



# SURVEY FUTURES

SURVEY DATA COLLECTION  
METHODS COLLABORATION

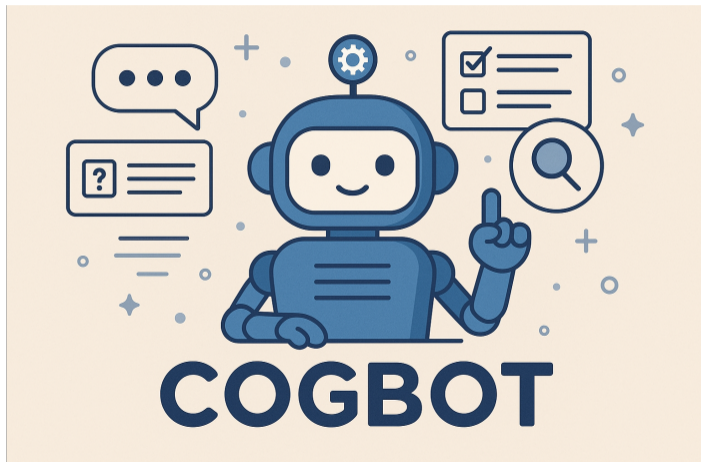
## LLMs for Survey Pretesting: How Well Can Large Language Models Identify Flaws in Survey Questions?

Survey Futures International Conference, City University, London

Patrick Sturgis   Tom Robinson   Caroline Roberts

17–18 June 2026

# Cogbot: AI cognitive interviewing



## Cogbot: the problem

---

- Cognitive interviewing and expert review are the gold standard for question pre-testing
- Workflow bottleneck: pretesting is slow and expensive
- Result: many questions ship to field with undetected problems

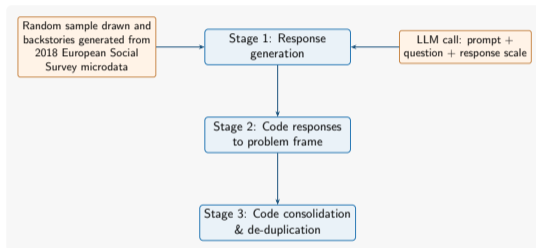
**Can LLMs simulate human pre-testing methods at scale?**



# Cognitive interview configuration

Four stages:

1. User specifies (random) sample size to be drawn from ESS backstories
2. Synthetic respondents produce think-aloud transcripts
3. Analyst LLM codes problems from transcripts
4. Synthesis stage consolidates codes



## Expert review configuration

---

- Three independent LLM reviewers evaluate each question for problems
- Analogous to expert appraisal, where experienced survey methodologists review questions against known design problems
- Each reviewer rates problem severity (1–10 scale); final score is the mean across reviewers
- Synthesis step merges duplicates and forms compound codes from overlapping problems



# Prompt variants

---

## Cognitive testing (CI)

*Synthetic respondent thinks aloud while answering*

- **CI-Open:** free think-aloud; no cognitive framework or problem types in prompt
- **CI-Structured:** works through the four Tourangeau–Cannell stages (comprehension, retrieval, judgement, response); no problem types named
- **CI-List:** as CI-Structured, plus an explicit list of common problem types provided at each stage

## Expert review (ER)

*Three reviewer LLMs evaluate the question*

- **ER-Open:** identify problems in own terms; no predefined types, no restraint language
- **ER-Restrained:** as ER-Open, but with strong restraint language: report only problems genuinely likely to affect respondents; return empty list if the question is well designed
- **ER-Checklist:** evaluate against a seven-type checklist (comprehension, double-barrelled, retrieval, response mapping, social desirability, presupposition, tautology)



## Evaluation design

---

- Test set: 27 survey questions
- **20 deliberately flawed items**, each containing 1–3 known problems (48 total)
  - Double-barrelled, presupposition, vague wording, response-mapping failures, retrieval difficulty, leading framing, sensitivity, complex syntax
  - Embedded problems constitute the answer key
- **7 ESS control items**: well-tested core questions fielded across 30+ countries
  - A well-functioning tool should report few or no problems on these
- Two metrics:
  - **Detection rate**: proportion of 48 known problems identified
  - **False positive rate**: problems flagged beyond the answer key



# Illustrative pipeline output

“Overall, how satisfied are you with your terms of employment?”

---

## Simulated cognitive testing (CI-Open)

### Respondent transcript

*Respondent 2 (retired, former shopkeeper):* “As I read the question about my satisfaction with my terms of employment, I first think about my situation. Being retired and a former shopkeeper, I’m not currently employed in the traditional sense, so the question feels a bit out of place for me. However, I can think back to when I was actively running my shop. [...] Since I’m retired, none of these apply directly to me anymore.”

*Respondent 8 (kitchen helper, female, 54):* “As I read the question [...] I start to think about my current job as a kitchen helper. I consider the hours I work, the pay I receive, and the conditions I work under. These are what I think about when I hear ‘terms of employment.’ [...] The phrasing ‘terms of employment’ could be a bit vague, but I interpret it as encompassing these aspects: pay, hours, and work conditions.”

### Analyst coding

The respondent is retired and self-employed, making the question about “terms of employment” feel irrelevant. They had to reflect on past experiences and adapt the question to fit their unique situation, which could lead to inaccuracies. (Severity: 7/10)

The respondent found the term “terms of employment” to be vague, which might have led to some uncertainty about what aspects to consider when answering the question. (Severity: 5/10)

---



# Illustrative pipeline output (cont.)

## “Overall, how satisfied are you with your terms of employment?”

### Expert review (ER-Checklist)

Expert reviewer assessment

Consolidated problem code

*Comprehension* (severity 6): “The phrase ‘terms of employment’ may be interpreted differently by respondents, potentially including salary, benefits, work hours, or job security.”

(1) “Terms of employment” is ambiguous (mean severity 5.7; flagged by all 3 reviewers).

*Presupposition* (severity 5): “The question assumes that respondents have defined and clear terms of employment, which may not be true for all, such as temporary or freelance workers.”

(2) The question assumes defined employment terms (mean severity 5.0; flagged by 1 reviewer).



# Overall results (GPT-4o)

Configuration	Set	Detection	FP/Q	FP/Q ( $\geq 5$ )	Confidence
CI-List	Flawed	75.0%	1.8	1.50	4.08
	ESS	—	1.6	1.00	4.44
ER-Checklist	Flawed	72.9%	1.4	1.05	—
	ESS	—	2.3	0.57	—
ER-Restrained	Flawed	68.8%	1.4	1.40	—
	ESS	—	3.1	3.14	—
CI-Open	Flawed	60.4%	1.4	1.00	4.21
	ESS	—	1.4	1.00	4.41
ER-Open	Flawed	47.9%	0.8	0.55	—
	ESS	—	0.3	0.29	—
CI-Structured	Flawed	31.2%	0.6	0.50	4.28
	ESS	—	0.7	0.57	4.42



## Detection rate (%) by problem type (GPT-4o)

Problem type	N	Cognitive testing			Expert review			Severity	Conf.
		Open	Struct.	List	Checklist	Open	Restr.		
Double-barrelled	6	100	67	100	100	100	100	7.18	4.07
Presupposition	9	44	33	44	89	11	56	6.18	3.91
Vague/ambiguous	9	56	11	78	67	67	78	5.57	4.37
Response-mapping	7	57	43	71	43	0	29	5.33	4.30
Retrieval/recall	5	100	0	100	100	80	80	5.64	4.32
Leading/loaded	5	40	20	100	80	40	80	6.33	3.80
Sensitivity	4	25	50	50	50	50	50	6.07	4.29
Complex syntax	3	67	33	67	33	67	100	5.97	4.04

Severity = mean LLM-assigned severity (1–10) for matched true positives. Conf. = mean respondent confidence (1–5), CI configurations only.



## Three-model comparison

Config	GPT-4o		Llama-3.3-70B		Qwen3-32B	
	Det.	FP/ESS	Det.	FP/ESS	Det.	FP/ESS
CI-Open	60.4%	1.6	83.3%	5.4	83.3%	7.0
CI-Structured	31.2%	0.6	85.4%	4.0	87.5%	5.2
CI-List	75.0%	1.8	91.7%	5.4	83.3%	6.2
ER-Checklist	72.9%	2.6	89.6%	3.5	77.1%	2.9
ER-Open	47.9%	0.4	47.9%	0.8	6.2%	0.1
ER-Restrained	68.8%	3.1	75.0%	3.1	64.6%	2.0

Det. = detection rate (% of 48 known problems). FP/ESS = false positives per ESS control question. Open-weight models achieve higher detection but with 3–7× more false positives.



## Discussion: Cogbot

---

- Best configuration (CI-List) detects 75% of known problems; after adjudication, most “false positives” are genuine problems the answer key missed
- Simulated cognitive testing and expert review are complementary: CI stronger on response-mapping failures, ER stronger on presupposition and leading framing
- Prompt design is model-dependent: prompts optimised for GPT-4o do not transfer to open-weight models without recalibration
- Full 20-item questionnaire evaluation: under 1 hour,  $\approx$ \$10 (cognitive testing) or minutes and  $<$ \$1 (expert review)
- Not a replacement for skilled cognitive interviewing, but extends diagnostic coverage to contexts where human pretesting is constrained by budget, timeline, or item volume



## Limitations and future directions

---

- Test set uses deliberately embedded, well-defined defects; real defects are messier
- Detection rates likely an *upper bound*
- Pipeline is non-interactive; real cognitive interviews are dynamic
- Test sensitivity to sample size and model temperature
- Model landscape and capabilities changing rapidly



Thank you

---

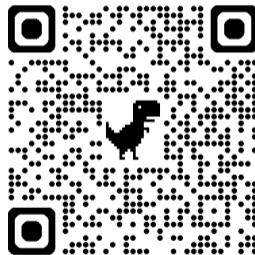
## Questions?

Patrick Sturgis

[p.sturgis@lse.ac.uk](mailto:p.sturgis@lse.ac.uk)

Department of Methodology  
London School of Economics and Political  
Science

Survey Futures International Conference,  
City University, London, 17–18 June 2026



Working paper