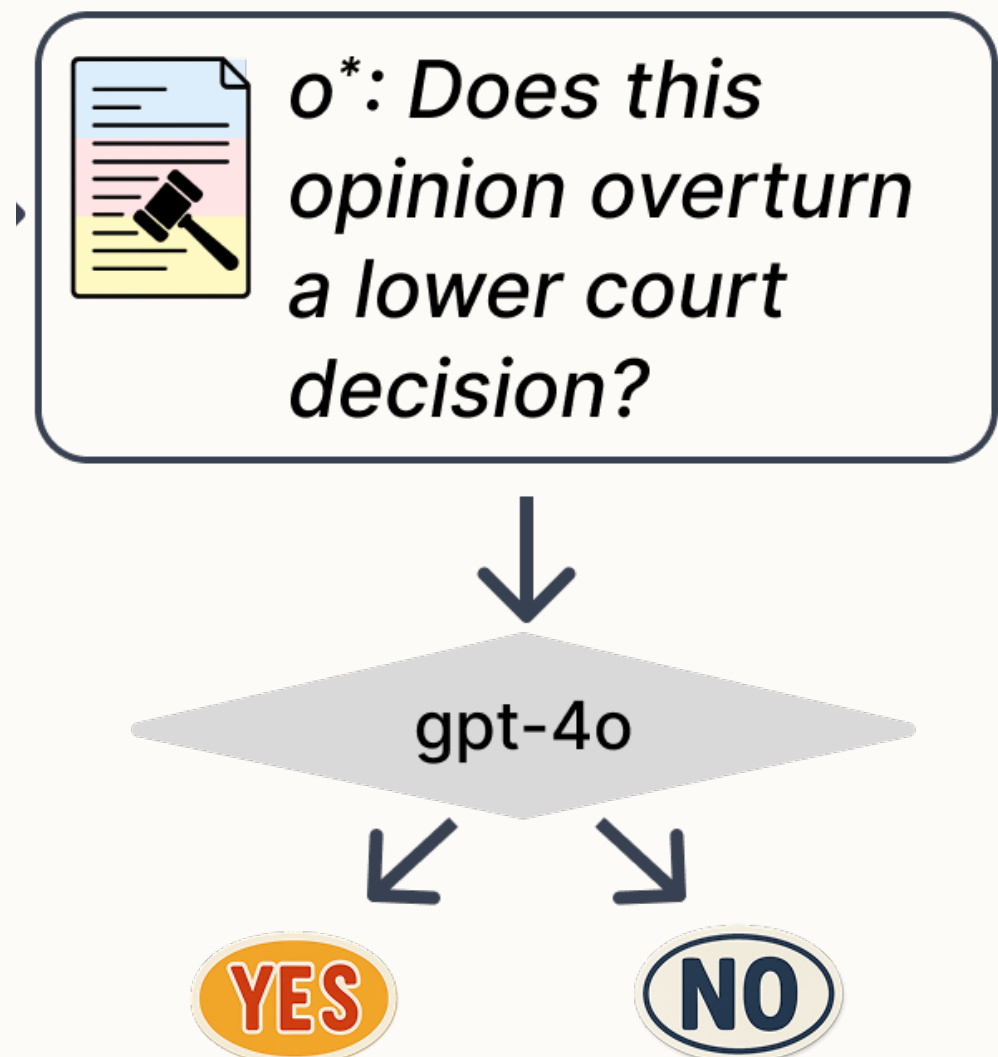
Does this opinion overturn a lower court decision?

Document



Oracle Task

$f=1.0$; $m=\text{gpt-4o}$; $o=o^*$

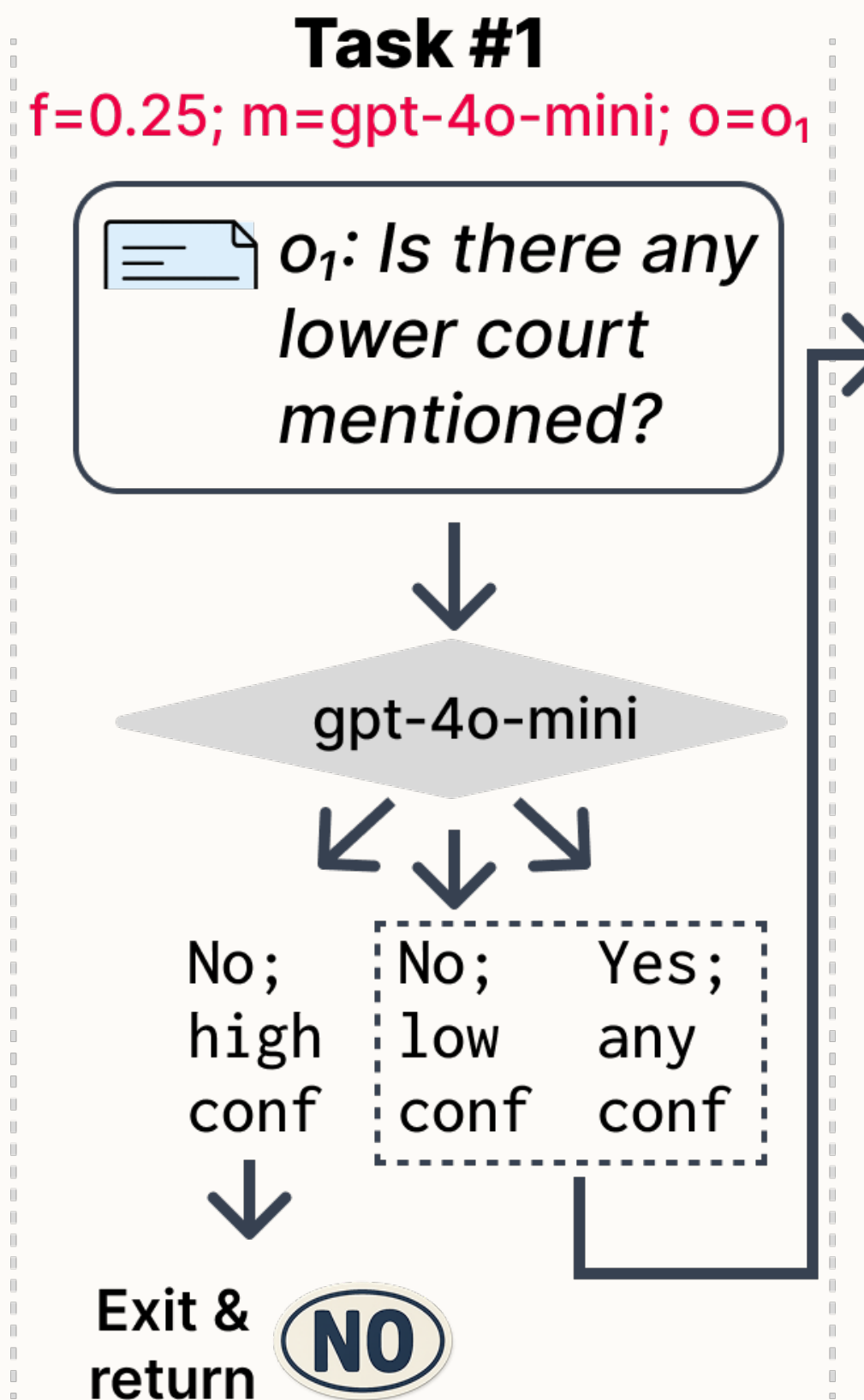


The Task Cascade Idea

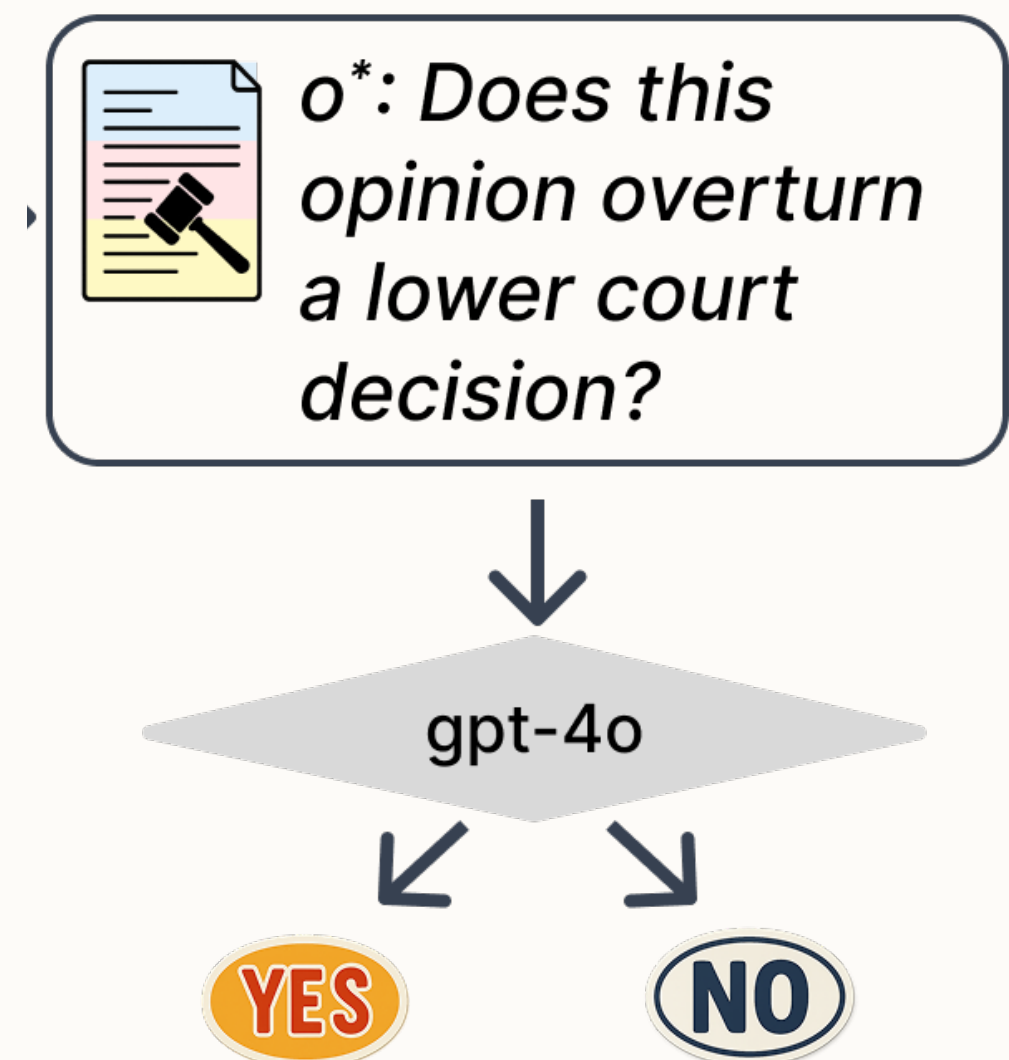
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Operation o^* :
Does this
opinion
overturn
a lower court
decision?

Document



Oracle Task
 $f=1.0$; $m=\text{gpt-4o}$; $o=o^*$

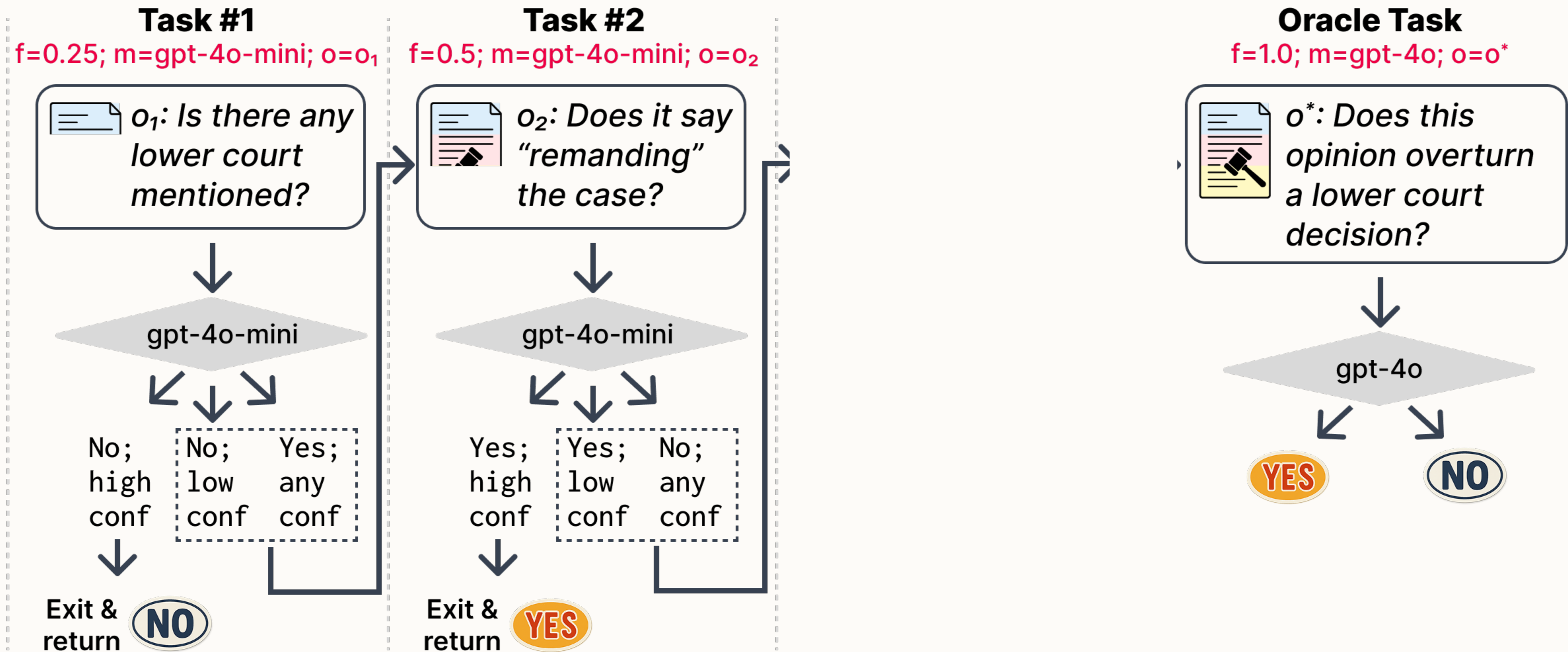


The Task Cascade Idea

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Operation o^* :
Does this opinion overturn a lower court decision?

Document

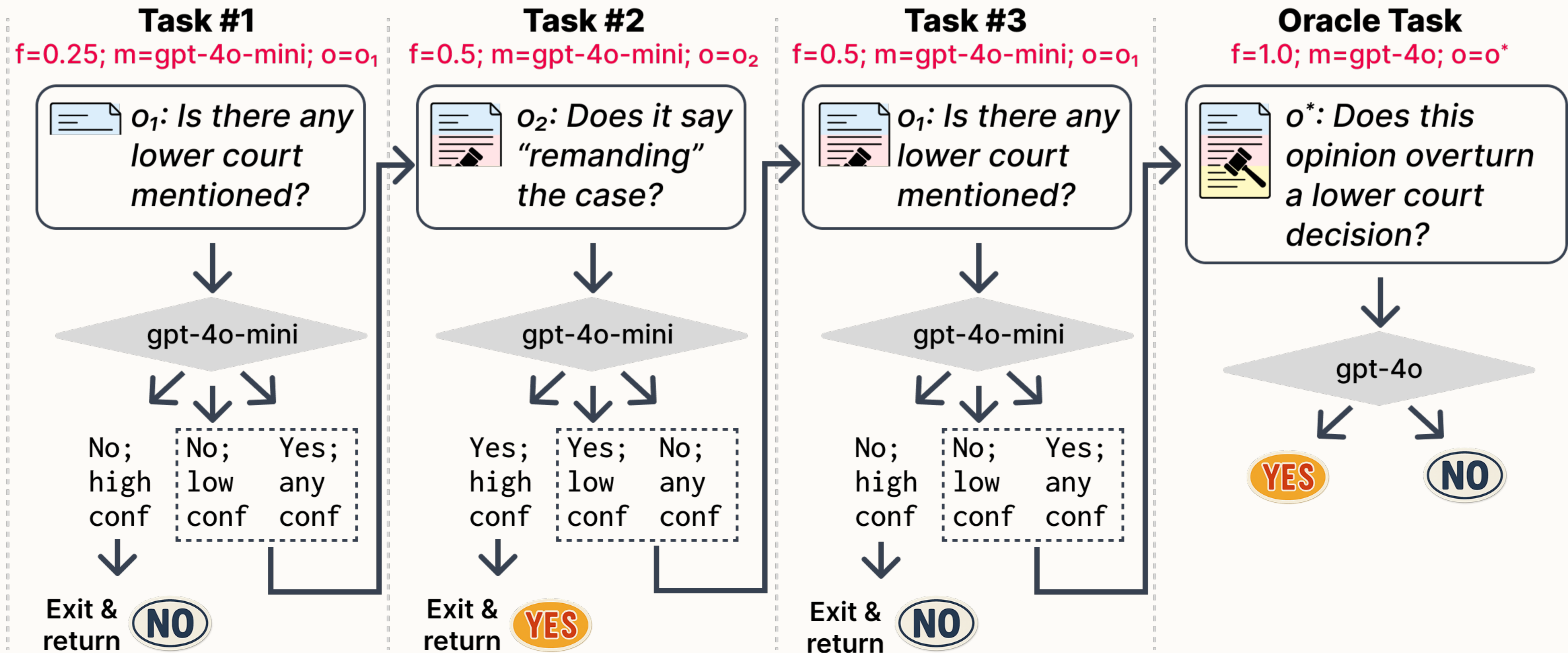


The Task Cascade Idea

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Operation o^* :
Does this opinion overturn a lower court decision?

Document



The Task Cascade Rewrite

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

`map` \Rightarrow `map*` \rightarrow `map`

`filter` \Rightarrow `map*` \rightarrow `filter`

Procedure:

1. Reorder documents for cheaper partial reads.
2. Generate lots of *surrogate* `map` operations.
3. Select tasks, thresholds, and order for the cascade.

The Task Cascade Rewrite

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

`map` \Rightarrow `map*` \rightarrow `map`

`filter` \Rightarrow `map*` \rightarrow `filter`

Procedure:

1. Reorder documents for cheaper partial reads.
2. Generate lots of *surrogate* `map` operations.
3. Select tasks, thresholds, and order for the cascade.

We show that:

Constructing an optimal task cascade is NP-Hard (in the task space).

The Task Cascade Rewrite

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

`map` \Rightarrow `map*` \rightarrow `map`

`filter` \Rightarrow `map*` \rightarrow `filter`

Procedure:

1. Reorder documents for cheaper partial reads.
2. Generate lots of *surrogate* `map` operations.
3. Select tasks, thresholds, and order for the cascade.

We show that:

Constructing an optimal task cascade is NP-Hard (in the task space).

We therefore do (3) with a greedy algorithm.

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Goal: Select cascade thresholds to minimize oracle model usage while guaranteeing accuracy $\geq T$ (relative to the oracle model).

We can get accuracy guarantees of the following form:

$$\Pr(A(\hat{Y}) \geq T) \geq 1 - \delta$$

where A is the accuracy function, \hat{Y} represent cascade outputs, and δ is a user-defined *failure rate*.

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Goal: choose a confidence threshold ρ that meets an accuracy target T with failure probability δ , using n sample documents and candidate threshold set C .

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Goal: choose a confidence threshold ρ that meets an accuracy target T with failure probability δ , using n sample documents and candidate threshold set C .

Statistic: for each document i , let $Z_i = \mathbf{1}[\hat{Y}_i(\rho) = Y_i]$, $\hat{A}(\rho) = \frac{1}{n} \sum_{i=1}^n Z_i$

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Goal: choose a confidence threshold ρ that meets an accuracy target T with failure probability δ , using n sample documents and candidate threshold set C .

Statistic: for each document i , let $Z_i = \mathbf{1}[\hat{Y}_i(\rho) = Y_i]$, $\hat{A}(\rho) = \frac{1}{n} \sum_{i=1}^n Z_i$

Need an estimator with bounded false positive rate: $\Pr(\hat{A}(\rho) - A(\rho) \geq \varepsilon) \leq \delta$

Example decision rule (Hoeffding): certify ρ if $\hat{A}(\rho) = T + \sqrt{\frac{\log(|C|/\delta)}{2n}}$

Final threshold: $\rho^* = \min\{\rho \in C : \rho \text{ is certified}\}$

Task Cascades with Guarantees

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

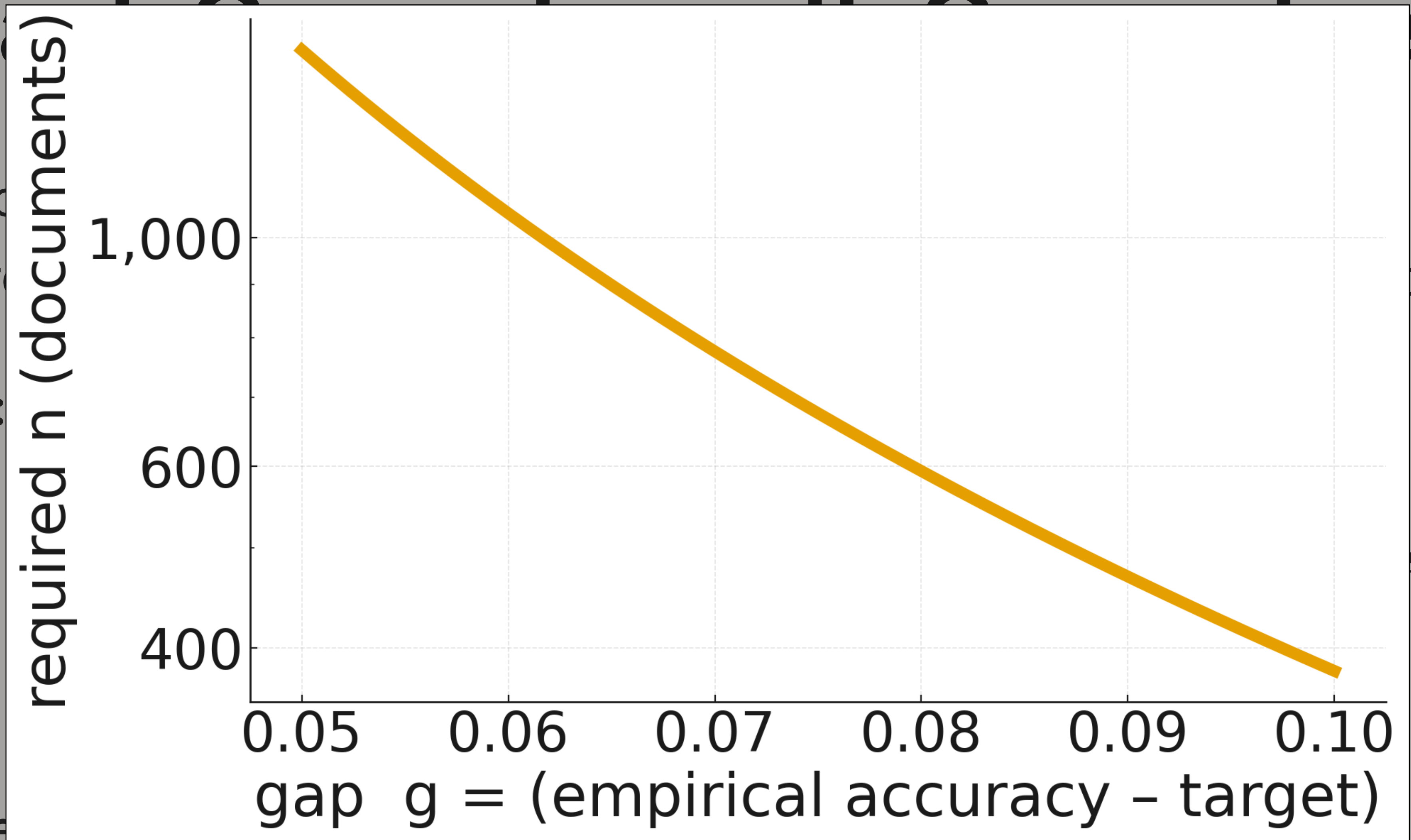
Goal: choose a confidence threshold ρ that meets an accuracy target T with failure probability δ , using n sample documents and candidate threshold set C .

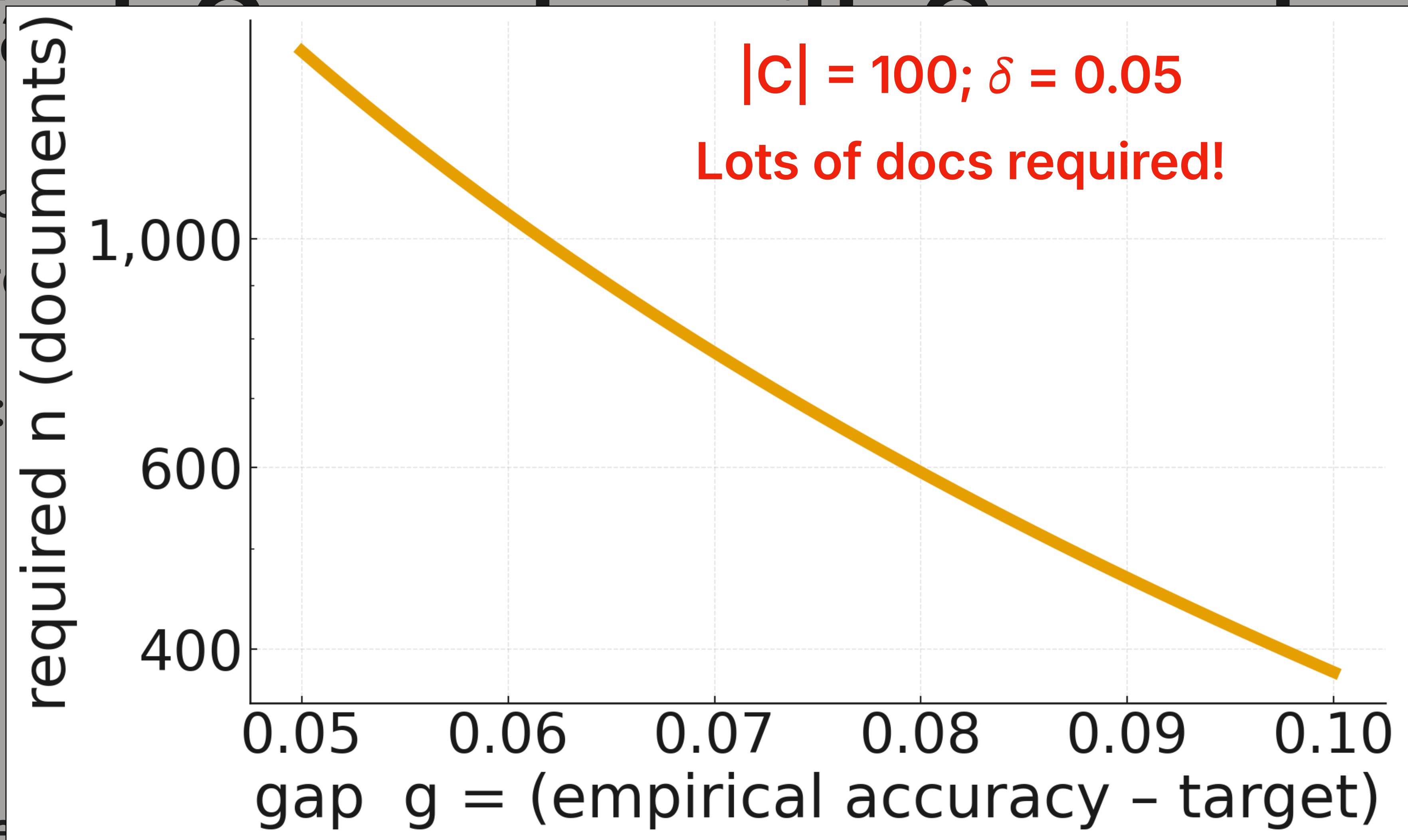
Statistic: for each document i , let $Z_i = \mathbf{1}[\hat{Y}_i(\rho) = Y_i]$, $\hat{A}(\rho) = \frac{1}{n} \sum_{i=1}^n Z_i$

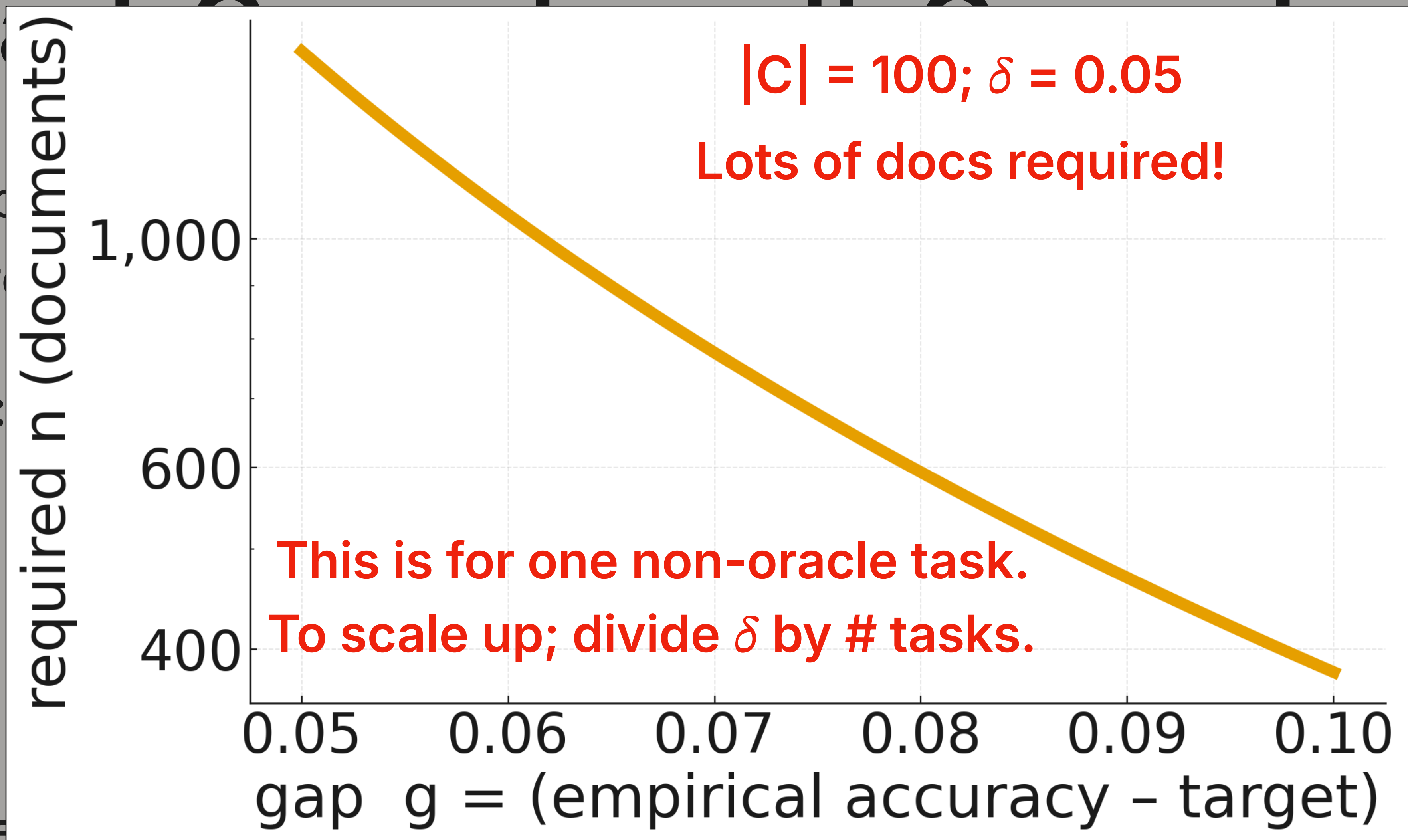
Need an estimator with bounded false positive rate: $\Pr(\hat{A}(\rho) - A(\rho) \geq \varepsilon) \leq \delta$

Example decision rule (Hoeffding): certify ρ if $\hat{A}(\rho) = T + \sqrt{\frac{\log(|C|/\delta)}{2n}}$

Final threshold: $\rho^* = \min\{\rho \in C : \rho \text{ is certified}\}$







Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds
 C , estimator E .

Problem: Number of docs
required grows with $\log |C|!$

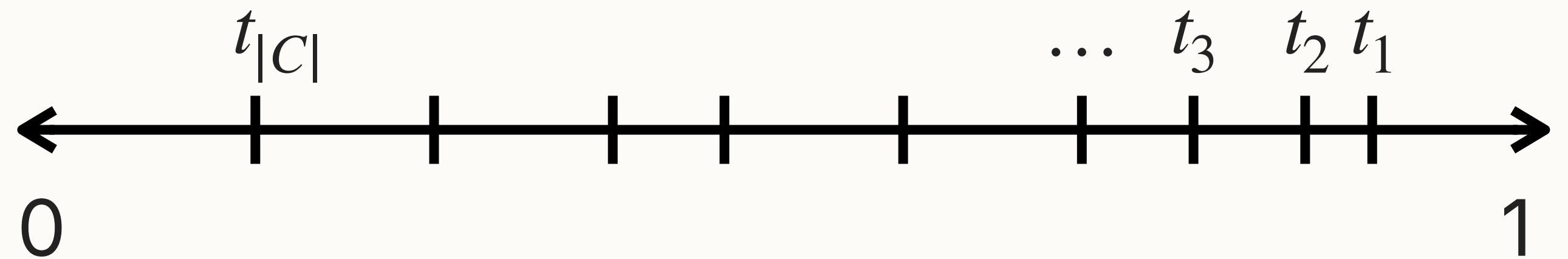
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|$!

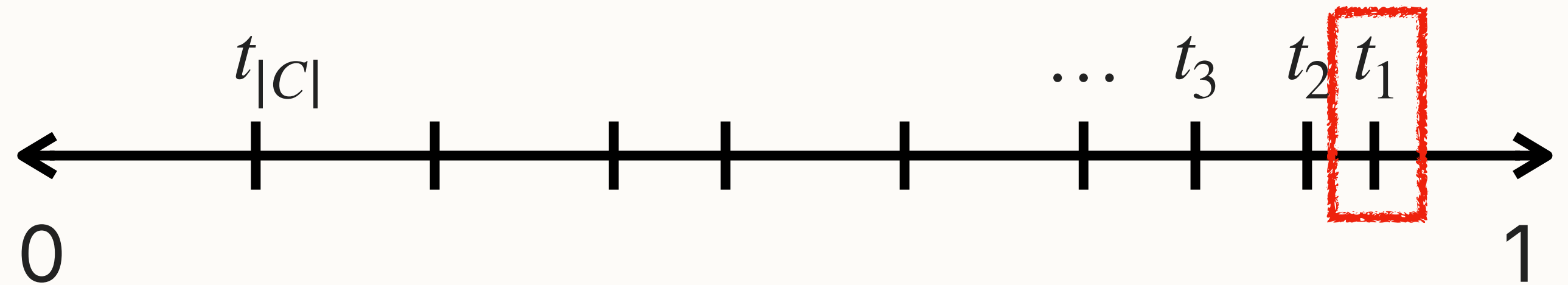
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|$!

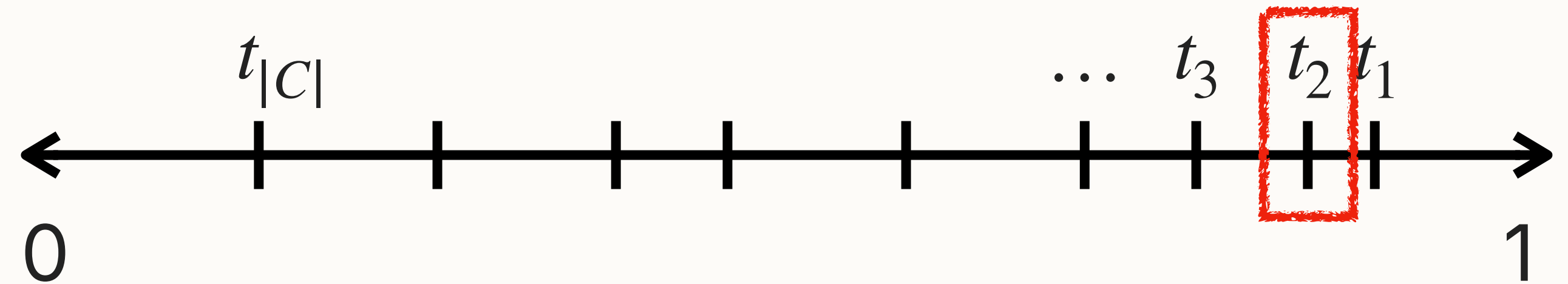
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|$!

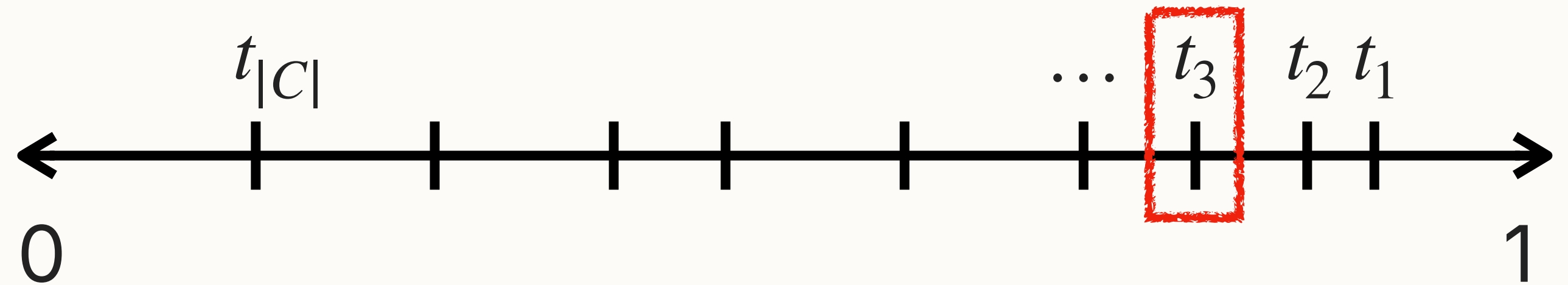
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|!$

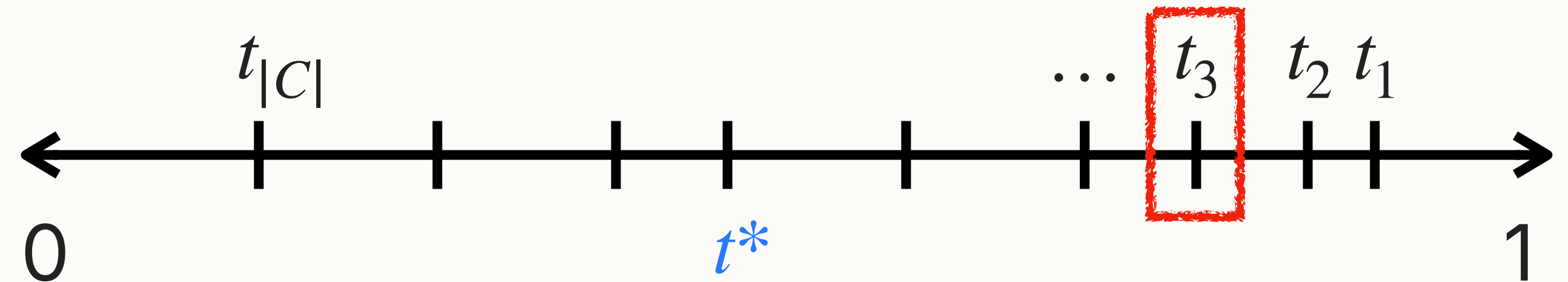
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|$!

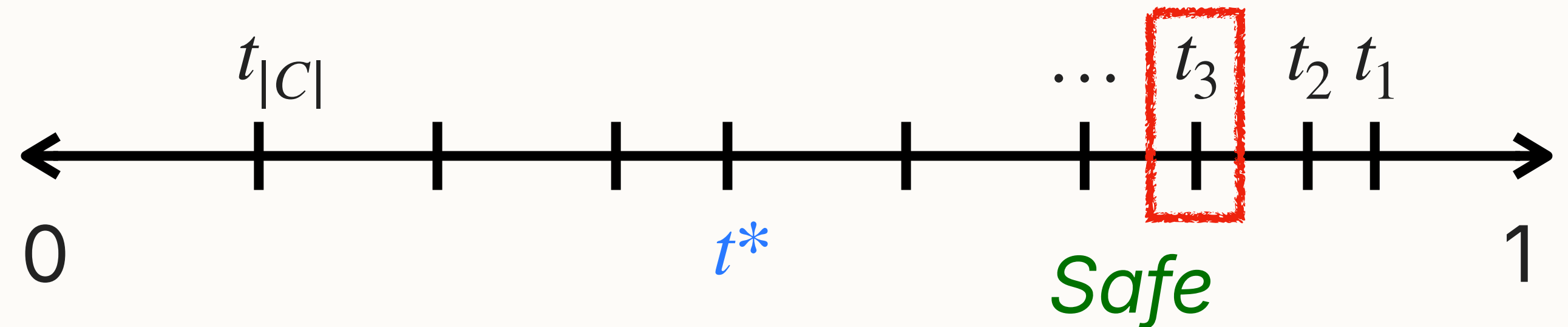
Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds C , estimator E .



Problem: Number of docs required grows with $\log |C|!$

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$

Sequential Test for Selecting Safe Thresholds

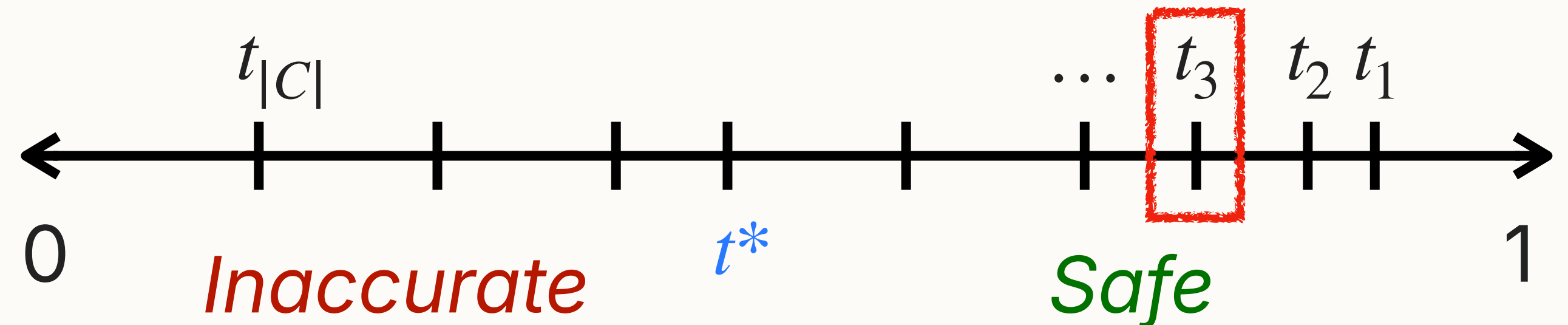
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds C , estimator E .

Problem: Number of docs required grows with $\log |C|$!

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$



Sequential Test for Selecting Safe Thresholds

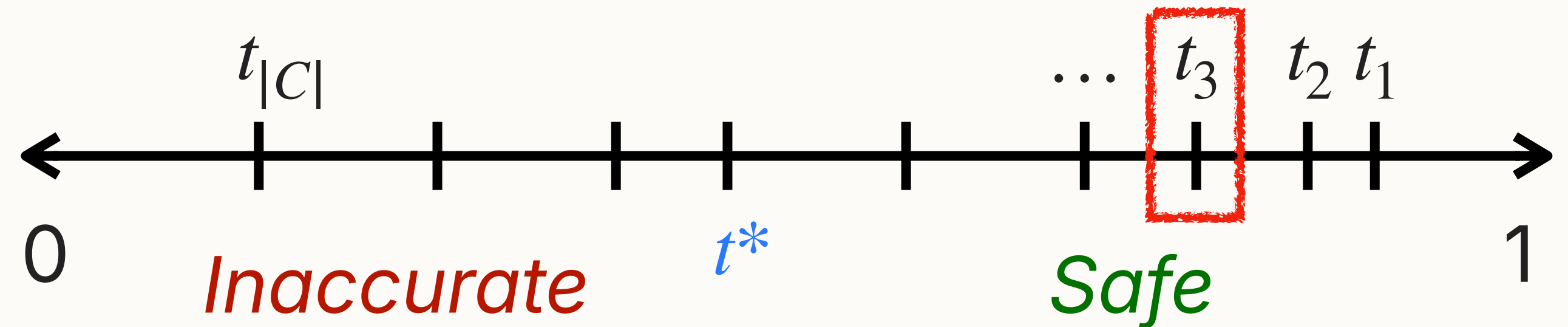
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .

Problem: Number of docs required grows with $\log |C|$!

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$



Proof that $\Pr(A(t_{\text{return}}) < T) \leq \delta$:

Sequential Test for Selecting Safe Thresholds

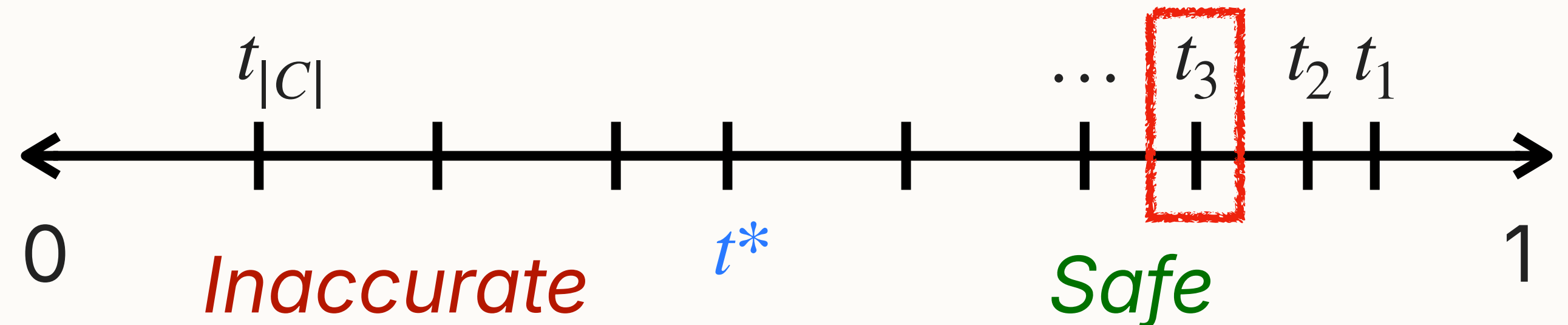
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds C , estimator E .

Problem: Number of docs required grows with $\log |C|$!

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$



Proof that $\Pr(A(t_{\text{return}}) < T) \leq \delta$:

♦ If $t_{\text{return}} > t_i^*$, then t_{return} is still in the "safe" zone.

Sequential Test for Selecting Safe Thresholds

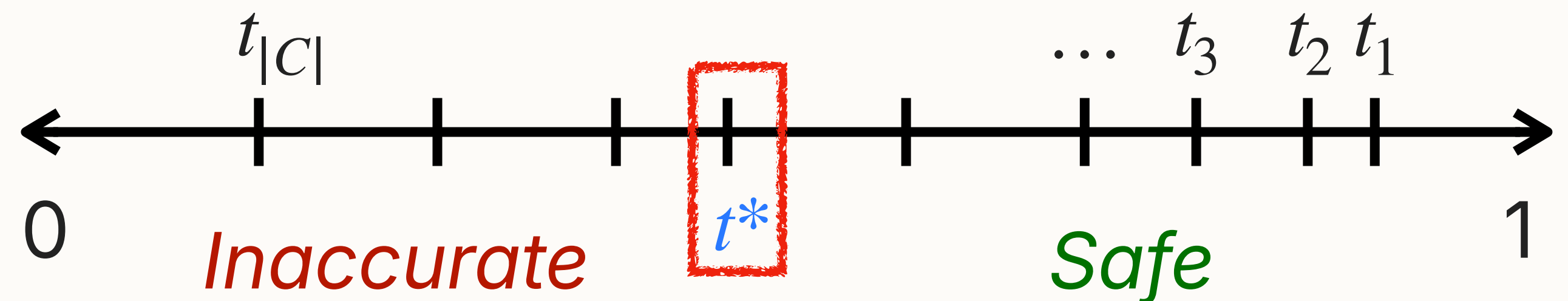
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Given: Confidence thresholds C , estimator E .

Problem: Number of docs required grows with $\log |C|$!

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$



Proof that $\Pr(A(t_{\text{return}}) < T) \leq \delta$:

- ◆ If $t_{\text{return}} > t_i^*$, then t_{return} is still in the "safe" zone.
- ◆ If $E(t_i^*) = 0$, the algorithm is correct.

Sequential Test for Selecting Safe Thresholds

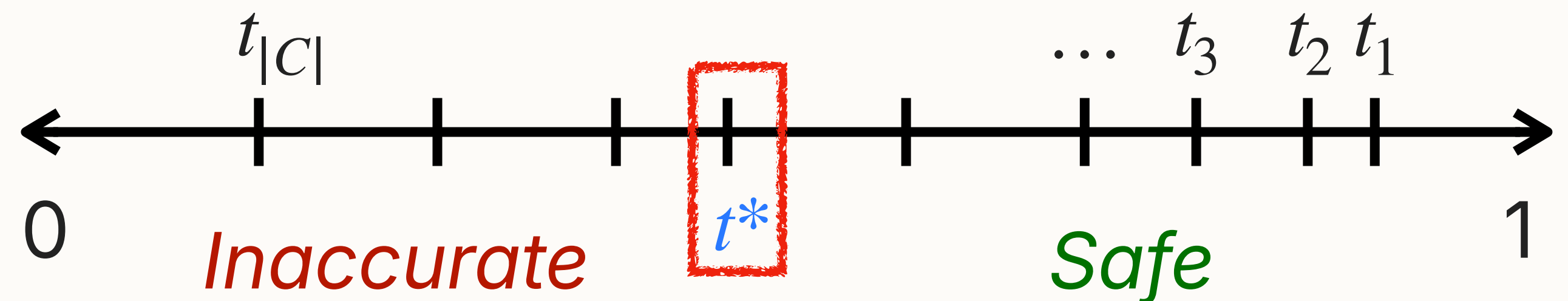
Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26.*

Given: Confidence thresholds C , estimator E .

Problem: Number of docs required grows with $\log |C|$!

Solution: Sequential procedure.

1. Start at t_1
2. While $E(t_i) = 1$, move to t_{i+1}
3. Stop at first t_j with $E(t_j) = 0$
4. $t_{\text{return}} \leftarrow t_{j-1}$



Proof that $\Pr(A(t_{\text{return}}) < T) \leq \delta$:

- ♦ If $t_{\text{return}} > t_i^*$, then t_{return} is still in the "safe" zone.
- ♦ If $E(t_i^*) = 0$, the algorithm is correct.
- ♦ If $E(t_i^*) = 1$, then
 $\Pr(E(t_i^*) = 1) \leq \delta \implies \Pr(A(t_{\text{return}}) < T) \leq \delta$.

Task Cascades: Recap & Results

Task Cascades. **Shankar**, Zeighami, Parameswaran. *Under Revision at SIGMOD '26*.

Motivation: Rewrite directives expose many execution strategies, but many are low-accuracy and evaluating accuracy is expensive.

Task Cascades—enabled by new algorithmic tools—safely replace expensive operators with cheaper ones, while guaranteeing accuracy.

◆ $\text{map} \Rightarrow \text{map}^* \rightarrow \text{map}$

◆ $\text{filter} \Rightarrow \text{map}^* \rightarrow \text{filter}$

Result across 8 real-world workloads: 48.5% cheaper than model cascade baselines and 86% cheaper than operators pre-rewrite.

Recap: Optimizing Semantic Data Pipelines

Recap: Optimizing Semantic Data Pipelines

For unstructured data and semantic operators:

♦ **Transformations:** semantic *rewrite directives* (not syntactic rules)

Recap: Optimizing Semantic Data Pipelines

For unstructured data and semantic operators:

- ♦ **Transformations:** semantic *rewrite directives* (not syntactic rules)
- ♦ **Search:**
 - * Complete plans (not necessarily subplans)
 - * Learned heuristics (UCT) + LLM agents
 - * Statistical tools to estimate whether rewrites will be helpful

Recap: Optimizing Semantic Data Pipelines

For unstructured data and semantic operators:

- ♦ **Transformations:** semantic *rewrite directives* (not syntactic rules)
- ♦ **Search:**
 - * Complete plans (not necessarily subplans)
 - * Learned heuristics (UCT) + LLM agents
 - * Statistical tools to estimate whether rewrites will be helpful

**How can data systems
reason over
unstructured data?**

Recap: Optimizing Semantic Data Pipelines

For unstructured data and semantic operators:

- ♦ **Transformations:** semantic *rewrite directives* (not syntactic rules)
- ♦ **Search:**
 - * Complete plans (not necessarily subplans)
 - * Learned heuristics (UCT) + LLM agents
 - * Statistical tools to estimate whether rewrites will be helpful

**How can data systems
reason over
unstructured data?**



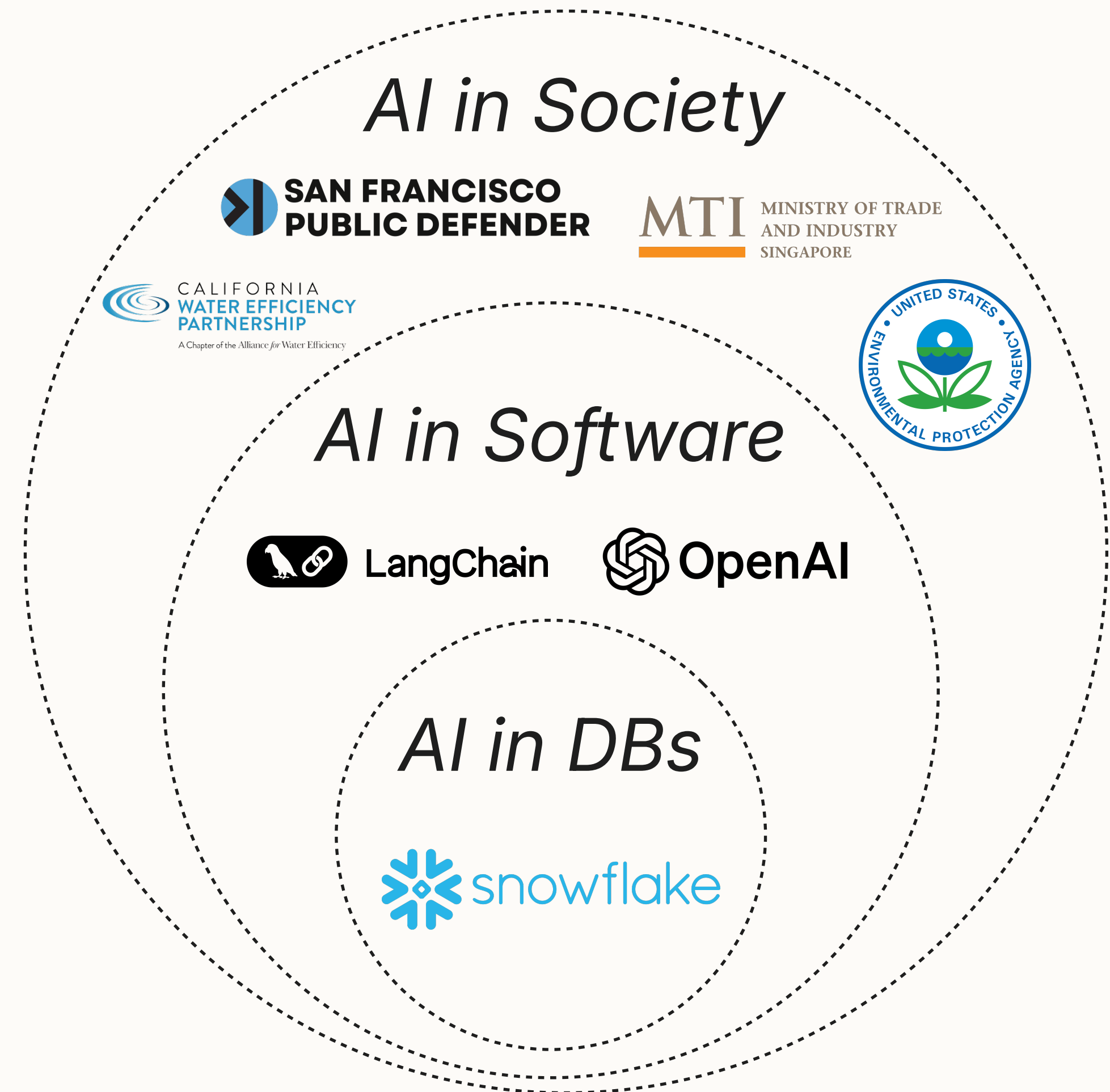
Build on database foundations, but adapt them—significantly!—for LLM realities.

Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)



Today's Talk

How can users **steer and debug** data systems that expose AI capabilities?

DocWrangler (UIST 🏆)

**Why is semantic data
processing so hard?**

DocWrangler as a Design Probe

- A: Pipeline editor
- B: Input & Output inspector (reminiscent of data wrangling, programming-by-example)

The screenshot displays the DocWrangler application interface. The top bar includes a menu (File, Edit, Help), a 'Quick Save' button, and a 'DocWrangler' logo. The main workspace is divided into several panels:

- FILES:** A sidebar on the left showing a file named 'medical_transcripts.json'.
- MedicalAnalysis:** The central panel, labeled 'A', contains a pipeline editor. It shows a 'map' operation using the 'gemini/gemini-2.0-flash' model with the prompt 'extract_discomfort_symptoms'. The prompt text is: 'extract discomfort information from the medical transcript. {{ input.src }} identify the discomfort level (low, medium, high), provide a brief description of the discomfort, and list the symptoms the patient complains about.' Below the prompt is an 'Output Schema' section with fields: 'discomfort_level' (enum with values 'low', 'medium', 'high'), 'discomfort_description' (string), and 'symptoms' (list, type 'string'). There is also a 'PDF URL Key' field with the example 'e.g. url or pdf_path'.
- OUTPUT - extract_discomfort_symptoms:** The bottom panel, labeled 'B', shows the output of the pipeline. It includes a 'Table' view with columns for 'symptoms', 'discomfort_level', 'discomfort_description', 'src', 'tgt', and 'file'. The 'symptoms' column shows an array of three items: 'nasal congestion', 'high blood pressure', and 'fluid in legs'. The 'discomfort_level' column shows 'medium'. The 'discomfort_description' column shows a detailed medical history. The 'src' column shows a patient's dialogue with a doctor. The 'tgt' column shows the extracted information. The 'file' column shows the source file path.
- Dataset Statistics:** A sidebar on the right showing statistics for 'medical_transcripts.json'. It includes a 'Word Count Distribution' histogram, 'Average Words: 1,626', 'Min Words: 714', 'Max Words: 3,376', and 'Std Deviation: 465'.

Cost: \$0.19

FilesOutputDataset

raw.json

Available Keys

Dataset Statistics

Search (min 5 characters)...

Q^v

No matches

```
1 [  
2   {  
3     "src": "[doctor] hi ,  
martha . how are you ?\n[patient] i'm  
doing okay . how are you ?\n[doctor]  
i'm doing okay . so , i know the nurse  
told you about dax . i'd like to tell  
dax a little bit about you , okay ?  
\n[patient] okay .\n[doctor] martha is  
a 50-year-old female with a past  
medical history significant for  
congestive heart failure , depression  
and hypertension who presents for her  
annual exam . so , martha , it's been  
a year since i've seen you . how are  
you doing ?\n[patient] i'm doing well  
 . i've been traveling a lot recently  
since things have , have gotten a bit  
lighter . and i got my , my vaccine ,  
so i feel safer about traveling . i've
```

Example Dataset: Conversation Transcripts between Doctors and Patients

Cost: \$0.19

Files

Output

Dataset

raw.json

> Available Keys

> Dataset Statistics

Search (min 5 characters)...

Q

^

v

No matches

1

2


3

[
 {
 "src": "[doctor] hi ,
martha . how are you ?\n[patient] i'm
doing okay . how are you ?\n[doctor]
i'm doing okay . so , i know the nurse
told you about dax . i'd like to tell
dax a little bit about you , okay ?
\n[patient] okay .\n[doctor] martha is
a 50-year-old female with a past
medical history significant for
congestive heart failure , depression
and hypertension who presents for her
annual exam . so , martha , it's been
a year since i've seen you . how are
you doing ?\n[patient] i'm doing well
 . i've been traveling a lot recently
since things have , have gotten a bit
lighter . and i got my , my vaccine ,
so i feel safer about traveling . i've

Example Dataset: Conversation Transcripts between Doctors and Patients

ick Save

Info

 **DocWrangler** (calm-bear-a2d6io0)

Cost: \$0.19

☐ Files

☐ Output

☐ Dataset

Medical_Analysis >

Add Operation +

Stop

Run Fresh

Run

map

gpt-4o-mini

extract_medications

Show Outputs

Improve Prompt

Enter prompt (must be a Jinja2 template)

Prompt must contain Jinja2 template syntax {{ and }}

^ Output Schema

medications


list

List type: string

**Semantic operators are very expressive!
At the same time: very hard to author.**

ick Save

Info

 **DocWrangler** (calm-bear-a2d6io0)

Cost: \$0.19

☐ Files

☐ Output

☐ Dataset

Medical_Analysis >

Add Operation +

Stop

Run Fresh

Run

map

gpt-4o-mini

extract_medications

Show Outputs

Improve Prompt

Enter prompt (must be a Jinja2 template)

Prompt must contain Jinja2 template syntax {{ and }}

^ Output Schema

medications

list

List type: string

**Semantic operators are very expressive!
At the same time: very hard to author.**

Observation: Long Tail of Issues

Observation: Long Tail of Issues

An Output

Q Search in cell...

^ **Array (2 items)**

0: "Lisinopril"

1: "Lasix 20 mg/day"

Observation: Long Tail of Issues

An Output

Q Search in cell...

^ **Array (2 items)**

0: "Lisinopril"

1: "Lasix 20 mg/day"

A User Reaction

I want the analysis to include dosages for all medications.

Observation: Long Tail of Issues

Observation: Long Tail of Issues

An Output

Q Search in cell...

^ **Array (2 items)**

0: "Ultram 50 mg every six hours as needed"

1: "Synthroid – continue on current dosage"

Observation: Long Tail of Issues

An Output

```
Q Search in cell...  
^ Array (2 items)  
  0: "Ultram 50 mg every six hours as needed"  
  1: "Synthroid – continue on current dosage"
```

A User Reaction

*"Current dosage" is not specific.
I want the exact dosage for
Synthroid...*

Observation: Long Tail of Issues

Observation: Long Tail of Issues

An Output

```
Q Search in cell...  
^ Array (5 items)  
  0: "CoQ10"  
  1: "Vitamin D"  
  2: "Vitamin C"  
  3: "Fish Oil"  
  4: "Elderberry Fruit"
```

Observation: Long Tail of Issues

An Output

```
Q Search in cell...  
^ Array (5 items)  
  0: "CoQ10"  
  1: "Vitamin D"  
  2: "Vitamin C"  
  3: "Fish Oil"  
  4: "Elderberry Fruit"
```

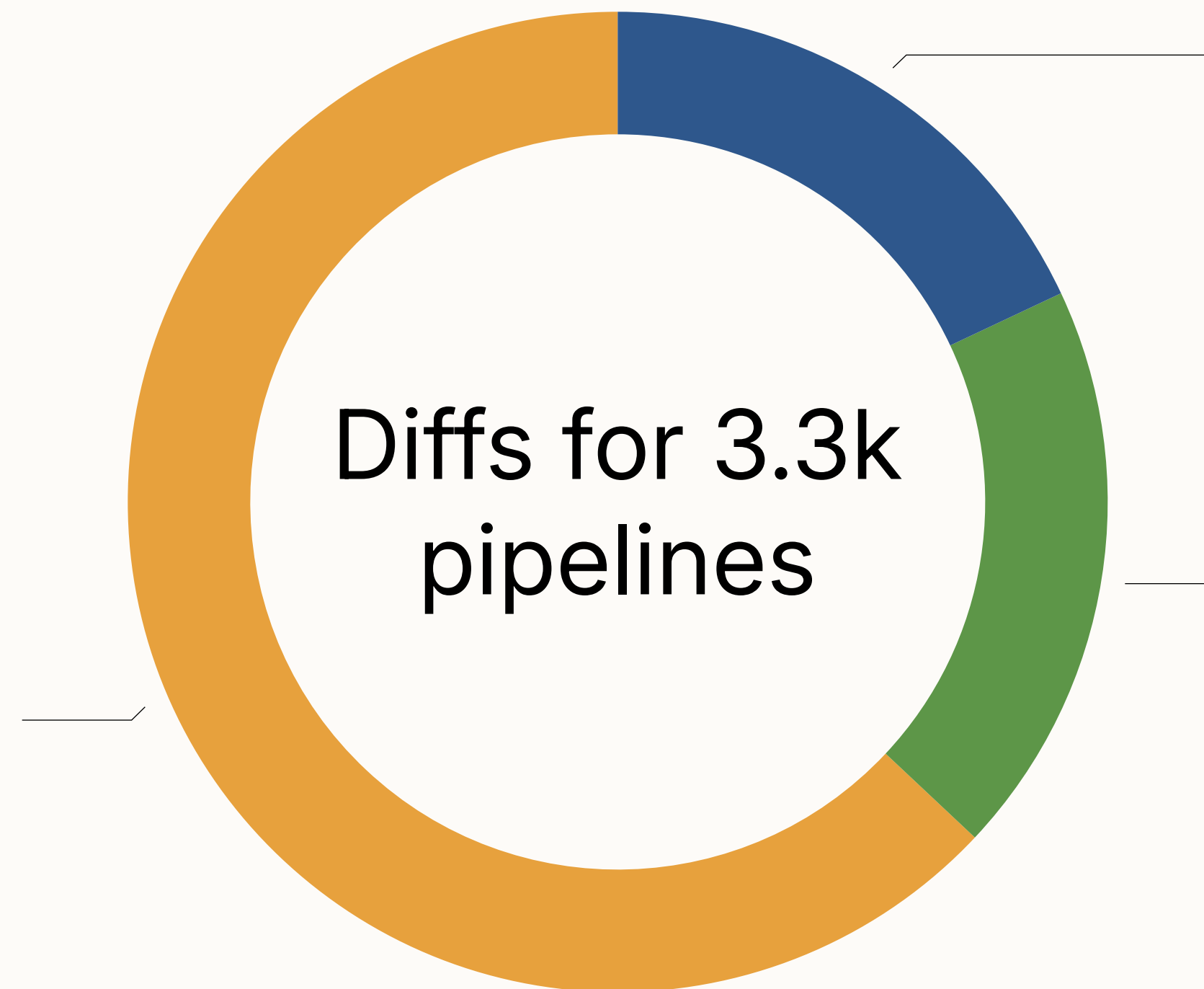
A User Reaction

I don't want the LLM to extract over-the-counter meds or supplements. Prescription meds only.

Pipelines are Discovered, not Declared

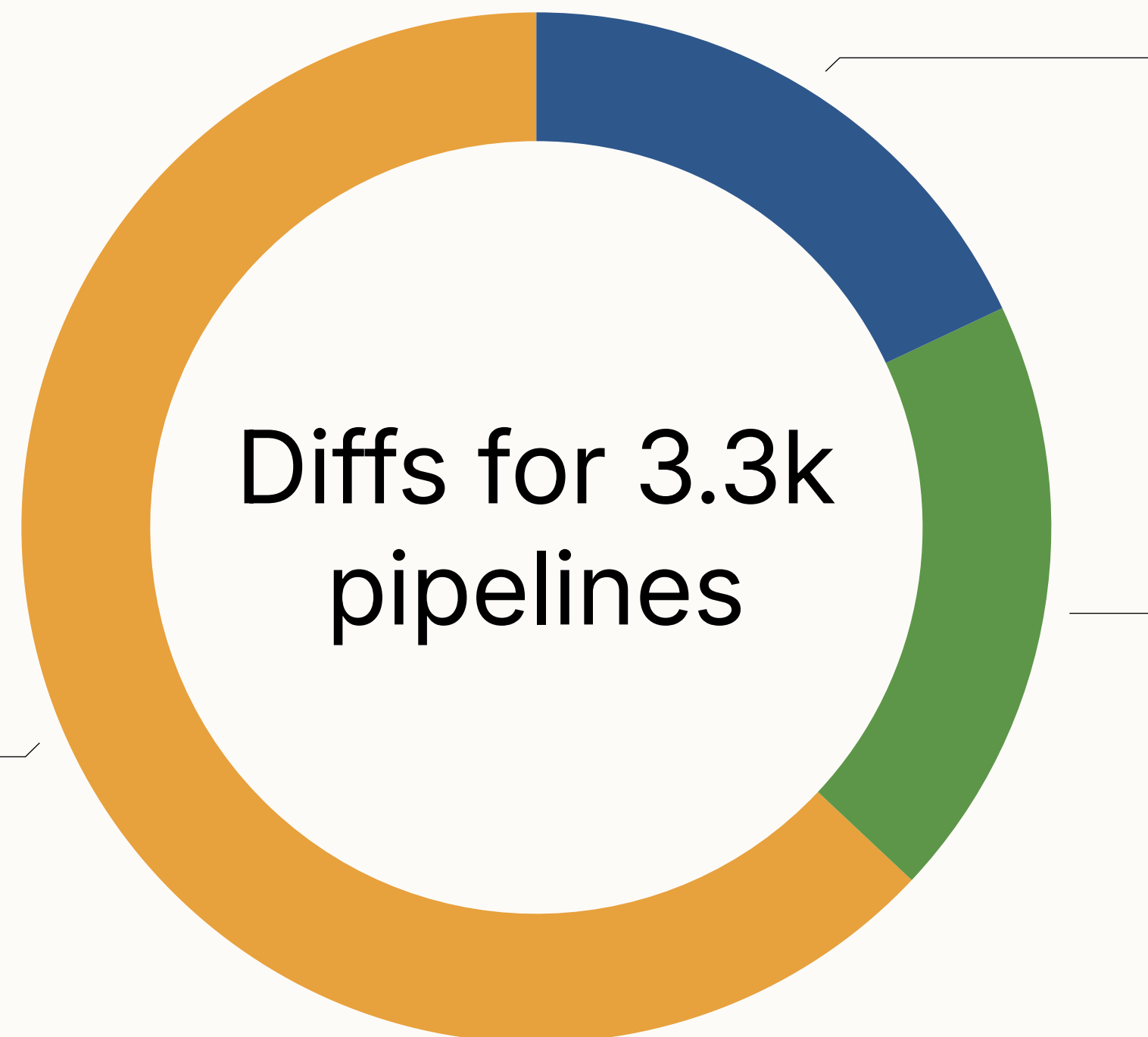
Pipelines are Discovered, not Declared

- No change
- Minor change
- >10 char diff in an operation



Pipelines are Discovered, not Declared

- No change
- Minor change
- >10 char diff in an operation



If semantic operators need authoring tools, what **discovery-oriented** tools do we build?

Why is semantic data processing so hard?

~~Why is semantic data
processing so hard?~~

~~Why is semantic data
processing so hard?~~

What tools do users need to
accelerate pipeline discovery?

Improve Prompt

Select the operation you want to improve the prompt for

extract_medications

Current Prompt:

from the following conversation transcript, list all the medications that the doctor prescribed to the patient: {{ input }}

Your Notes:

No feedback or bookmarks found for this operation.

Additional Instructions (optional)

Add specific instructions for improving the prompt (e.g., 'Make it more concise', 'Add more examples')

Leave blank to let the AI follow default improvement guidelines

Continue to Analysis

One idea: an *interactive* assistant to improve semantic operator prompts

Improve Prompt

Select the operation you want to improve the prompt for

extract_medications

Current Prompt:

from the following conversation transcript, list all the medications that the doctor prescribed to the patient: {{ input }}

Your Notes:

No feedback or bookmarks found for this operation.

Additional Instructions (optional)

Add specific instructions for improving the prompt (e.g., 'Make it more concise', 'Add more examples')

Leave blank to let the AI follow default improvement guidelines

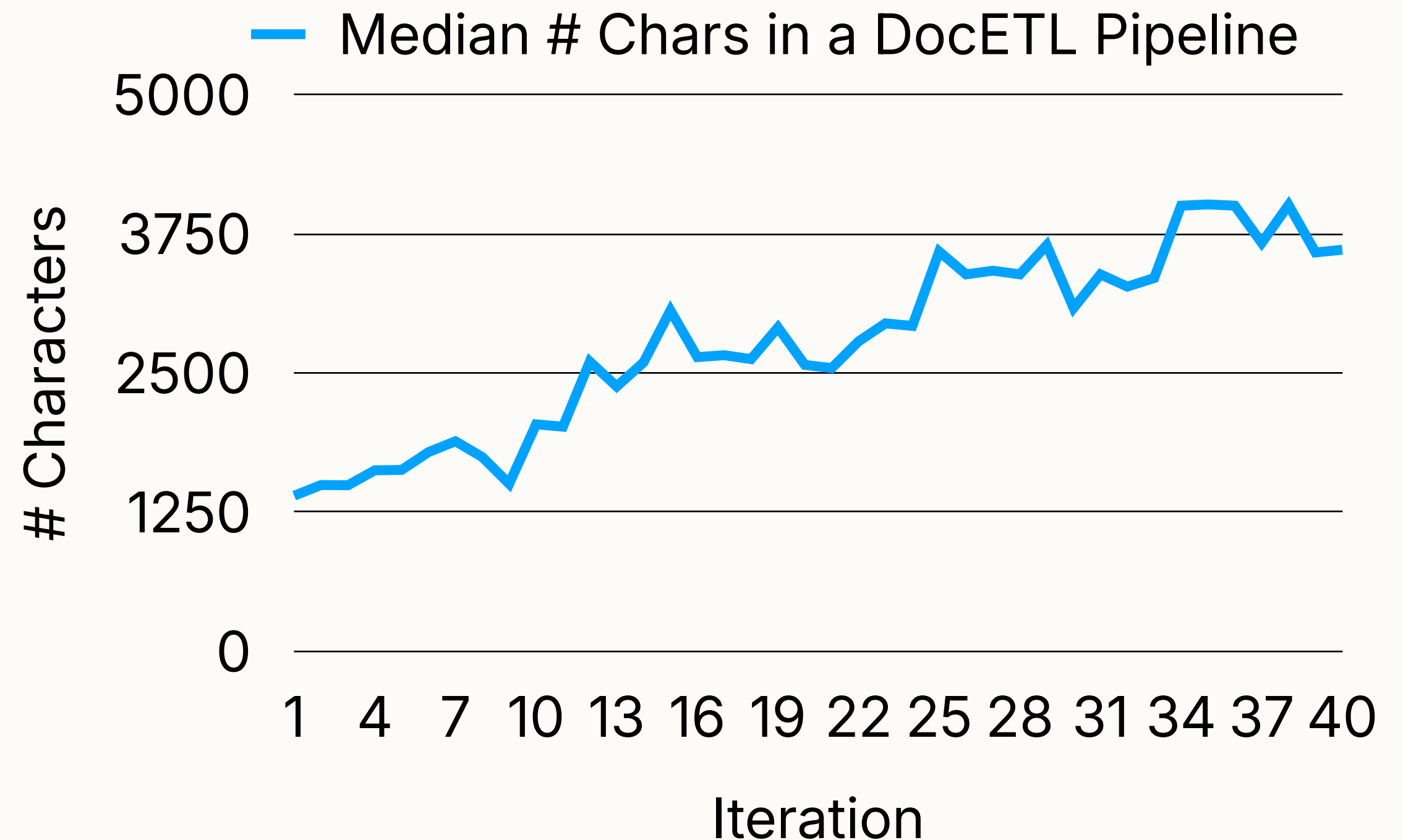
Continue to Analysis

One idea: an *interactive* assistant to improve semantic operator prompts

Towards a Principled Understanding

Building one tool doesn't solve all problems; e.g.,

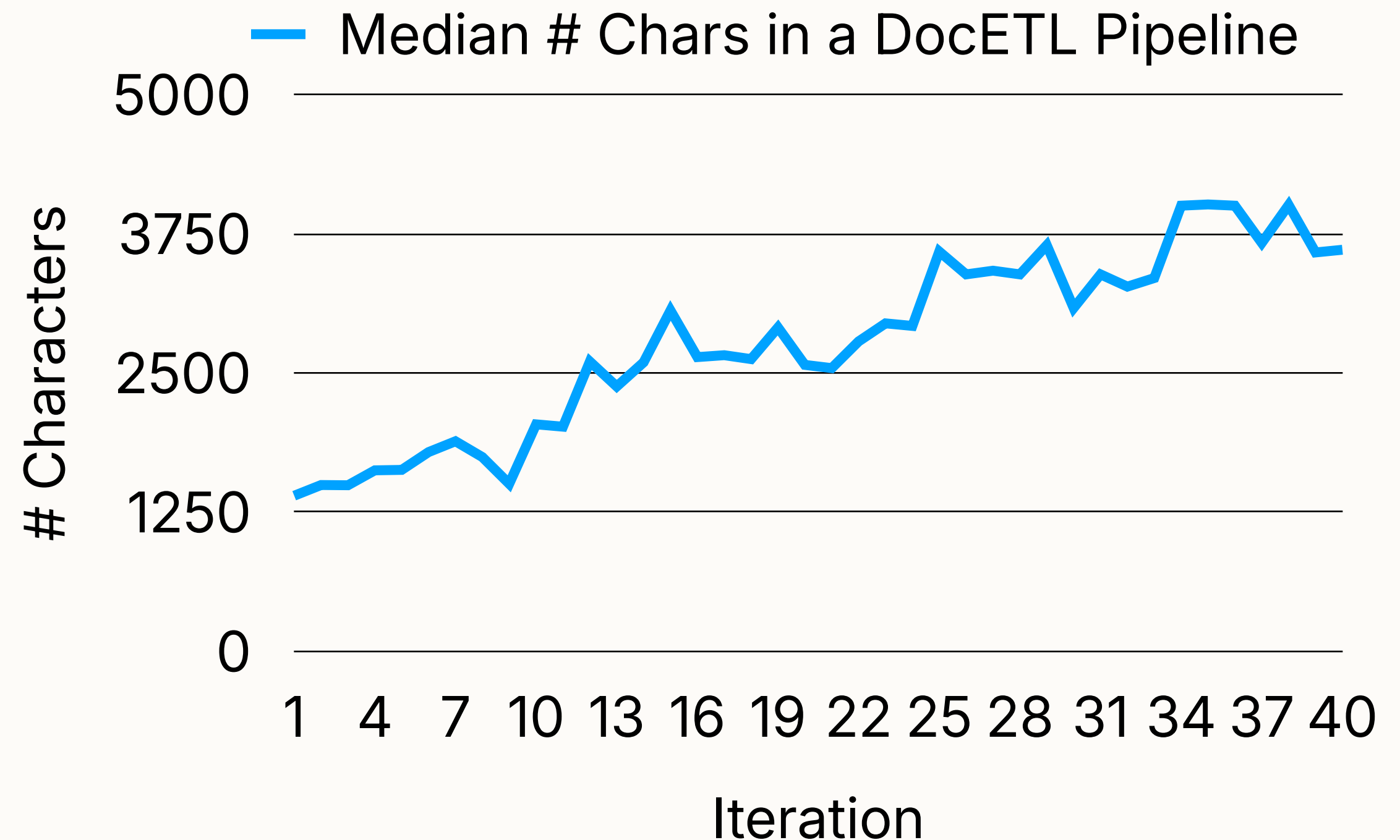
- ◆ It's hard to remember all the changes to make;
- ◆ Each edit adds more detail;
- ◆ Operators become bloated and hard to validate and manage.



Towards a Principled Understanding

Building one tool doesn't solve all problems; e.g.,

- ◆ It's hard to remember all the changes to make;
- ◆ Each edit adds more detail;
- ◆ Operators become bloated and hard to validate and manage.



We don't just need better tools; we need a theory for understanding where tools are needed.

~~Why is semantic data
processing so hard?~~

What tools do users need to
accelerate pipeline discovery?

~~Why is semantic data
processing so hard?~~

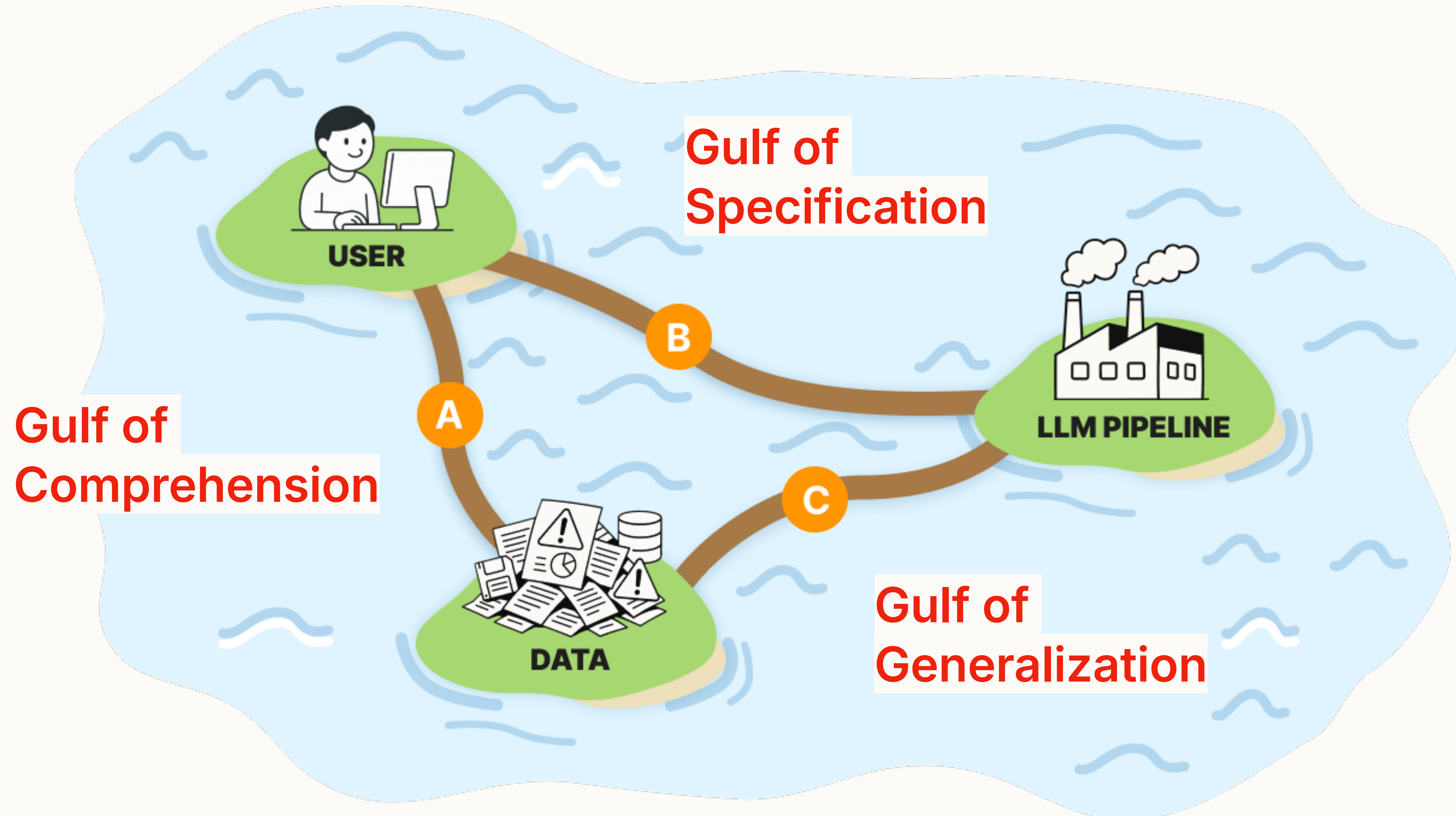
~~What tools do users need to
accelerate pipeline discovery?~~

~~Why is semantic data
processing so hard?~~

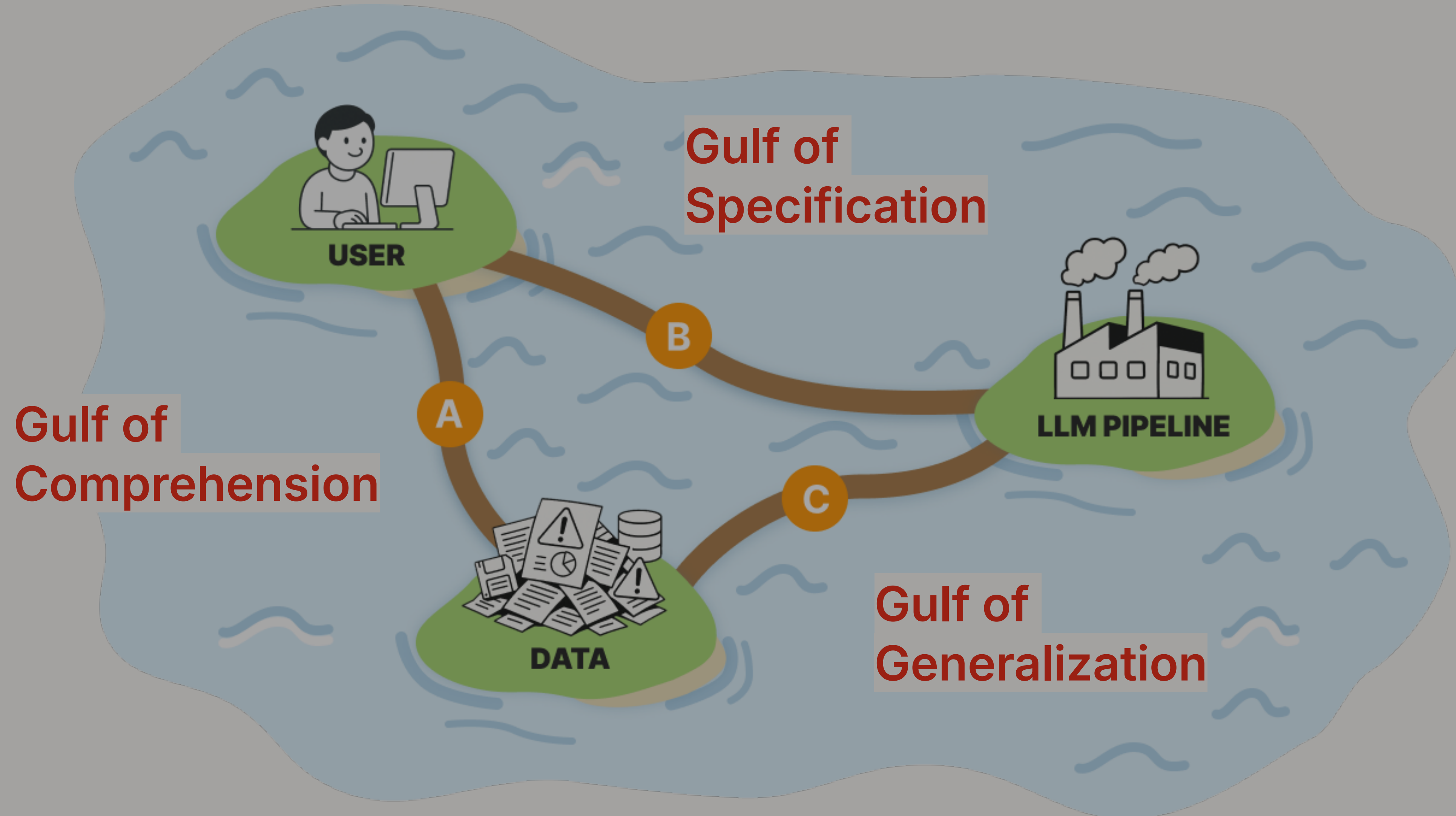
~~What tools do users need to
accelerate pipeline discovery?~~

How do we map the space of
user challenges?

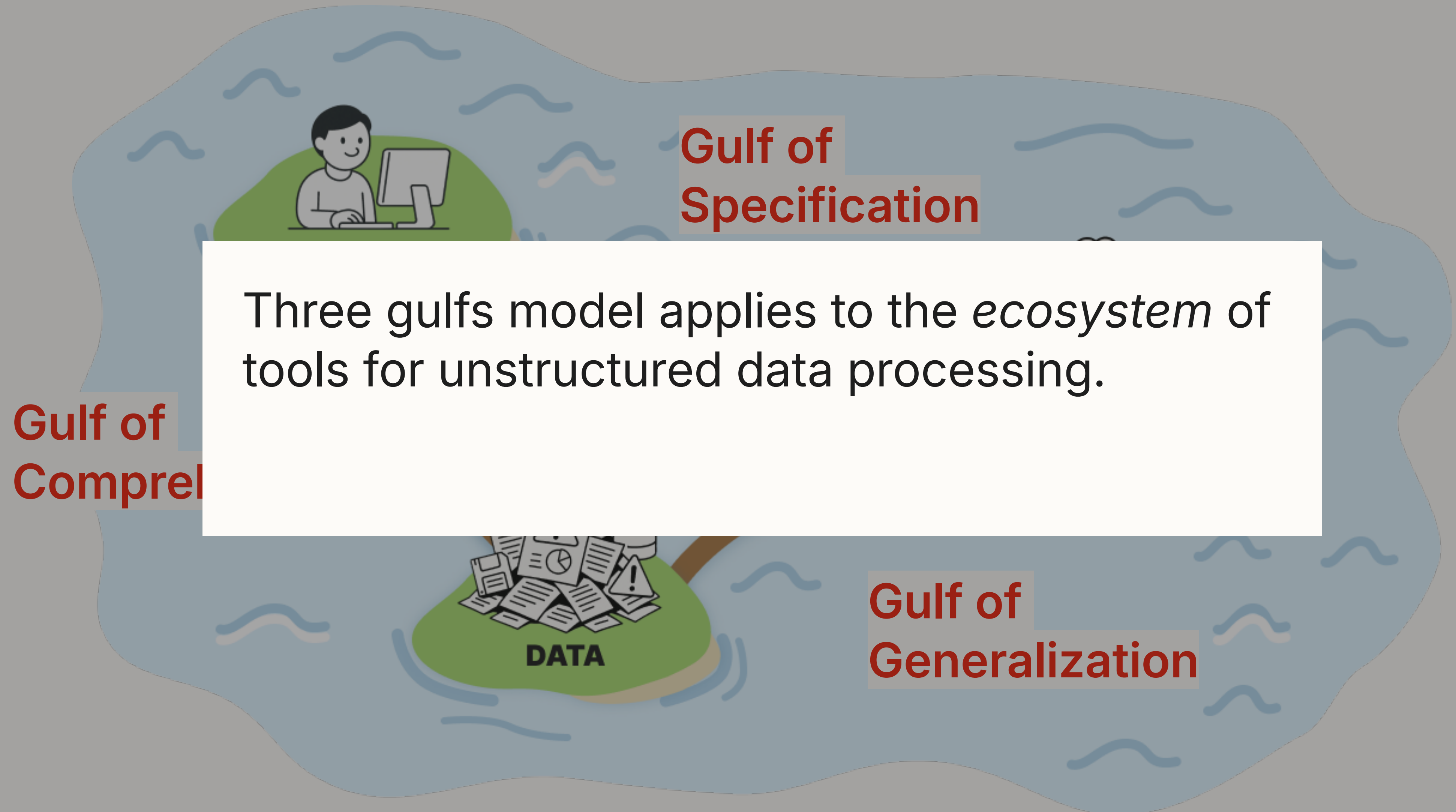
Three Gulfs Model



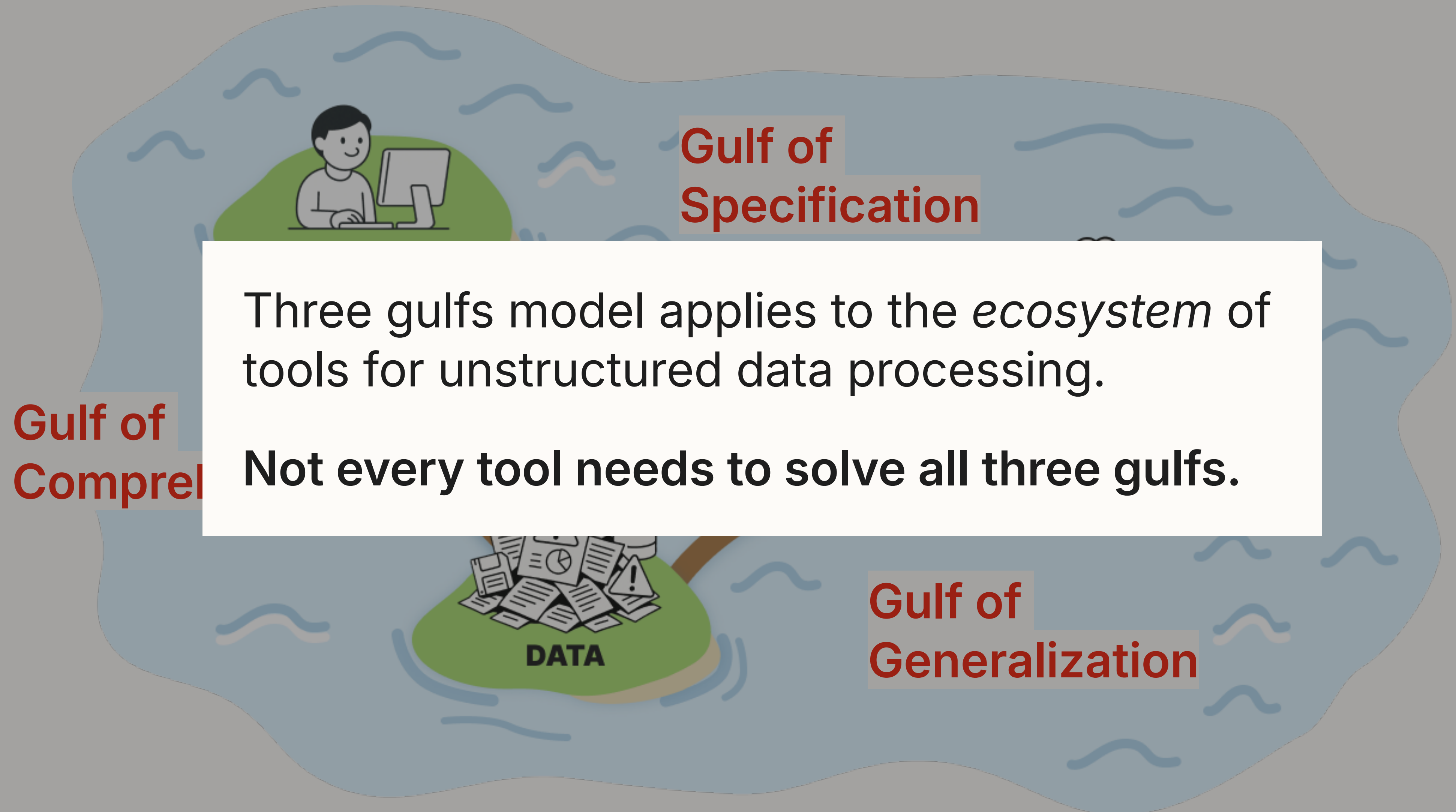
Three Gulfs Model



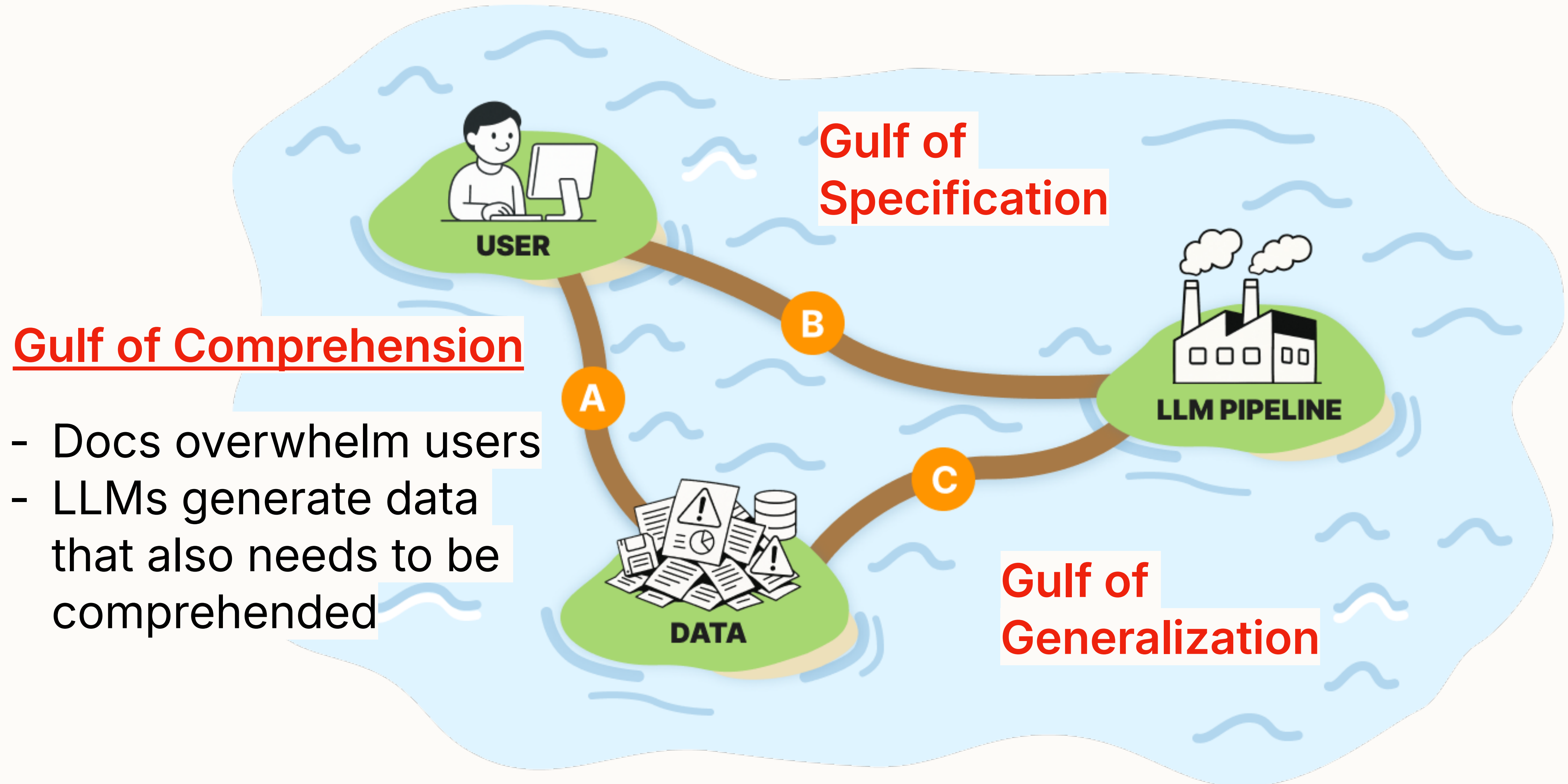
Three Gulfs Model



Three Gulfs Model



Three Gulfs Model



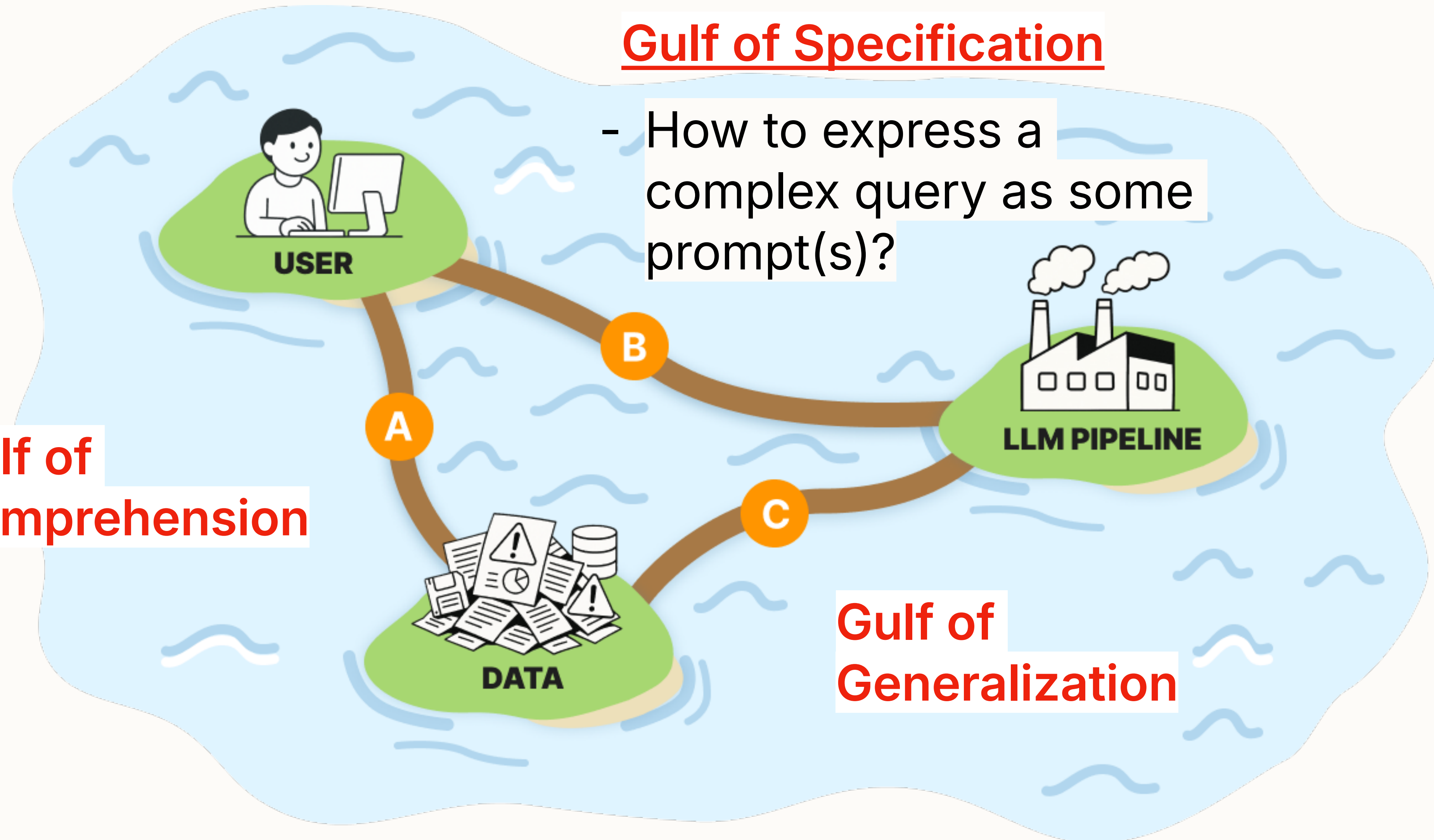
Three Gulfs Model

Gulf of Specification

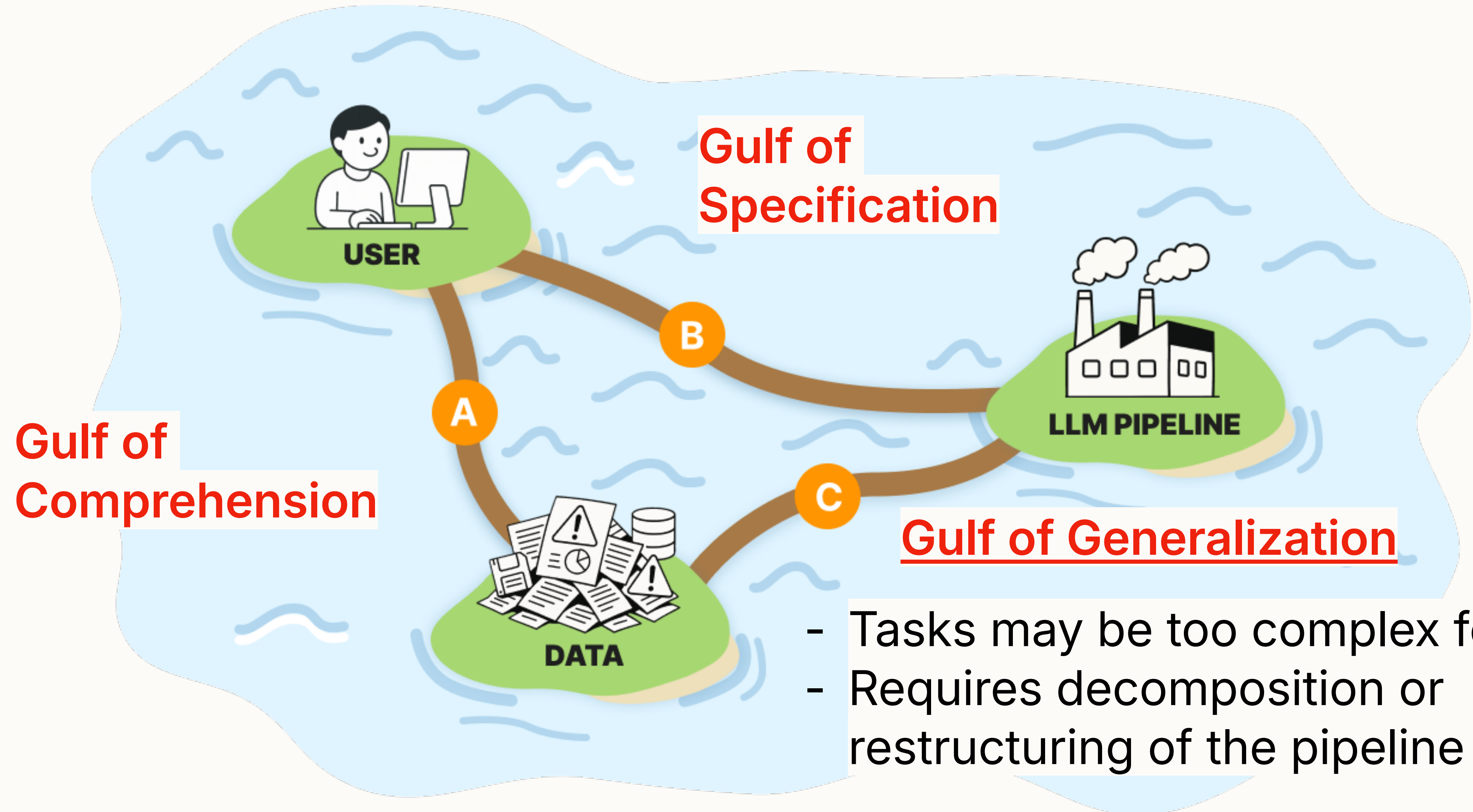
- How to express a complex query as some prompt(s)?

Gulf of Comprehension

Gulf of Generalization



Three Gulfs Model



Why the Three Gulfs Model Matters

Steering Semantic Data Processing with DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆

Why the Three Gulfs Model Matters

Steering Semantic Data Processing with DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆

Systems Perspective

- ✱ Precise diagnosis of user-facing challenges
- ✱ Enables better solutions and evaluation

Why the Three Gulfs Model Matters

Steering Semantic Data Processing with DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆

Systems Perspective

- ✱ Precise diagnosis of user-facing challenges
- ✱ Enables better solutions and evaluation

HCI Perspective

- ✱ New methodology: “lateral” deployment
- ✱ Empirical foundation for studying how people build, adapt, and reason about AI

Impact: DocWrangler Users

Steering Semantic Data Processing with DocWrangler. **Shankar***, Chopra* et al. *UIST '25*. 🏆



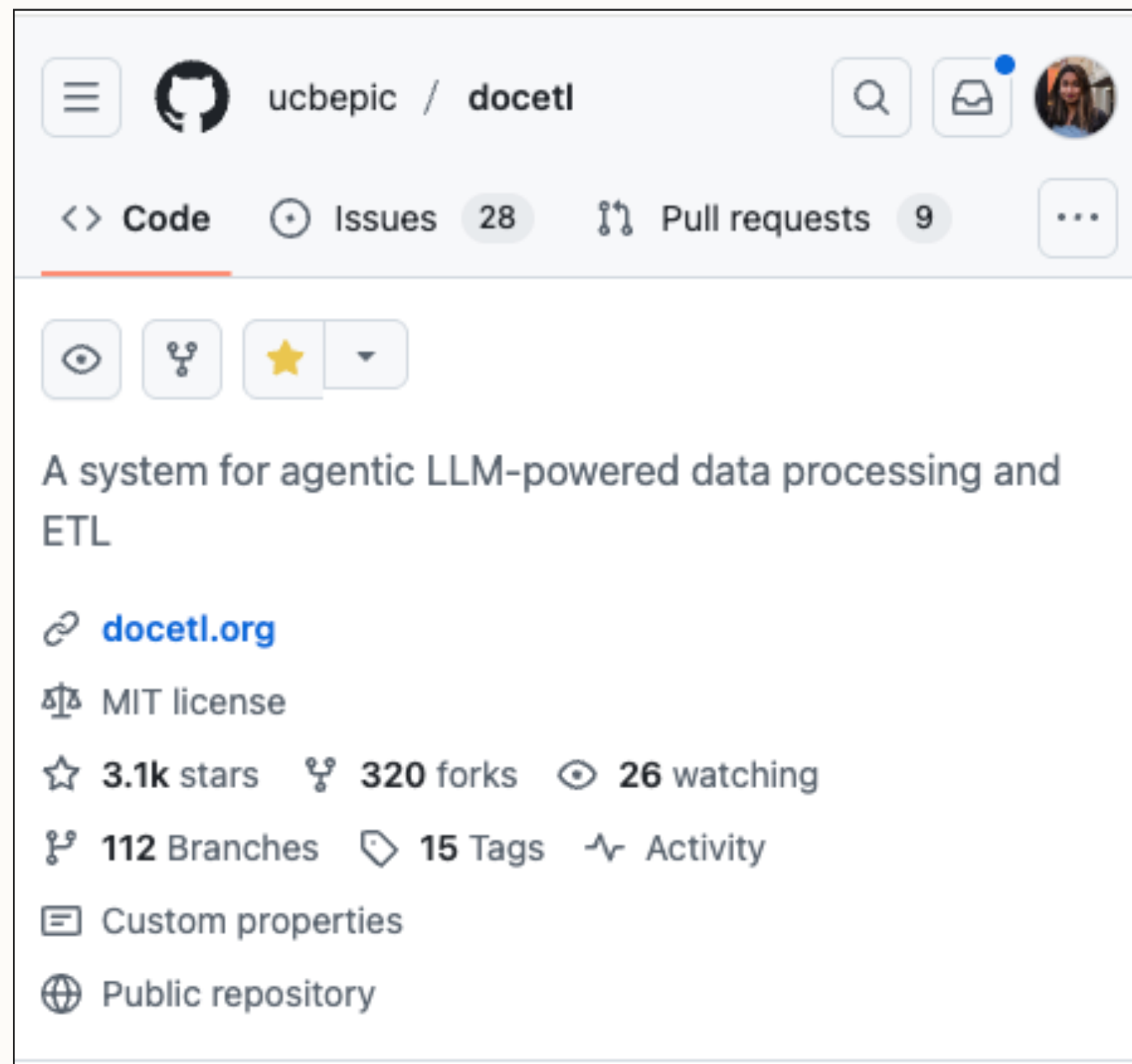
Impact: DocWrangler Users

Steering Semantic Data Processing with DocWrangler. **Shankar***, Chopra* et al. **UIST '25.** 🏆

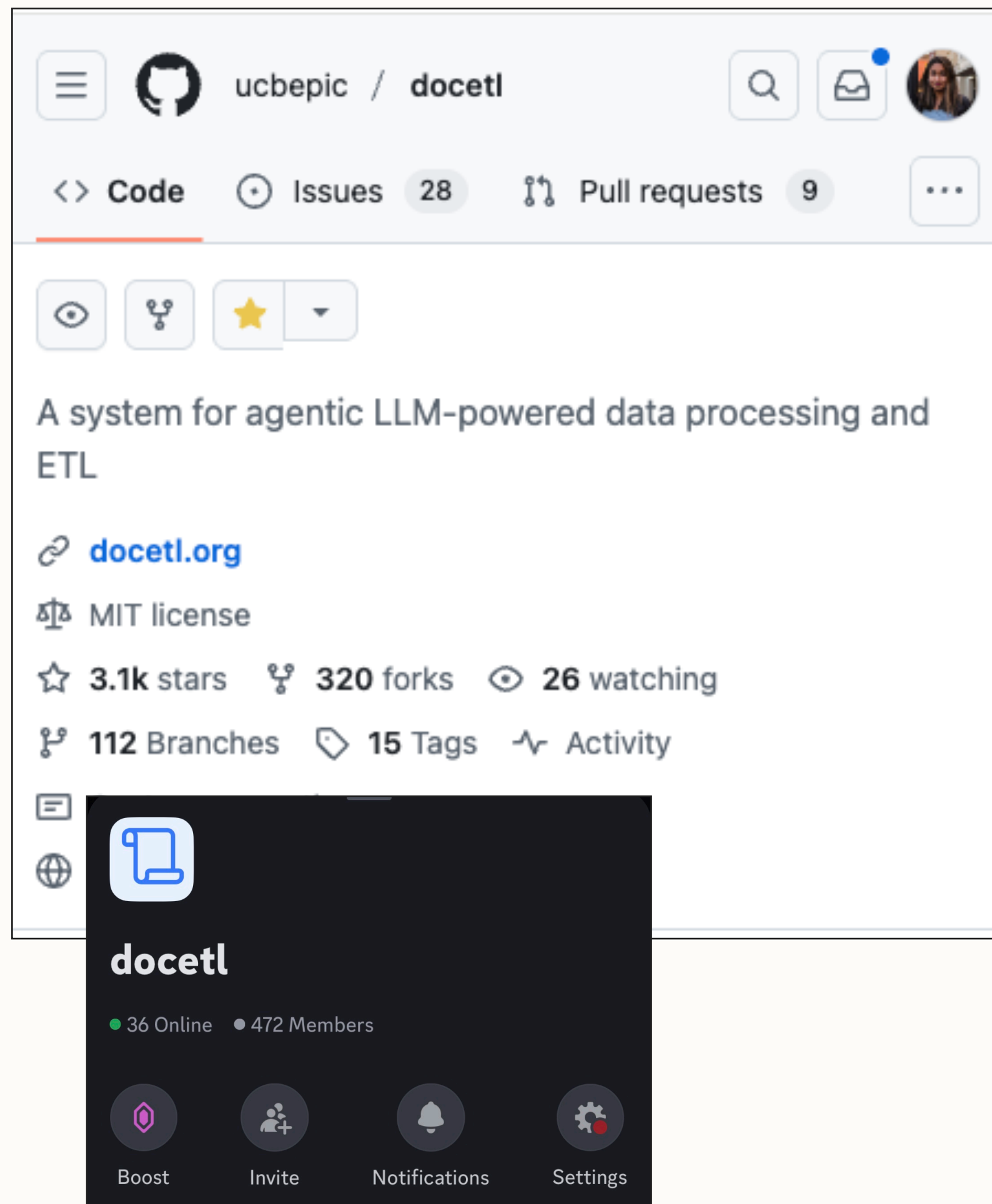


Impact: DocETL Community

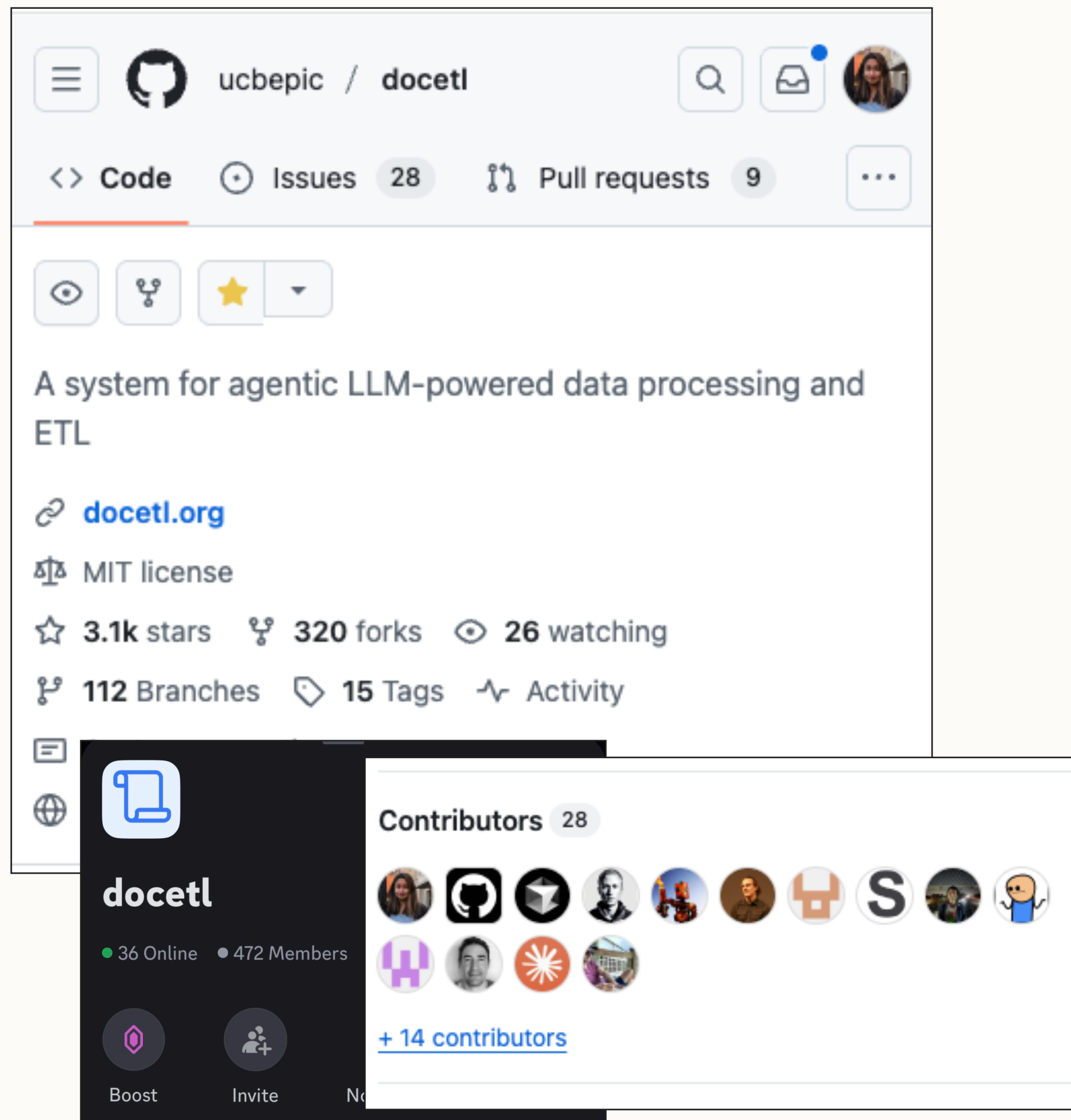
Impact: DocETL Community



Impact: DocETL Community



Impact: DocETL Community



Impact: DocETL Community

The image is a collage of three screenshots illustrating the DocETL community and its tools.

Top Left: GitHub Repository
The screenshot shows the GitHub repository for 'ucbepic / docetl'. It features the repository name, a search bar, and navigation tabs for 'Code', 'Issues' (28), and 'Pull requests'. Below the repository name, it states 'A system for agentic LLM-powered data processing ETL'. The repository has 3.1k stars, 320 forks, and 26 watchers. It also shows 112 branches, 15 tags, and an activity graph.

Bottom Left: Discord Server
The screenshot shows the 'docetl' Discord server. It displays the server name, a description, and a list of 28 contributors. The server has 36 online members and 472 total members. There are buttons for 'Boost' and 'Invite'.

Right: DocWrangler Interface
The screenshot shows the 'DocWrangler' web interface. It has a top bar with 'File', 'Edit', 'Help', 'Quick Save', and 'Info'. The main area is titled 'Blog Post LLM Context' and shows a list of files. A 'NOTES' section on the left contains a list of notes. The right pane shows a 'Blog Post LLM Context' with a prompt and an example output. The output is a structured JSON object with 'key_points', 'filename', and 'markdown' fields. Below the output, there are three charts: 'key_points', 'filename', and 'markdown'. A text overlay at the bottom of the interface reads: 'which I think is pretty cool, is I'.

Impact: DocETL Community

ucbepic / docetl

<> Code

Issues 28

Pull requests

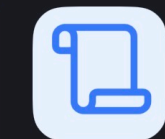
A system for agentic LLM-powered data processing ETL

[docetl.org](#)

MIT license

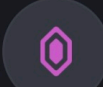
3.1k stars 320 forks 26 watching

112 Branches 15 Tags Activity



docetl

36 Online 472 Members

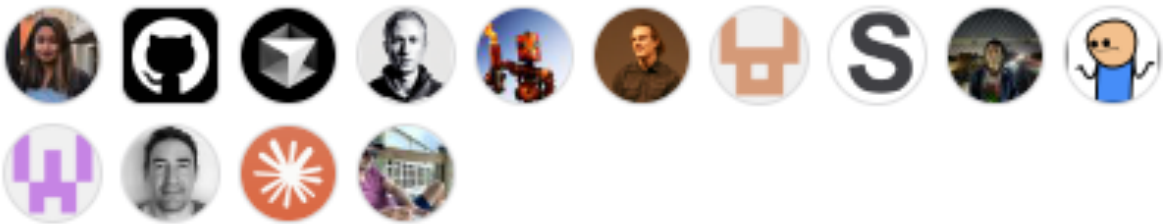


Boost



Invite

Contributors 28



+ 14 contributors

DocETL - AI-Powered Document Processing

Claude

[2410.12189] DocETL: Agent...

localhost:3000/playground

File Edit Help Quick Save Info

DocWrangler (blog-post-llm-context) Cost: \$0.07 Files Output Dataset

FILES

Tip: Right-click files to view, download or delete them

converted_2025-09-17T16-02-59-013Z

2021-04-12-fastai-chapter-6.md

2021-04-25-fastai-chapter-6-bear-classifier.md

NOTES

Tip: Click in any to add notes

Note: Notes are only visible in the current prompt, not in operation

Search...

The chunk is small enough where the takeaways are pretty...

No explanatory text is in this chunk, so the LLM has inferred...

The document title is not included in this set of key points.

The key points are valid and they match the Markdown chunk...

The key points here are unnecessary and somewhat...

I want document title removed from the LLM prompt.

Blog Post LLM Context

Overview System Prompts 5 Add Operation Stop Run Fresh Run

Please follow these steps:

Summarize Information: Condense the extracted information from this section into concise summaries. Aim for 3-5 key points.

Output Format:

The final output should be structured as follows:

Section Key Points:

Summary of key point 1

Summary of key point 2

Summary of key point 3

Example Output:

Document Title: "Market Analysis 2023"

Key Points:

Point 1: "Market growth projected at 10% YoY."

Point 2: "Emerging competitors include Company A and Company B."

Point 3: "Consumer preferences shifting towards eco-friendly products."

OUTPUT - Untitled Map 0 Console Table Visualize Input Distribution 99 in 99 out 1.00x

key_points filename markdown

Filter...

3 items avg: 4 6 items 5 distinct values

Search in cell...

2021-04-12-fastai-chapter-6

title: "fast.ai Chapter 6: Classification Models" date: "2021-04-12"

An example image from the image dataset used in this lesson. The image has a train going

introduced two more classification models:

classification, for when you want to predict more than one or no label per image

Regression, for when you want to predict a quantity instead of a category for an image

which I think is pretty cool, is

@EkShunya 1 month ago

nice to see you back , docwrangler +1

Reply

@alexkelly757 1 month ago

I was wondering about resolve myself..decided to go for reduce. It was interesting your thoughts process on map and reduce for the last operation. I would have gone for reduce but map worked. Thanks for the video! Try the gpt5 nano model, it's so cheap and good.

Reply

1 reply

@vishal_learner 1 month ago

Thanks I'll try out gpt5 nano for sure. I need to experiment with more models. Regarding the final map: the reason I went with map and not reduce was that there was only 1 input (the string with a bulleted list of key points across all document chunks) so it's not like it was aggregating across multiple documents, it was summarizing within a single document. But I'm sure conceptually it works either way, will test out both to see if the outputs are better/worse.

1

Reply

Impact: DocETL Community

ucbepic / docetl

<> Code

Issues 28

Pull requests

1:37

1:37

1:37

1:37

A system for agentic LLM-powered data processing

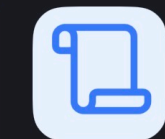
ETL

docetl.org

MIT license

3.1k stars 320 forks 26 watching

112 Branches 15 Tags Activity



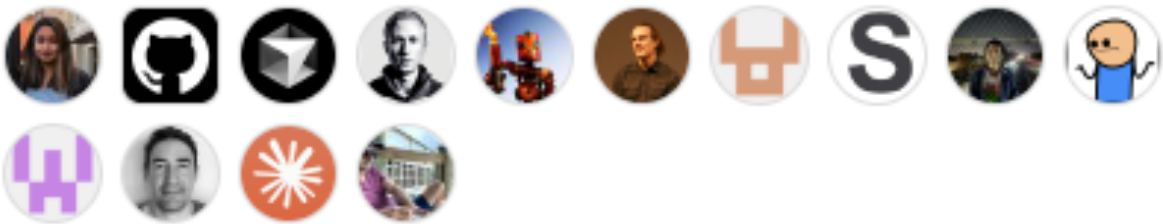
docetl

36 Online 472 Members



Boost Invite

Contributors 28



+ 14 contributors

DocETL - AI-Powered Document Processing

DocWrangler (blog-post-llm-context)

Cost: \$0.00

FILES

converted_2025-09-17T16-02-59-013Z

2021-04-12-fastai-chapter-6.md

2021-04-25-fastai-chapter-6-bear-classifier.md

NOTES

Tip: Click in any output

Note: Notes are on prompts

Search...

The chunk is small enough where the takeaways are pretty...

No explanatory text is in this chunk, so the LLM has inferred...

The document title is not included in this set of key points.

The key points are valid and they match the Markdown chunk...

The key points here are unnecessary and somewhat...

I want document title removed from the LLM prompt.

Blog Post LLM Context

Overview System Prompts 5 Add Operation

Please follow these steps:

Summarize Information: Condense the extracted information from this section into key points.

Output Format:

- The final output should be structured as follows:

- **Section Key Points**:

- [Summary of key point 1]

- [Summary of key point 2]

- [Summary of key point 3]

Example Output:

- **Document Title**: "Market Analysis 2023"

- **Key Points**:

- Point 1: "Market growth projected at 10% YoY."

- Point 2: "Emerging competitors include Company A and Company B."

- Point 3: "Consumer preferences shifting towards eco-friendly products."

OUTPUT - Untitled Map 0

Console Table Visualize Input Distribution

key_points filename markdown

Filter...

3 items avg: 4 6 items 5 distinct values

Search in cell...

2021-04-12-fastai-chapter-6.md

title: "fast.ai Chapter 6: Classification"

which I think is pretty cool, is

Using DocWrangler to Process Blog Posts

vishal 1.02K subscribers

Subscribe

@EkShunya 1 month ago

nice to see you back, docwrangler +1

👍 🗨️ 🔄 Reply

@alexkelly757 1 month ago

I was wondering about resolve myself..decided to go for reduce. It was interesting your thoughts process on map and reduce

👍 🗨️ 🔄 Reply

@vishal_learner 1 month ago

Thanks I'll try out gpt5 nano for sure. I need to experiment with more models. Regarding the final map: the reason I was aggregating across multiple documents, it was summarizing within a single document. But I'm sure conceptually it was

👍 1 🗨️ 🔄 Reply

Solvelt + FastHTML: Creating Interactive Search Applications

AI NLP

AUTHOR
Alex Kelly

PUBLISHED
October 4, 2025

Displaying enriched docETL data using solvit/fasthtml

The reason for this project is I have data that is currently in Excel, i want to build a front end website to display the information and use advanced information retrievable. The data was enriched using DocETL which enables you to build a series of LLM operations to extract and enrich information from unstructured data. I see these LLM operations been key in every organization and being able to display and manipulate the data as a key skill. In this case it is a fictitious set of email data. Lets dive into the code :

Building a FastHTML + LanceDB Search App Using Solvelt (Jeremy H...)

Watch later Share

solvelt • Testing fasthtml json_final_cut

Code 01 (02:14:48)

Code 02 (02:14:48)

Code 03 (02:14:48)

Code 04 (02:14:48)

Code 05 (02:14:48)

Code 06 (02:14:48)

Code 07 (02:14:48)

Code 08 (02:14:48)

Code 09 (02:14:48)

Code 10 (02:14:48)

Code 11 (02:14:48)

Code 12 (02:14:48)

Code 13 (02:14:48)

Code 14 (02:14:48)

Code 15 (02:14:48)

Code 16 (02:14:48)

Code 17 (02:14:48)

Code 18 (02:14:48)

Code 19 (02:14:48)

Code 20 (02:14:48)

Code 21 (02:14:48)

Code 22 (02:14:48)

Code 23 (02:14:48)

Code 24 (02:14:48)

Code 25 (02:14:48)

Code 26 (02:14:48)

Code 27 (02:14:48)

Code 28 (02:14:48)

Code 29 (02:14:48)

Code 30 (02:14:48)

Code 31 (02:14:48)

Code 32 (02:14:48)

Code 33 (02:14:48)

Code 34 (02:14:48)

Code 35 (02:14:48)

Code 36 (02:14:48)

Code 37 (02:14:48)

Code 38 (02:14:48)

Code 39 (02:14:48)

Code 40 (02:14:48)

Code 41 (02:14:48)

Code 42 (02:14:48)

Code 43 (02:14:48)

Code 44 (02:14:48)

Code 45 (02:14:48)

Code 46 (02:14:48)

Code 47 (02:14:48)

Code 48 (02:14:48)

Code 49 (02:14:48)

Code 50 (02:14:48)

Code 51 (02:14:48)

Code 52 (02:14:48)

Code 53 (02:14:48)

Code 54 (02:14:48)

Code 55 (02:14:48)

Code 56 (02:14:48)

Code 57 (02:14:48)

Code 58 (02:14:48)

Code 59 (02:14:48)

Code 60 (02:14:48)

Code 61 (02:14:48)

Code 62 (02:14:48)

Code 63 (02:14:48)

Code 64 (02:14:48)

Code 65 (02:14:48)

Code 66 (02:14:48)

Code 67 (02:14:48)

Code 68 (02:14:48)

Code 69 (02:14:48)

Code 70 (02:14:48)

Code 71 (02:14:48)

Code 72 (02:14:48)

Code 73 (02:14:48)

Code 74 (02:14:48)

Code 75 (02:14:48)

Code 76 (02:14:48)

Code 77 (02:14:48)

Code 78 (02:14:48)

Code 79 (02:14:48)

Code 80 (02:14:48)

Code 81 (02:14:48)

Code 82 (02:14:48)

Code 83 (02:14:48)

Code 84 (02:14:48)

Code 85 (02:14:48)

Code 86 (02:14:48)

Code 87 (02:14:48)

Code 88 (02:14:48)

Code 89 (02:14:48)

Code 90 (02:14:48)

Code 91 (02:14:48)

Code 92 (02:14:48)

Code 93 (02:14:48)

Code 94 (02:14:48)

Code 95 (02:14:48)

Code 96 (02:14:48)

Code 97 (02:14:48)

Code 98 (02:14:48)

Code 99 (02:14:48)

Code 100 (02:14:48)

Code 101 (02:14:48)

Code 102 (02:14:48)

Code 103 (02:14:48)

Code 104 (02:14:48)

Code 105 (02:14:48)

Code 106 (02:14:48)

Code 107 (02:14:48)

Code 108 (02:14:48)

Code 109 (02:14:48)

Code 110 (02:14:48)

Code 111 (02:14:48)

Code 112 (02:14:48)

Code 113 (02:14:48)

Code 114 (02:14:48)

Code 115 (02:14:48)

Code 116 (02:14:48)

Code 117 (02:14:48)

Code 118 (02:14:48)

Code 119 (02:14:48)

Code 120 (02:14:48)

Code 121 (02:14:48)

Code 122 (02:14:48)

Code 123 (02:14:48)

Code 124 (02:14:48)

Code 125 (02:14:48)

Code 126 (02:14:48)

Code 127 (02:14:48)

Code 128 (02:14:48)

Code 129 (02:14:48)

Code 130 (02:14:48)

Code 131 (02:14:48)

Code 132 (02:14:48)

Code 133 (02:14:48)

Code 134 (02:14:48)

Code 135 (02:14:48)

Code 136 (02:14:48)

Code 137 (02:14:48)

Code 138 (02:14:48)

Code 139 (02:14:48)

Code 140 (02:14:48)

Code 141 (02:14:48)

Code 142 (02:14:48)

Code 143 (02:14:48)

Code 144 (02:14:48)

Code 145 (02:14:48)

Code 146 (02:14:48)

Code 147 (02:14:48)

Code 148 (02:14:48)

Code 149 (02:14:48)

Code 150 (02:14:48)

Code 151 (02:14:48)

Code 152 (02:14:48)

Code 153 (02:14:48)

Code 154 (02:14:48)

Code 155 (02:14:48)

Code 156 (02:14:48)

Code 157 (02:14:48)

Code 158 (02:14:48)

Code 159 (02:14:48)

Code 160 (02:14:48)

Code 161 (02:14:48)

Code 162 (02:14:48)

Code 163 (02:14:48)

Code 164 (02:14:48)

Code 165 (02:14:48)

Code 166 (02:14:48)

Code 167 (02:14:48)

Code 168 (02:14:48)

Code 169 (02:14:48)

Code 170 (02:14:48)

Code 171 (02:14:48)

Code 172 (02:14:48)

Code 173 (02:14:48)

Code 174 (02:14:48)

Code 175 (02:14:48)

Code 176 (02:14:48)

Code 177 (02:14:48)

Code 178 (02:14:48)

Code 179 (02:14:48)

Code 180 (02:14:48)

Code 181 (02:14:48)

Code 182 (02:14:48)

Code 183 (02:14:48)

Code 184 (02:14:48)

Code 185 (02:14:48)

Code 186 (02:14:48)

Code 187 (02:14:48)

Code 188 (02:14:48)

Code 189 (02:14:48)

Code 190 (02:14:48)

Code 191 (02:14:48)

Code 192 (02:14:48)

Code 193 (02:14:48)

Code 194 (02:14:48)

Code 195 (02:14:48)

Code 196 (02:14:48)

Code 197 (02:14:48)

Code 198 (02:14:48)

Code 199 (02:14:48)

Code 200 (02:14:48)

Code 201 (02:14:48)

Code 202 (02:14:48)

Code 203 (02:14:48)

Code 204 (02:14:48)

Code 205 (02:14:48)

Code 206 (02:14:48)

Code 207 (02:14:48)

Code 208 (02:14:48)

Code 209 (02:14:48)

Code 210 (02:14:48)

Code 211 (02:14:48)

Code 212 (02:14:48)

Code 213 (02:14:48)

Code 214 (02:14:48)

Code 215 (02:14:48)

Code 216 (02:14:48)

Code 217 (02:14:48)

Code 218 (02:14:48)

Code 219 (02:14:48)

Code 220 (02:14:48)

Code 221 (02:14:48)

Code 222 (02:14:48)

Code 223 (02:14:48)

Code 224 (02:14:48)

Code 225 (02:14:48)

Code 226 (02:14:48)

Code 227 (02:14:48)

Code 228 (02:14:48)

Code 229 (02:14:48)

Code 230 (02:14:48)

Code 231 (02:14:48)

Code 232 (02:14:48)

Code 233 (02:14:48)

Code 234 (02:14:48)

Code 235 (02:14:48)

Code 236 (02:14:48)

Code 237 (02:14:48)

Code 238 (02:14:48)

Code 239 (02:14:48)

Code 240 (02:14:48)

Code 241 (02:14:48)

Code 242 (02:14:48)

Code 243 (02:14:48)

Code 244 (02:14:48)

Code 245 (02:14:48)

Code 246 (02:14:48)

Code 247 (02:14:48)

Code 248 (02:14:48)

Code 249 (02:14:48)

Code 250 (02:14:48)

Code 251 (02:14:48)

Code 252 (02:14:48)

Code 253 (02:14:48)

Code 254 (02:14:48)

Code 255 (02:14:48)

Code 256 (02:14:48)

Code 257 (02:14:48)

Code 258 (02:14:48)

Code 259 (02:14:48)

Code 260 (02:14:48)

Code 261 (02:14:48)

Code 262 (02:14:48)

Code 263 (02:14:48)

Code 264 (02:14:48)

Code 265 (02:14:48)

Code 266 (02:14:48)

Code 267 (02:14:48)

Code 268 (02:14:48)

Code 269 (02:14:48)

Code 270 (02:14:48)

Code 271 (02:14:48)

Code 272 (02:14:48)

Code 273 (02:14:48)

Code 274 (02:14:48)

Code 275 (02:14:48)

Code 276 (02:14:48)

Code 277 (02:14:48)

Code 278 (02:14:48)

Code 279 (02:14:48)

Code 280 (02:14:48)

Code 281 (02:14:48)

Code 282 (02:14:48)

Code 283 (02:14:48)

Code 284 (02:14:48)

Code 285 (02:14:48)

Code 286 (02:14:48)

Code 287 (02:14:48)

Code 288 (02:14:48)

Code 289 (02:14:48)

Code 290 (02:14:48)

Code 291 (02:14:48)

Code 292 (02:14:48)

Code 293 (02:14:48)

Code 294 (02:14:48)

Code 295 (02:14:48)

Code 296 (02:14:48)

Code 297 (02:14:48)

Code 298 (02:14:48)

Code 299 (02:14:48)

Code 300 (02:14:48)

Code 301 (02:14:48)

Code 302 (02:14:48)

Code 303 (02:14:48)

Code 304 (02:14:48)

Code 305 (02:14:48)

Code 306 (02:14:48)

Code 307 (02:14:48)

Code 308 (02:14:48)

Code 309 (02:14:48)

Code 310 (02:14:48)

Code 311 (02:14:48)

Code 312 (02:14:48)

Code 313 (02:14:48)

Code 314 (02:14:48)

Code 315 (02:14:48)

Code 316 (02:14:48)

Code 317 (02:14:48)

Code 318 (02:14:48)

Code 319 (02:14:48)

Code 320 (02:14:48)

Code 321 (02:14:48)

Code 322 (02:14:48)

Code 323 (02:14:48)

Code 324 (02:14:48)

Code 325 (02:14:48)

Code 326 (02:14:48)

Code 327 (02:14:48)

Code 328 (02:14:48)

Code 329 (02:14:48)

Code 330 (02:14:48)

Code 331 (02:14:48)

Code 332 (02:14:48)

Code 333 (02:14:48)

Code 334 (02:14:48)

Code 335 (02:14:48)

Code 336 (02:14:48)

Code 337 (02:14:48)

Code 338 (02:14:48)

Code 339 (02:14:48)

Code 340 (02:14:48)

Code 341 (02:14:48)

Code 342 (02:14:48)

Code 343 (02:14:48)

Code 344 (02:14:48)

Code 345 (02:14:48)

Code 346 (02:14:48)

Code 347 (02:14:48)

Code 348 (02:14:48)

Code 349 (02:14:48)

Code 350 (02:14:48)

Code 351 (02:14:48)

Code 352 (02:14:48)

Code 353 (02:14:48)

Code 354 (02:14:48)

Code 355 (02:14:48)

Code 356 (02:14:48)

Code 357 (02:14:48)

Code 358 (02:14:48)

Code 359 (02:14:48)

Code 360 (02:14:48)

Code 361 (02:14:48)

Code 362 (02:14:48)

Code 363 (02:14:48)

Code 364 (02:14:48)

Code 365 (02:14:48)

Code 366 (02:14:48)

Code 367 (02:14:48)

Code 368 (02:14:48)

Code 369 (02:14:48)

Code 370 (02:14:48)

Code 371 (02:14:48)

Code 372 (02:14:48)

Code 373 (02:14:48)

Code 374 (02:14:48)

Code 375 (02:14:48)

Code 376 (02:14:48)

Code 377 (02:14:48)

Code 378 (02:14:48)

Code 379 (02:14:48)

Code 380 (02:14:48)

Code 381 (02:14:48)

Code 382 (02:14:48)

Code 383 (02:14:48)

Code 384 (02:14:48)

Code 385 (02:14:48)

Code 386 (02:14:48)

Code 387 (02:14:48)

Code 388 (02:14:48)

Code 389 (02:14:48)

Code 390 (02:14:48)

Code 391 (02:14:48)

Code 392 (02:14:48)

Code 393 (02:14:48)

Code 394 (02:14:48)

Code 395 (02:14:48)

Code 396 (02:14:48)

Code 397 (02:14:48)

Code 398 (02:14:48)

Code 399 (02:14:48)

Code 400 (02:14:48)

Code 401 (02:14:48)

Code 402 (02:14:48)

Code 403 (02:14:48)

Code 404 (02:14:48)

Code 405 (02:14:48)

Code 406 (02:14:48)

Code 407 (02:14:48)

Code 408 (02:14:48)

Code 409 (02:14:48)

Code 410 (02:14:48)

Code 411 (02:14:48)

Code 412 (02:14:48)

Code 413 (02:14:48)

Code 414 (02:14:48)

Code 415 (02:14:48)

Code 416 (02:14:48)

Code 417 (02:14:48)

Code 418 (02:14:48)

Code 419 (02:14:48)

Code 420 (02:14:48)

Code 421 (02:14:48)

Code 422 (02:14:48)

Code 423 (02:14:48)

Code 424 (02:14:48)

Code 425 (02:14:48)

Code 426 (02:14:48)

Code 427 (02:14:48)

Code 428 (02:14:48)

Code 429 (02:14:48)

Code 430 (02:14:48)

Code 431 (02:14:48)

Code 432 (02:14:48)

Code 433 (02:14:48)

Code 434 (02:14:48)

Code 435 (02:14:48)

Code 436 (02:14:48)

Code 437 (02:14:48)

Code 438 (02:14:48)

Code 439 (02:14:48)

Code 440 (02:14:48)

Code 441 (02:14:48)

Code 442 (02:14:48)

Code 443 (02:14:48)

Code 444 (02:14:48)

Code 445 (02:14:48)

Code 446 (02:14:48)

Code 447 (02:14:48)

Code 448 (02:14:48)

Code 449 (02:14:48)

Code 450 (02:14:48)

Code 451 (02:14:48)

Code 452 (02:14:48)

Code 453 (02:14:48)

Code 454 (02:14:48)

Code 455 (02:14:48)

Code 456 (02:14:48)

Code 457 (02:14:48)

Code 458 (02:14:48)

Code 459 (02:14:48)

Code 460 (02:14:48)

Code 461 (02:14:48)

Code 462 (02:14:48)

Code 463 (02:14:48)

Code 464 (02:14:48)

Code 465 (02:14:48)

Code 466 (02:14:48)

Code 467 (02:14:48)

Code 468 (02:14:48)

Code 469 (02:14:48)

Code 470 (02:14:48)

Code 471 (02:14:48)

Code 472 (02:14:48)

Code 473 (02:14:48)

Code 474 (02:14:48)

Code 475 (02:14:48)

Code 476 (02:14:48)

Code 477 (02:14:48)

Code 478 (02:14:48)

Code 479 (02:14:48)

Code 480 (02:14:48)

Code 481 (02:14:48)

Code 482 (02:14:48)

Code 483 (02:14:48)

Code 484 (02:14:48)

Code 485 (02:14:48)

Code 486 (02:14:48)

Code 487 (02:14:48)

Code 488 (02:14:48)

Code 489 (02:14:48)

Code 490 (02:14:48)

Code 491 (02:14:48)

Code 492 (02:14:48)

Code 493 (02:14:48)

Code 494 (02:14:48)

Code 495 (02:14:48)

Code 496 (02:14:48)

Code 497 (02:14:48)

Code 498 (02:14:48)

Code 499 (02:14:48)

Code 500 (02:14:48)

Code 501 (02:14:48)

Code 502 (02:14:48)

Code 503 (02:14:48)

Code 504 (02:14:48)

Code 505 (02:14:48)

Code 506 (02:14:48)

Code 507 (02:14:48)

Code 508 (02:14:48)

Code 509 (02:14:48)

Code 510 (02:14:48)

Code 511 (02:14:48)

Code 512 (02:14:48)

Code 513 (02:14:48)

Code 514 (02:14:48)

Code 515 (02:14:48)

Code 516 (02:14:48)

Code 517 (02:14:48)

Code 518 (02:14:48)

Code 519 (02:14:48)

Code 520 (02:14:48)

Code 521 (02:14:48)

Code 522 (02:14:48)

Code 523 (02:14:48)

Code 524 (02:14:48)

Code 525 (02:14:48)

Code 526 (02:14:48)

Code 527 (02:14:48)

Code 528 (02:14:48)

Code 529 (02:14:48)

Code 530 (02:14:48)

Code 531 (02:14:48)

Code 532 (02:14:48)

Code 533 (02:14:48)

Code 534 (02:14:48)

Code 535 (02:14:48)

Code 536 (02:14:48)

Code 537 (02:14:48)

Code 538 (02:14:48)

Code 539 (02:14:48)

Code 540 (02:14:48)

Code 541 (02:14:48)

Code 542 (02:14:48)

Code 543 (02:14:48)

Code 544 (02:14:48)

Code 545 (02:14:48)

Code 546 (02:14:48)

Code 547 (02:14:48)

Code 548 (02:14:48)

Code 549 (02:14:48)

Code 550 (02:14:48)

Code 551 (02:14:48)

Code 552 (02:14:48)

Code 553 (02:14:48)

Code 554 (02:14:48)

Code 555 (02:14:48)

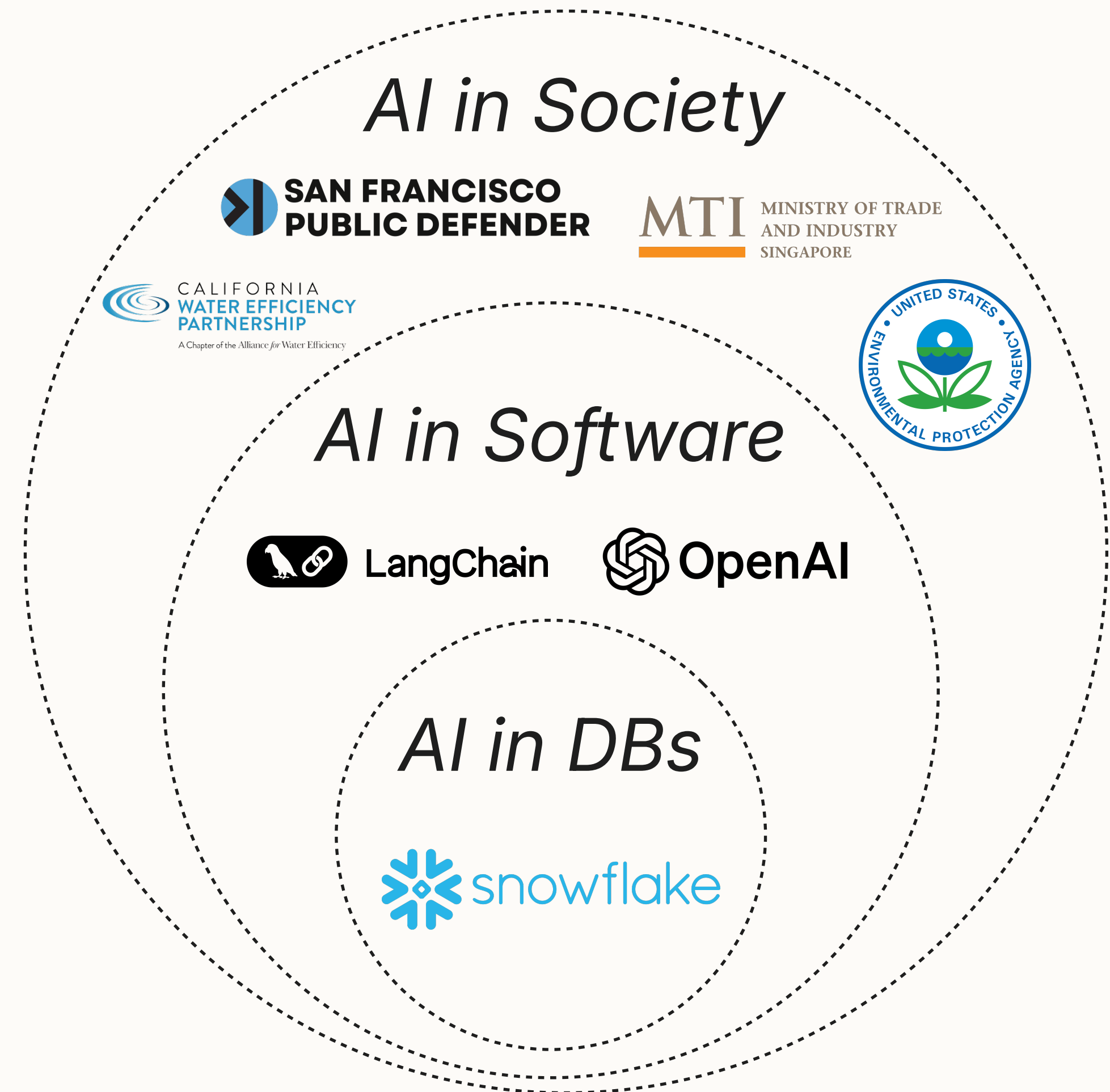
Code 556 (02:

Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)



Today's Talk

How can we measure the **reliability** of AI-generated outputs? *EvalGen & a course (3.5k practitioners)*

**How can we evaluate LLM outputs
reliably and cheaply at scale?**

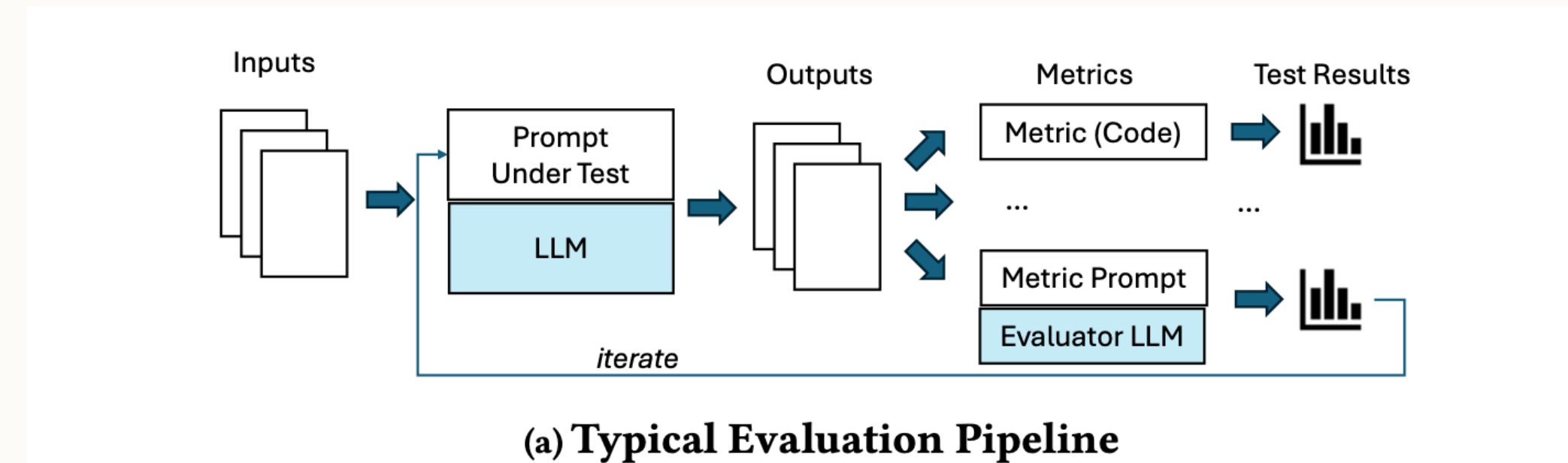
How can we evaluate LLM outputs reliably and cheaply at scale?

Algorithms are one solution.

My question: what can the *interface* do?

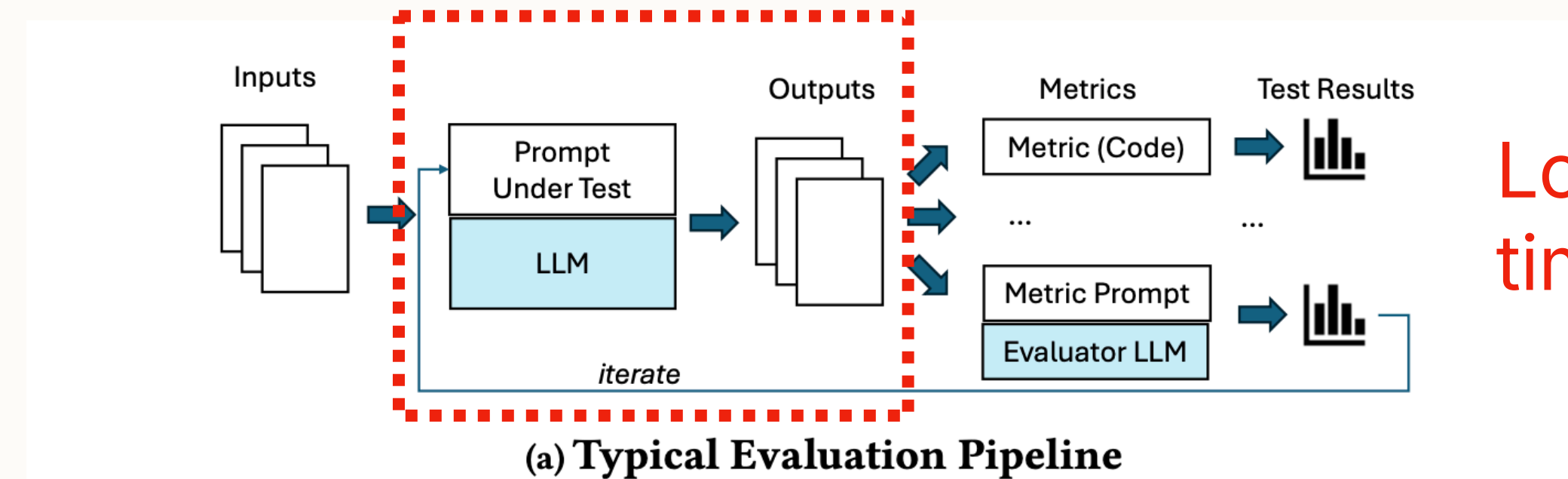
Idea: Get Annotations During Wait Time

Who Validates the Validators? **Shankar** et al. *Most cited paper at UIST '24.*



Idea: Get Annotations During Wait Time

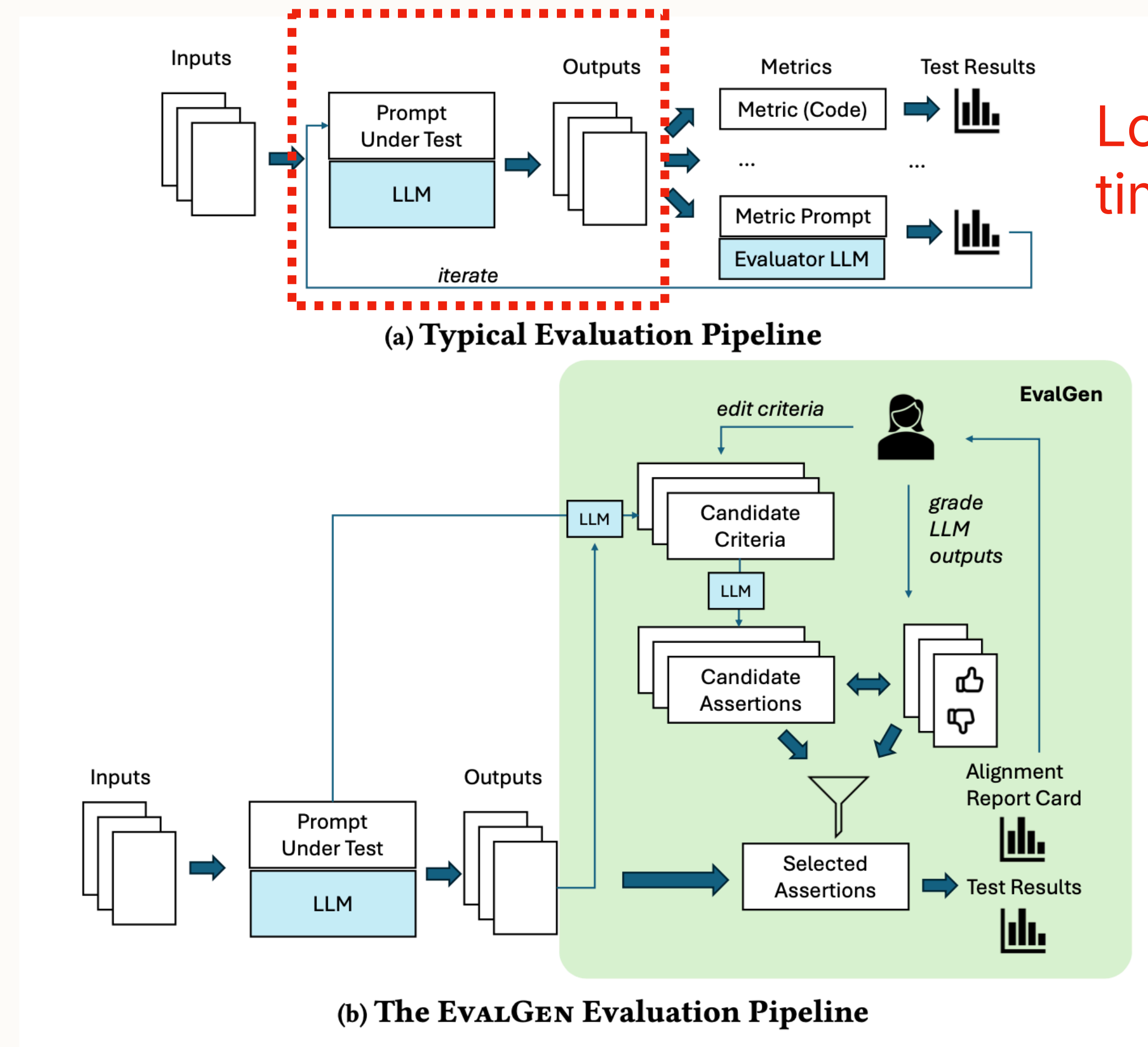
Who Validates the Validators? **Shankar** et al. *Most cited paper at UIST '24.*



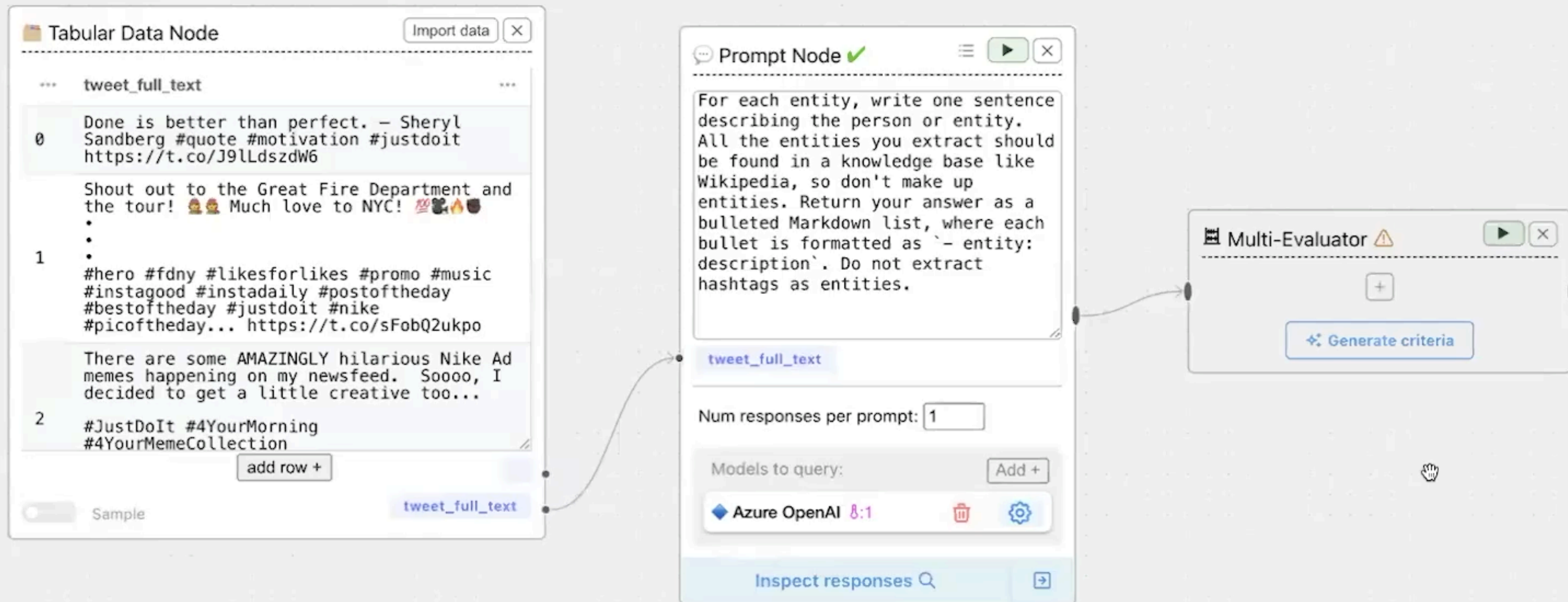
Lots of wait time here

Idea: Get Annotations During Wait Time

Who Validates the Validators? **Shankar** et al. *Most cited paper at* **UIST '24**.



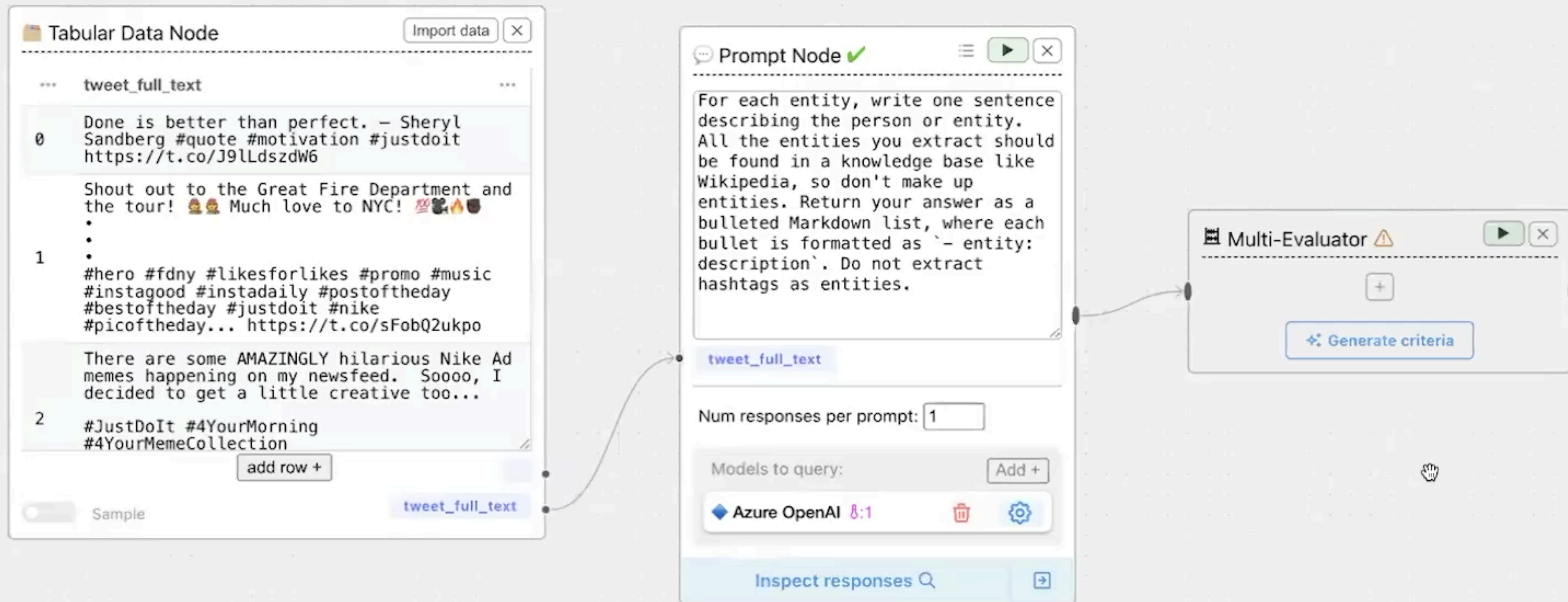
Lots of wait time here



[Send us feedback](#)

React Flow

Who Validates the Validators? **Shankar et al.** *Most cited paper at* **UIST '24**.




[Send us feedback](#)

React Flow

Who Validates the Validators? **Shankar et al.** *Most cited paper at* **UIST '24**.

Adoption from Tech Companies

 LangChain

Case StudiesIn the LoopLangChainDocsChangelogSign inSubscribe

Motivating research

There were two pieces of motivating research that led us to implement a solution.

The first piece is nothing new: language models are adept at few-shot learning. If you give LLMs examples of things done correctly, they will imitate the correct behavior. This method is widely-adopted in our client LLM applications; it's particularly effective in cases where it's tough to explain in instructions how the LLM should behave, and where the output is expected to have a particular format. Evaluations fit both these criteria!

The other piece of research is new: a paper out of Berkeley by Shreya Shankar titled Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. This paper addresses the same problem, and, though it proposes a different solution than ours, it helped motivate our usage of feedback collection as a way to programmatically align LLM evaluations with human preferences.

So - how did we take these two ideas and build our “self-improving” evaluators?

Generative Benchmarking

answering tasks. This contrasts with the ambiguity of real data, where queries are often vague with only partial matches to their relevant documents. Lastly, most embedding models have likely seen these benchmark datasets during training, which makes it difficult to distinguish true retrieval capabilities from memorization.

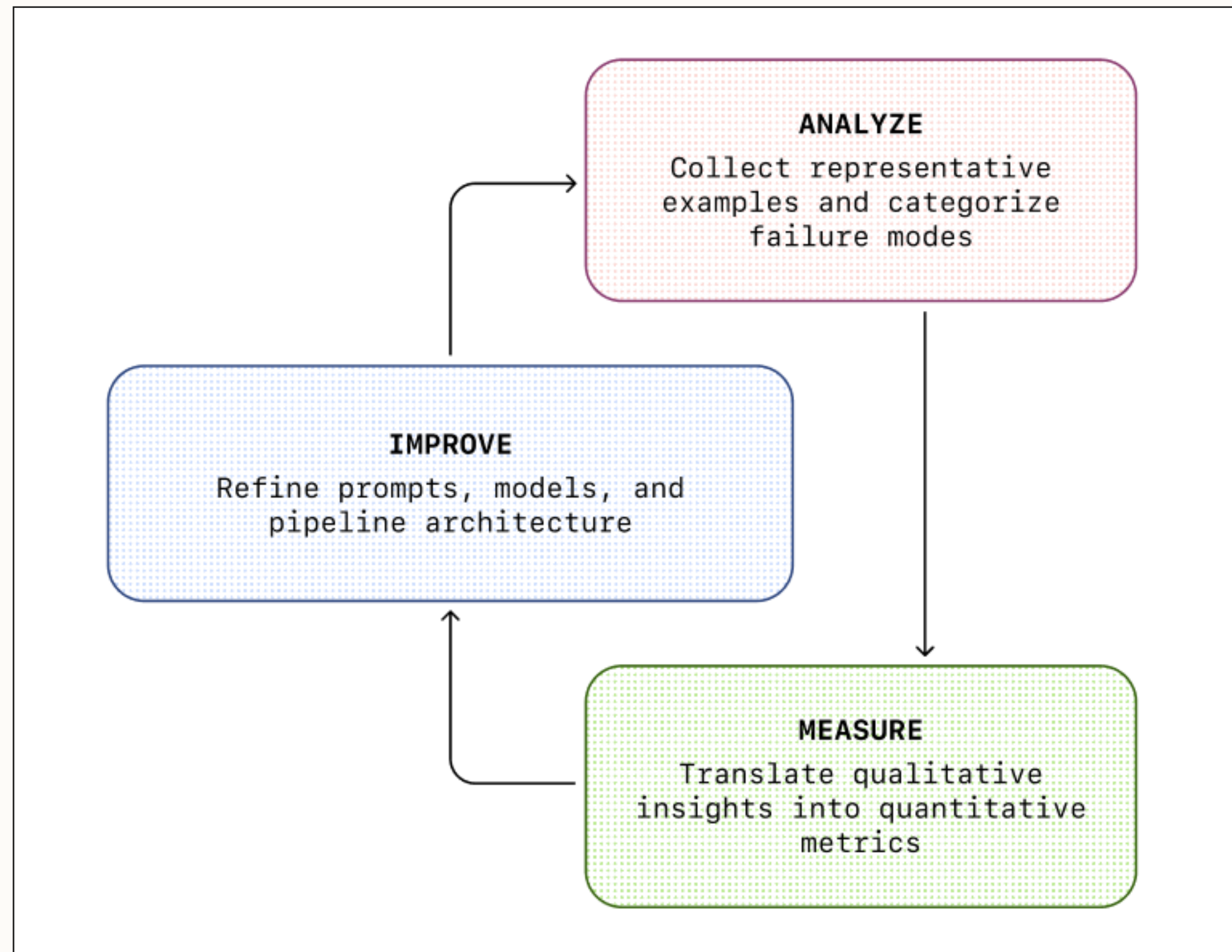
Synthetic dataset generation for information retrieval has been an active area of research, such as the work by InPars [4] and Promptagator [5], which leverage few-shot prompting for generating synthetic training data. These methods are aimed at improving retrieval models, often focusing on improvements in metrics such as Recall. This contrasts with our motivation to evaluate models in a more realistic manner rather than to improve them.

RAGAS [6] and AIR-Bench [7] are more aligned to our focus on evaluation, as both aim to generate testsets for evaluating retrieval in real-world scenarios. Both approaches focus on the diversity of generated testsets, with RAGAS focusing on diversity in query type and AIR-Bench aiming for diversity in domains beyond public benchmarks. However, an area that has not been extensively explored in prior work is how well these synthetically generated queries represent real user queries from production. Our experiments focus on synthetically generating queries that are representative of the ground truth, which we believe is the objective of a golden dataset when evaluating retrieval systems.

In our work, we also employ Large Language Models (LLMs) as judges for labeling tasks. LLM judges allow for a cost-effective and consistent way of labeling data, however, they come with known problems around alignment. We cannot guarantee that LLMs will have the same judgments as humans would, largely due to their high sensitivity to minor changes in prompting and the difficulty in articulating ambiguous concepts such as “relevance”.

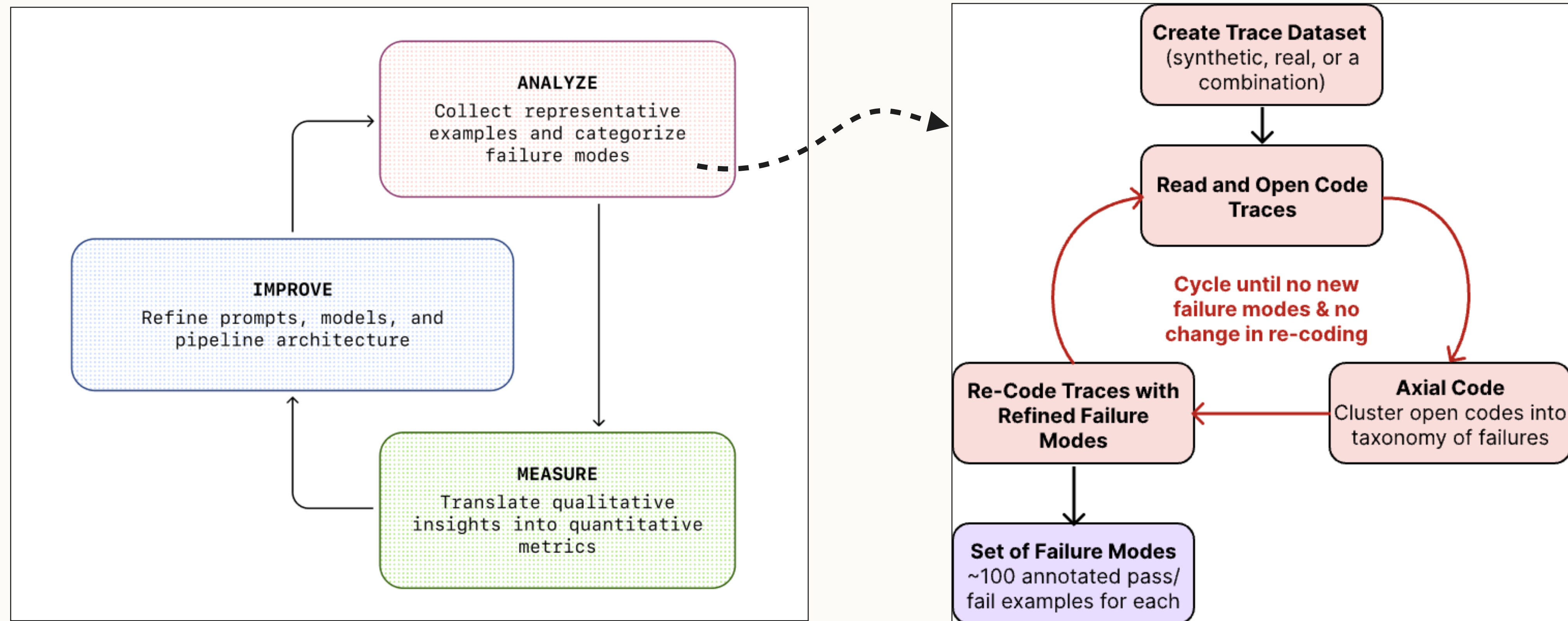
To better align our LLM judge with human judgements, we use an adapted version of EvalGEN [8]. EvalGEN is a framework for validating LLM outputs through iterating on a set of criteria based human inputs. We use this process to align our LLM judge for document filtering, the first step in our generative benchmarking process.

Teaching Evals to 3.5k+ Practitioners



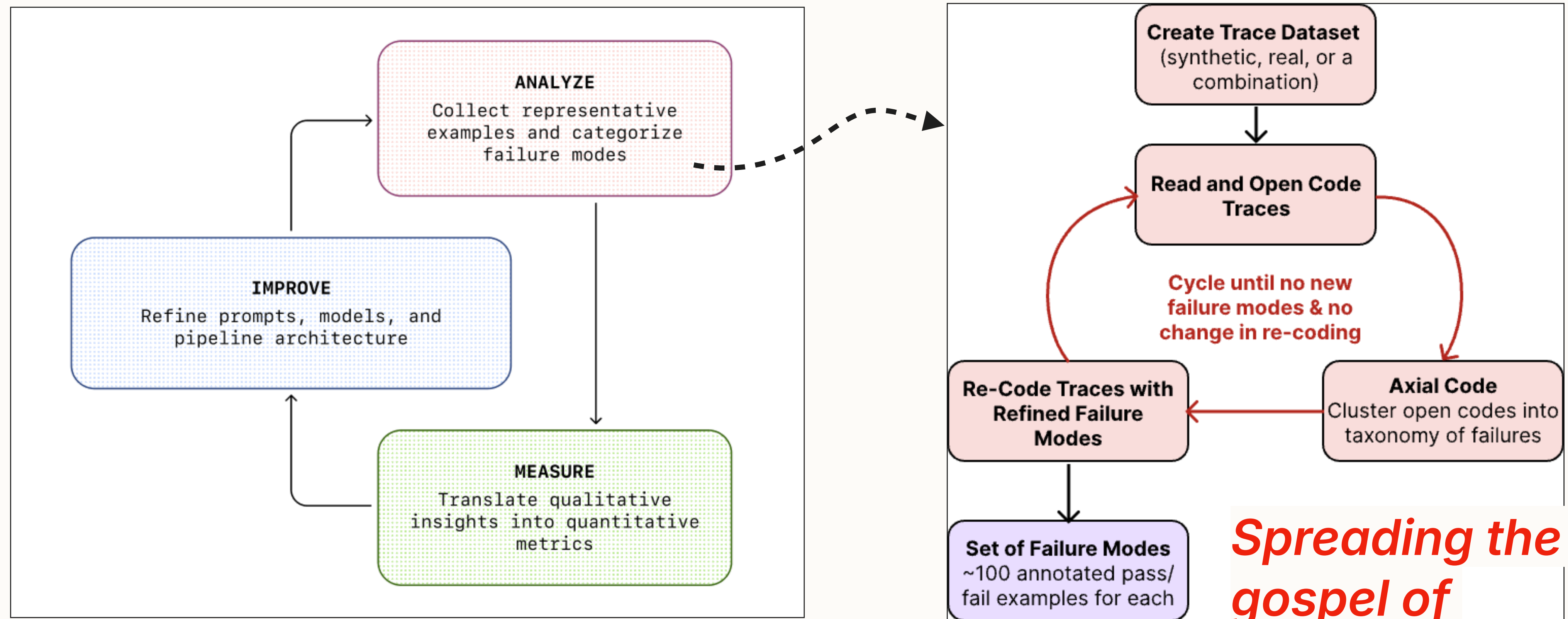
Kapse et al., "Building Resilient Prompts Using an Evaluation Flywheel", OpenAI Cookbook, Oct. 6 2025.

Teaching Evals to 3.5k+ Practitioners



Kapse et al., "Building Resilient Prompts Using an Evaluation Flywheel", OpenAI Cookbook, Oct. 6 2025.

Teaching Evals to 3.5k+ Practitioners



Kapse et al., "Building Resilient Prompts Using an Evaluation Flywheel", OpenAI Cookbook, Oct. 6 2025.

Teaching Evals to 3.5k+ Practitioners

★★★★★

If you are building AI infused product or service, then understanding how build credible evals is fundamental. I always say “vibe check ≠ validation”. In order to build credibility evals, you need to deeply understand the workflow of build evals for your AI software. Hamel & Shreya’s did exactly for me, it gave me structure, vocabulary and workflow to build evals with confidence. I feel more in control of my product destiny as a result of attending this course. I would highly recommend this course to all AI Builders.



Ashish COHORT 3
Senior Product Manager
Audible (Amazon)

★★★★★

This course is incredibly valuable! Thoughtful and clear coursework, excellent resources, brilliant instructors. I am excited to incorporate the learnings, and refer back to the sessions in the future! Thank you Hamel and Shreya!



Liz COHORT 3
Software Engineer, Dev AI
Airbnb

★★★★★

Excellent Course - Highly Recommend I took this course with beginner/intermediate knowledge and found it valuable for all skill levels. Hamel and Shreya are outstanding instructors who give thoughtful answers to everyone's questions. The curriculum is very comprehensive, covering practical exercises, theoretical frameworks, benchmarking, and real-world applications. I learned as much from other students' questions and challenges as I did from the course material itself. I genuinely leveled up taking this course. Highly recommended whether you're just starting out or looking to evaluate your current process.



Kara COHORT 3
Senior Technical Product Manager
Comcast

★★★★★

Before taking the course, I didn't realise how detailed and involved creating effective evaluations for AI applications can be. Now that I have completed the course, I thought I had a good understanding of evaluations, but I discovered several gaps in my knowledge that would...

[Read more](#)




JO COHORT 3
Senior AI Engineer
JO

Teaching Evals to 3.5k+ Practitioners

★★★★★


If you are building AI infused product or service, then u
build credible evals is fundamental. I always say “vibe check + validation”.
In order to build credibility evals, you need to deeply understand the
workflow of build evals for your AI software. Hamel & Shreya’s did exactly
for me, it gave me structure, vocabulary and workflow to build evals with
confidence. I feel more in control of my product destiny as a result of
attending this course. I would highly recommend this course to all AI
Builders.



Ashish COHORT 3
Senior Product Manager
Audible (Amazon)

★★★★★

This course is incredibly valuable! Thoughtful and clear coursework,
excellent resources, brilliant instructors. I am excited to incorporate the
learnings, and refer back to the sessions in the future! Thank you Hamel
and Shreya!




Liz COHORT 3
Software Engineer, Dev AI
Airbnb

★★★★☆ 4.7 (624)

Highly Recommend I took this course with
knowledge and found it valuable for all skill levels.

Hamel and Shreya are outstanding instructors who give thoughtful
answers to everyone's questions. The curriculum is very comprehensive,
covering practical exercises, theoretical frameworks, benchmarking, and
real-world applications. I learned as much from other students' questions
and challenges as I did from the course material itself. I genuinely leveled
up taking this course. Highly recommended whether you're just starting
out or looking to evaluate your current process.




Kara COHORT 3
Senior Technical Product Manager
Comcast

★★★★★

Before taking the course, I didn't realise how detailed and involved
creating effective evaluations for AI applications can be. Now that I have
completed the course, I thought I had a good understanding of
evaluations, but I discovered several gaps in my knowledge that would...

[Read more](#)



JO COHORT 3
Senior AI Engineer
JO

Teaching Evals to 3.5k+ Practitioners

★★★★★

If you are building AI infused product or service, then u
build credible evals is fundamental. I always say “vibe check + validation”.
In order to build credibility evals, you need to deeply understand the
workflow of build evals for your AI software. Hamel & Shreya’s did exactly
for me, it gave me structure, vocabulary and workflow to build evals with
confidence. I feel more in control of my product destiny as a result of
attending this course. I would
Builders.

Ashish

COHORT 3

Senior Product Manager
Audible (Amazon)

★★★★★

This course is incredibly valuable
excellent resources, brilliant
learnings, and refer back to
and Shreya!

Liz

COHORT 3

Software Engineer, Dev
Airbnb

★★★★★

4.7 (624)

BOOK

Evals for AI Engineers

by Shreya Shankar, Hamel Husain

October 2026

Intermediate To
Advanced

275 Pages

Early Release

RAW & UNEDITED

Shreya Shankar
& Hamel Husain

Senior AI Engineer
JO

Highly Recommend I took this course with
knowledge and found it valuable for all skill levels.
Hamel and Shreya are outstanding instructors who give thoughtful
answers to everyone's questions. The curriculum is very comprehensive,
covering practical exercises, theoretical frameworks, benchmarking, and
real-world applications. I learned as much from other students' questions
myself. I genuinely leveled
up after you're just starting
and involved
can be. Now that I have
standing of
knowledge that would...

Building Effective AI-Powered Data Systems. Shreya Shankar, 2025.

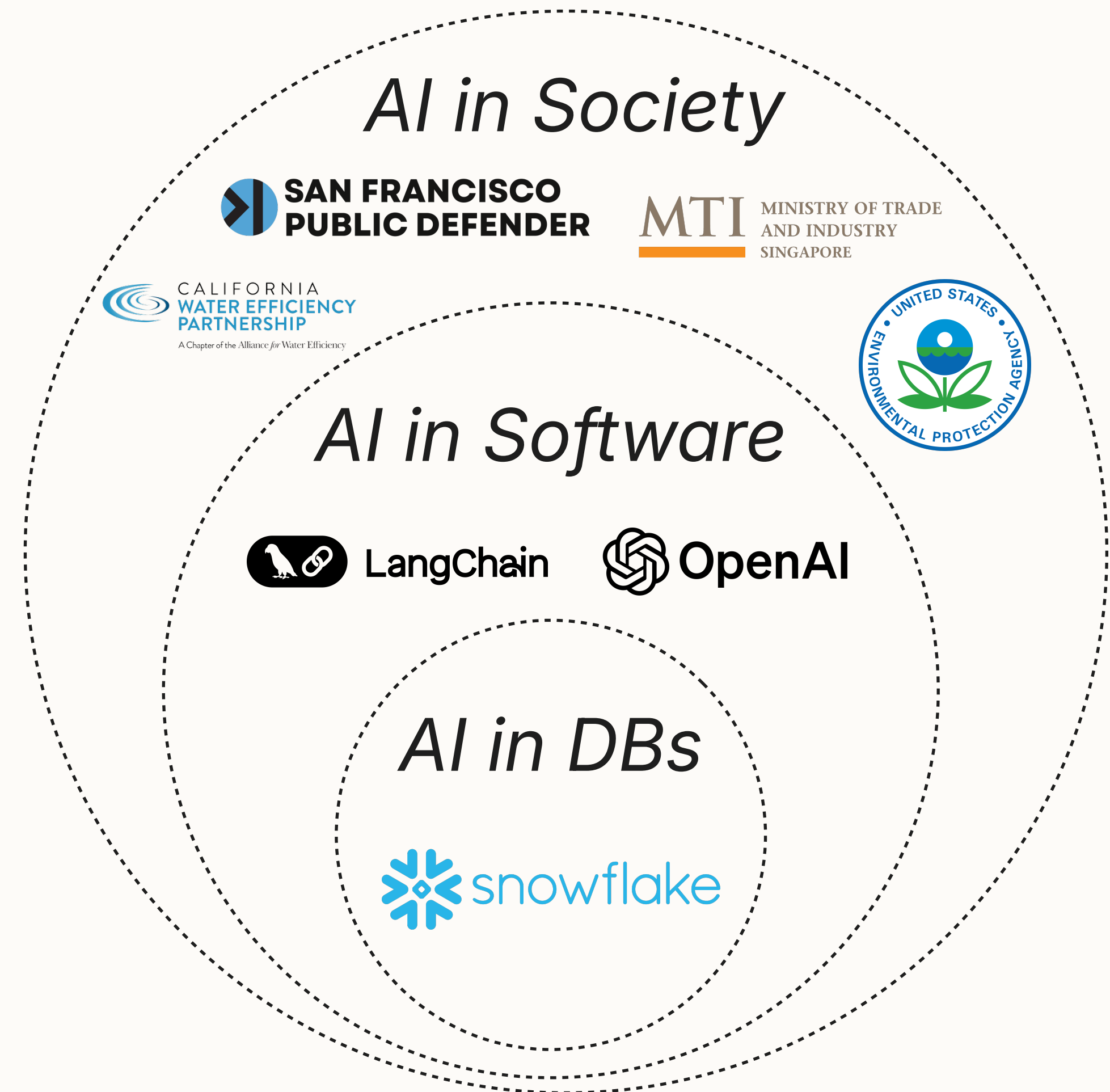
78

Today's Talk

How can data systems **reason** over unstructured data? *DocETL* (3.1k ★)

How can users **steer and debug** data systems that expose AI capabilities? *DocWrangler* (UIST 🏆)

How can we measure the **reliability** of AI-generated outputs? *EvalGen* & a course (3.5k practitioners)



Future: Full-Stack for Unstructured Data

Future: Full-Stack for Unstructured Data

DocETL

Interface

Optimizer

Execution Engine

Storage

Future: Full-Stack for Unstructured Data

DocETL

Interface

Optimizer

*Takes a long
time*

Execution Engine

*Opportunities for
acceleration*

Storage

*Data is too
big*

Future: Full-Stack for Unstructured Data

DocETL

Interface

Optimizer

*Takes a long
time*

*RL agents to learn
query rewrites*

Execution Engine

*Opportunities for
acceleration*

*Open-source LMs;
custom KV caching*

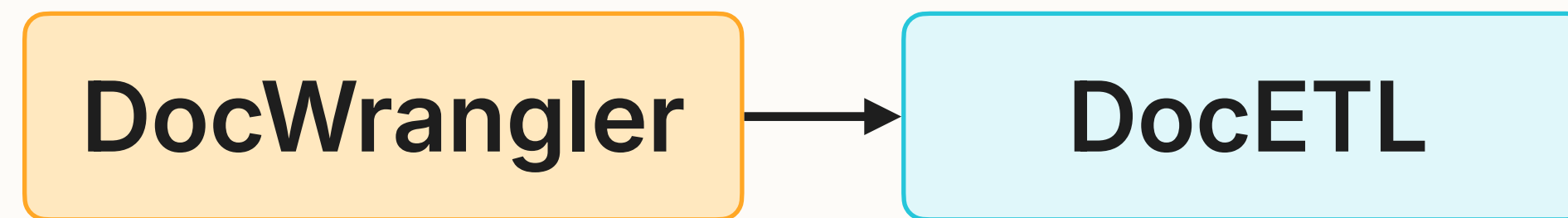
Storage

*Data is too
big*

*Compression for
unstructured data*

Future: Full-Stack for Unstructured Data

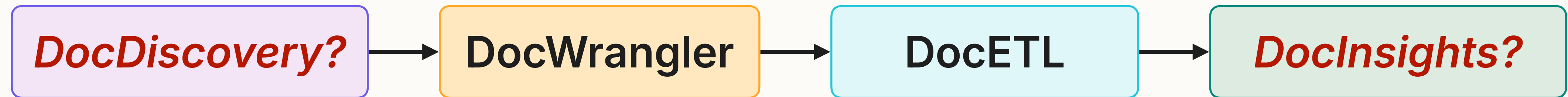
Future: Full-Stack for Unstructured Data



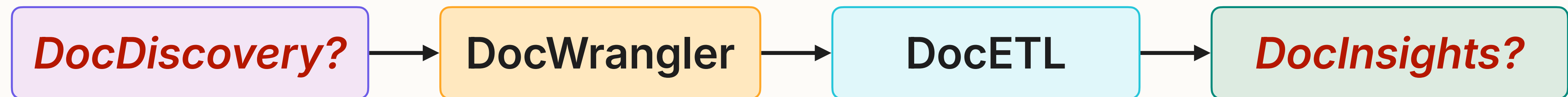
Future: Full-Stack for Unstructured Data



Future: Full-Stack for Unstructured Data



Future: Full-Stack for Unstructured Data



Lifecycle gaps:

- ◆ Before: tools to help decide *what* to analyze
- ◆ During: *collaborative* environments for iterative refinement
- ◆ After: interfaces to communicate interpretive, non-deterministic findings

Moonshot: How AI Reshapes Knowledge Work

Two observations from DocETL:

- ◆ AI capabilities shape **questions**
- ◆ AI capabilities shape **what types of data can exist**

How do workflows evolve as users discover AI possibilities and limitations?

Can we build systems to support newly emerging research methodologies?

Requires collaboration: e.g., economists, policymakers, scientists, social scientists, theorists

Building Effective AI-Powered Data Systems



Shreya Shankar

UC Berkeley EECS

November 2025

docetl.org



A system for LLM-powered data processing

Email: shreyashankar@berkeley.edu

Website: sh-reya.com

