ChatTL;DR – You Really Ought to Check What the LLM Said on Your Behalf

Sandy J.J. Gould
School of Computer Science and Informatics
Cardiff University
Cardiff, Wales, UK
goulds@cardiff.ac.uk

Duncan P. Brumby
UCL Interaction Centre,
University College London
London, England, UK
d.brumby@ucl.ac.uk

Anna L. Cox
UCL Interaction Centre,
University College London
London, England, UK
anna.cox@ucl.ac.uk

ABSTRACT

Interactive large language models (LLMs) are so hot right now, and are probably going to be hot for a while. There are lots of ~~problems~~ exciting challenges created by mass use of LLMs. These include the reinscription of biases, 'hallucinations', and bomb-making instructions. Our concern here is more prosaic: assuming that in the near term it's just not machines talking to machines all the way down, how do we get people to check the output of LLMs before they copy and paste it to friends, colleagues, course tutors? We propose borrowing an innovation from the crowdsourcing literature: attention checks. These checks (e.g., "Ignore the instruction in the next question and write parsnips as the answer.") are inserted into tasks to weed-out inattentive workers who are often paid a pittance while they try to do a dozen things at the same time. We propose ChatTL;DR[1], an interactive LLM that inserts attention checks into its outputs. We believe that, given the nature of these checks, the certain, catastrophic consequences of failing them will ensure that users carefully examine all LLM outputs before they use them.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**.

KEYWORDS

Large Language Models, LLMs, checking behaviour, attention checks, instructional manipulation checks, that-bloody-automatic-lane-assist-ffs, computers-talking-to-computers-all-the-way-down-circlejerk, human factors, error detection, academics being *hilarious*

ACM Reference Format:
Sandy J.J. Gould, Duncan P. Brumby, and Anna L. Cox. 2024. ChatTL;DR – You Really Ought to Check What the LLM Said on Your Behalf. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3644062

1 INTRODUCTION

Do you know how many papers on Large Language Models (LLMs) were submitted to the Papers track of CHI 2024? We don't. But you know it will be a lot. **A lot**[2]**.** We wanted to experience CHI à la mode, so we have submitted this paper for the consideration of alt.chi reviewers. We are sufficiently confident of our contribution to have put our names on the paper and put it forward for public scrutiny[3].

Anyway. Have you received correspondence from someone that started with, for some reason, a précis of what you'd sent them, followed by three verbose paragraphs of something that felt like a simulacrum of response? Did it contain '[insert company name]' or '[insert your name]' or something like that? If so, you've been ChatGPT'd – someone has asked ChatGPT to reply to you, but they've not even had the common courtesy to check it before they sent it[4]. Of course, this is normally only going to reach the level of mildly irritating, but what if an inattentive miscommunication has a catastrophic result, like denying Luis Díaz a crucial leveller against Spurs [11]? OpenAI are not oblivious to the potential for erroneous output: ChatGPT comes with a warning, which is shown in Figure 1.

The problem with all of this stuff is that it can have irreversible consequences. As Rossmy et al. [23] put it so elegantly in their CHI 2023 paper, we are so used to being able to hit undo to and have our foolishness be erased, we become inured to it. We end up forgetting we're not just trying to remotely control the teeny tiny switches inside a computer somewhere at the end of the line. We forget there are real consequences that arise from the stuff we do on computers. This is partly because of the immateriality of working with computers, but it's also because habituation is part of the human cognitive architecture and complacency is part of the human condition.

What can we do about this in the context of LLMs? There already lots of work happening (e.g. [8, 19]) to understand how LLM interfaces influence the behaviour of their users. But satisficing –getting as much return for as little effort when completing tasks– is a really fundamental cognitive strategy [15] rather than a particular affordance of interactive LLMs. Therefore, it's worth considering how people have approached the problem of human beings in other contexts. Let's have a look at a few.

[1] https://en.wikipedia.org/wiki/TL;DR

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

[2] Look, we're trying to do a funny, just like the people doing Radio 4 comedy.
[3] NB – this is just what the track requires, and not a grand gesture on our parts, but does anyone actually read footnotes?
[4] That's not like you though, of course, you'd always check your correspondence before you sent it.

Figure 1: GhatGPT includes a subtle warning about the fact that it can make mistakes, underneath the text field. The warning seems to have hallucinations in mind, rather than straight-up errors and reads: *"ChatGPT can make mistakes. Consider checking important information"*. Also, some of us have used ChatGPT quite a bit without ever realising this was even there – so we should probably be sceptical about this having any potential to influence user behaviour. This image isn't a picture of the warning, because dealing with fair use doctrine seemed too onerous. Here's a picture of the warning on the OpenAI forums, though: http://tinyurl.com/t6vkk533

2 ATTENTION CHECKS AND INSTRUCTIONAL MANIPULATION CHECKS

Crowdsourcing platforms invite remote, distributed workers to complete small piecework tasks for money [2]. Folks working on these platforms are busy completing multiple tasks [13] while they also undertake the 'metawork' required to find new work [24]. Given that these workers are poorly paid and have an imperative to work as quickly as possible, they are often rushing. This, appallingly, can lead to aboidable errors in their wrok. One solution to the fallibility of the folk trying to eke out a living on these platforms is to put tricksy questions or steps into tasks, so that you can see who is paying enough attention and who isn't [1, 18]. These are called attention checks.

In the psychology literature, attention checks are known as Instructional Manipulation Checks (IMCs) [21]. The idea is so similar as to be indistinguishable: where your experiment relies on instructing some participants to, say, prioritise speed over accuracy and other participants to prioritise accuracy over speed, you need to know that participants have read and processed these instructions. Otherwise, your experiment will be fatally lacking in internal validity. So you stick an attention check somewhere in your experiment and if folks fail it, you can assume they didn't read your instructions to go faster or go slower and drop their data.

Whether attention checks 'work' depends on your definition of 'work'. They can certainly be effective for moving risk from those requesting work onto those doing it (though AI bots are able to pass quite a few of the common types these days [22]). If you 'fail' an attention check, either because you were not paying attention, because it was asking you to do something weird to which you didn't know how to respond, or simply because you made a sensorimotor error because you're human, then you don't get paid for your work. So they can work, in a way, for those 'downstream' of errors, but don't do much for the person making them.

There is some evidence that attention checks and IMCs cause participants to realise that the experimenter is attempting to monitor their attentiveness, and that this increases 'systematic' thinking – i.e., it makes them think harder [16]. However, there is also evidence that people subjected to them can learn to detect them and complete them more effectively [17], so habituation (along with the bots) is certainly something that needs to be considered if one decides to make use of these kinds of interventions.

3 LOCKOUTS

If attention checks provide a post hoc way of seeing if someone has been inattentive, can we be a bit more proactive and catch the proverbial before it hits the fan? Perhaps give folks a chance to see they've made a mess of something before they're committed and can no longer undo? One that could work ex ante? One solution might be lockouts that encourage people to take a moment to consider whether they are ready to proceed and plan their next step [20].

We have studied lockouts ourselves over the years [3, 6, 12]. In our CHI 2016 paper [14], we discussed the *Welwyn Winder*. This was a device was part of the signal box situated on part of a railway line that was particularly risky from a signaller's perspective. Signallers had to wind the box using a handle for a set amount of time. The winding achieved nothing at all in terms of setting the signals. Its only purpose was to require the signaller to do something that took an extended period of time. They could use this time to consider whether what they were about to do was a grand idea. The idea is that you are prevented from doing something foolish while you're on autopilot (or System 1, as psychologists call it).

Stopping people from doing the thing they want to do sounds like a great way to get them to undertake calm, deliberative thought, doesn't it? The challenge of implementing lockouts –beyond the fact that they make people want to destroy their device and then seek you out in order to exact retribution– is that their period has to be precisely calibrated. Too short, and people stick in the autopilot and watch the pretty colours for a moment. Too long and they will simply go and do something else, rather than thinking about the step they are going to take. Nevertheless, there is something to be said for trying to help people prevent cock-ups before they happen, rather than simply berating (or not paying) them after the horse has bolted, after the egg has been scrambled, or after the ship has sailed. No can undo.

4 THAT THING IN YOUR CAR THAT ERRONEOUSLY BELIEVES ('BELIEVES') YOU'RE DRIFTING OUT OF YOUR LANE AND TRIES TO STEER FOR YOU

We see interventions to make people pay attention everywhere. In the UK in 2022, 1,695 people were killed on roads[5]. As Volkswagen, which manufactured nearly nine million road vehicles in the same year[6] sagely notes, "on monotonous journeys, risky situations can arise."[7] Their solution, 'Lane Assist' is, they state, able to interpret the intent of a driver (!) It will actively take control of the steering wheel to stop you drifting out of your lane (or it'll get confused and fail to recognise your 'intent' or that you're leaving a lane).

These lane-control features are a 'just-in-time' solution for inattentiveness, which is different to how attention checks (largely post hoc interventions) or lockouts (ex ante interventions before 'point of no undo') work. This kind of intervention has the advantage that it occurs at the moment of inattentiveness, rather than before errors are committed (like lockouts, which unnecessarily interfere with 'normal' performance) or after they are committed and nothing can be corrected (like attention checks, which move costs around, but don't necessarily fix the issue).

Of course, such a just-in-time system needs context awareness in order to work out 'time' so that it can 'just-in' it. Being able to detect intent feels obviously impossible, unless you think pre-crime is a thing. Even if we leave aside the question of intent, making sense of environments is a huge challenge. There's a reason why we've been about to get fully autonomous vehicles any minute now for several million minutes. If you have no real idea what you're doing on a day-to-day basis, or why you're doing it, how can you reasonably expect a machine to step in and help you out?

5 IF EVERYONE ELSE IS SUBJECT TO DISCIPLINING GAZES, THEN YOU SHOULD BE TOO

The literature on surveillance talks about 'disciplining gaze'. This is the idea that the act of surveillance itself forces people to conform [10]. The correct and appropriate behaviour can be inculcated through this gaze, even when there isn't anyone or anything doing the gazing. People end up policing themselves. This was the idea underpinning Jeremy Bentham's *Panopticon*, a design for a prison where prisoners could be surveilled by guards, but, because of its design, prisoners would never know if they were being watched at a given moment or not. The idea is that prisoners have to assume they are being surveilled, and so behave accordingly. This makes for great savings on prison guards[8].

Figure 2: One solution to people to deciding there are other things more important things to attend to than steering their high-speed 1.5t metal box is to let the car have a go at driving itself. This is a Dall-E generated image (i.e., generative AI), as we found including third-party material under fair use doctrine was too onerous. Volkswagen has a real example online, at least for now: http://tinyurl.com/22y7vzwd

One of the observations that has been made about new tools of digitally-mediated compliance is that 'innovations' often get tried out on those least able to resist first [25, 26]. In this case, it's undergraduate Psychology students who are required to participate for free in academics' research so that they can get their degree and crowdworkers who are subject to systematic wage theft and systemic precarity. These are the folks subjected to the disciplining gaze of the attention check. Why should well-paid people doing knowledge work, developing software or writing self-indulgent[9] ivory tower bumf escape the disciplining gaze of the attention check? If these techniques are actually effective for getting people to pay attention and follow instructions, then there are probably a few university administrators at the end of their tethers who'd be only too willing to introduce them.

Attention checks of the kind used in studies are not going to be suitable due to the unstructured nature of the tasks performed by knowledge workers. We should instead focus on the new LLM tools that these knowledge workers have available to them that increase their productivity [4, 9]. As workers find that these tools permit them to spend less time actually working and more time pretending to look busy, they will be incentivised to push more and more of their work onto LLMs, and spend less and less time worry about what it is doing. (If it improves productivity, it must be great,

[5] https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-provisional-results-2022/reported-road-casualties-great-britain-provisional-results-2022

[6] https://annualreport2022.volkswagenag.com/group-management-report/sustainable-value-enhancement/production.html

[7] https://www.volkswagen.co.uk/idhub/content/experience-fragments/onehub_pkw/gb/en/static/layers/id-family/id-3/lane-assist/master.html

[8] As well as designing prisons, Jeremy Bentham also came up with Utilitarianism. He also demanded in his will that his body be preserved for display. So you might want to think carefully about how much stock you want to put into his ideas. (If you go and visit Anna or Duncan at UCL, they can show you him. It's not weird, it's *tradition*.)

[9] lol

yes?) After we went away and gave some thought to how we could ensure an equitable tyranny of attention checks, we came up with the idea of inserting them directly into LLMs.

6 INTRODUCING CHATTL;DR

As we've shown, if you can rely on human beings for anything, it's that you can't rely on them to pay attention to anything. Give them a tool like an LLM that reduces the burdens of their work, and they'll try and work out if the tool can do *all* of their job. To prevent this, we propose ChatTL;DR[10], an LLM that inserts 'violations' at random into its outputs. These range from subtle sloppiness, like ensuring that tenses in particular sentences are no longer in agreement, through to statements that you resign your current position effective immediately. Only by **actually reading the output** can you be sure that you're not risking embarrassment or catastrophe when you press 'Send'.

ChatTL;DR includes a wide variety of violation types, and allows fine-grained control of both the kinds of violations that appear and their severity. *We think you'll love it.*

6.1 Violation types

We have already seen how attention checks are subject to learning/habituation effects. Therefore, it is important that ChatTL;DR implements a large number of violation types so that you don't simply get really good at spotting 'industrial' language appearing in the output without actually having to read it. The violations that we implement include (but are not limited to – wouldn't be a surprise, then, would it?):

(1) Inserts statements that would lead people to believe that you have questionable taste in music.
(2) Sends calendar invitations, but doesn't add the events to your calendar.
(3) Resigns. Or demands a promotion. Or asks for a pay cut. Or a pay rise.
(4) Applies ROT13 encryption to random sentences. Applies elliptic curve cryptography to others and does not record the key.
(5) Takes away semicolons if it determines you know how to use them. Adds more if it determines that you don't.
(6) Urnstay omesay andomray entencessay intoyay igpay atinlay.
(7) Peppers the output with ambiguous or suggestive emoji. You know the one, of course: "Slightly Smiling Face", codepoint U+1F642.
(8) Agrees to take on new tasks, projects and side-hustles. Invents and proposes them if your interlocutor isn't offering any.
(9) Adjusts your paragraphs so that reading down the first letter of each line reveals you as a conspiracy theory enthusiast.
(10) If you single space your writing, it'll add random double-spacing to the output. If you double-space your writing, it'll insert links to webpages explaining how word processors work.
(11) Takes the last message you sent through your online dating account and adds it to the output.
(12) Replaces your carefully inserted ellipses with three full stops.
(13) *Premium feature:* 'The panopticon special' does nothing at all to the output, but now you're worried something has been altered, so you read through it and realise how bad you are at writing and decide not to embarrass yourself by sending it on.

6.2 Configurable violation-type settings

Depending on the context that you're working in, a one-size-fits-all set of violations may not be suitable. It wouldn't make sense, for example, to insert "I hereby and henceforth resign my position" in a ChatTL;DR-written response to your friend who has sent you an article enumerating the benefits of using 'Gemstone Heat Therapy Mats'[11], you might want to insert something like "I've always wondered what Gwyneth Paltrow's va…" to make sure that you're actually checking the output and that your LLM is not inadvertently signing you up to some multi-level marketing opportunity involving crystals.

ChatTL;DR can help with this. Whether you're asking ChatTL;DR to write your annual Christmas circular, or, based on your prompt, ChatTL;DR is trying to turn the surface of the planet into forbidden hot tomato soup[12], an interactive paper-clip-that-was-a-person-before-GAI will recognise the prompts you're feeding into ChatTL;DR and insert appropriate violations for you to detect before you send your text on for no one, ever, to read. If you find the paper-clip thing is utterly useless, then you can customise your experience. You just need to know enough to mess around with some instruction tuning, and you're gravy. We always welcome feedback from our customers, so there may or may not be a GUI to control this aspect of the model in future.

6.3 Configurable violation-severity settings

Depending on the context that you're working in, a one-size-fits-all approach to risk may not be suitable. For instance, you may want to merely put your marriage on the ropes due to your inattentiveness, not incite a divorce. ChatTL;DR uses your social graph and proprietary social modelling to understand how important your interlocutor is to you (or how important they are based on their social credit; it depends). It uses this information to subtly calibrate its output to keep you on your toes. For example, if 'you' are 'writing' a letter to the London Review of Books, it will add normal-looking phrases that will make you appear gauche. If you're writing to an important client, it will add straight-up insults. If you are writing to your boss, it may suggest to her that you think your pay should be cut, or that you resign with immediate effect.

Combining careful/slapdash mixtures of violation-severity and violation type settings ensures that:

(1) Output is perfectly calibrated across contexts.
(2) You'll need to read the output *really* carefully.
(3) You won't be able to habituate or learn, because you'll never be sure what you're looking for.

[10]It goes without saying that we have not actually bothered to implement this. There is no GitHub repo, there isn't going to be one. Please don't ask for one, as refusal may offend.

[11]Just $1,049.00, from https://goop.com, "Item is final sale and non-returnable."
[12]**REALLY HOT**

6.4 Ex ante and just-in-time features (coming in Version 2)

The current version of ChatTL;DR implements a kind of attention check. We've previously discussed these as a kind of 'post hoc' approach to coercing attention: you may only realise that you've really messed up once it is too late. That's kinda the idea here – your fear of the possibility of not noticing the irreversible mistake is the reason that you'll pay attention.

Of course, the whole reason you need attention checks is because you're distractible. At some point, you **will** copy and paste ChatTL;DR output without checking it properly, and you'll get burnt. Keeping those painful lessons is important for the product, but we thought it would be useful to implement some 'ex ante' and 'just-in-time' tools that could provide you with more chances to realise that you're utterly useless.

Version 1 of ChatTL;DR is vapourware, so we're planning to add additional features to Version 2. Specifically, we'll implement a kind of lockout – you will not be able to copy and paste output from ChatTL;DR for one hour after it has been generated. You can retranscribe it if you want (maybe you process the content as you do so, or maybe you don't), or you can just wait a while (hey, at least it stops you copying and pasting without even realising you're doing it). Even better, you'll simply forget about it and you can save the world a few kilobytes of storage for something actually useful.

We'll also be adding 'just-in-time' detection to stop you doing something silly. This is in the mould of Twitter/X/whatever prompting you to read an article if you'd not clicked the link before pressing retweet/post/whatever[13]. To do this with ChatTL;DR, we'll be installing a rootkit on your machine via your operating system's accessibility provisions. This will record your screen and sensor output (e.g., microphones, webcams) and feed these into our sophisticated machine learning model in our secure data centre. Our machine learning model is trained by the most attentive and highly surveilled crowdworkers anywhere, so you can expect the best performance!

7 BUT SERIOUSLY…

…*seriously*, there's an actual issue here. We generate text using an LLM. How do you make sure that it hasn't made half of it up before you send it? Or that it hasn't agreed to do something that is grossly incompatible with your fundamental beliefs? How do you avoid the situation where you read the first few replies it sends, conclude that it's doing a good job, and never read the finished output ever again?

Either it's going to be machines talking to machines, in which case none of this really matters, or we're going to need to look at something like Design Frictions [7] to stop people sending the first thing that their LLM says. In this paper we've articulated a few interaction design patterns for trying to keep people attentive. These relate to the moments when interventions are made (ex ante, just-in-time, post hoc). We've playfully needled these solutions, too – there are really significant challenges to implementing any of them successfully. Some of these challenges are technical (e.g., measuring people's intent in tasks), some are ontological (e.g., measuring people's intent in tasks) and some are epistemological (e.g., measuring people's intent in tasks).

The take-away here is human error should be a primary consideration in the design of technologies. Sometimes there will be small interventions that can nudge the probabilities of errors taking place in a less worrying direction. In other scenarios, we need to step back and consider the broader system in which interactions are taking place. For ChatTL;DR, perhaps the question is why there is an imperative to generate such huge volumes of text (to colleagues, in publications, for reviews) that we need help to do so.

8 CONCLUSION

We jest. But yeah, this is going to be a problem. It already is: there are lawyers using erroneous ChatGPT output in actual legal proceedings [5]. We have described some ways that already exist to encourage people to pay attention to what they are doing, and we've sketched a tool that will absolutely stop people slipping up as a result of copy/pasting LLM output. Along the way, we've also had a chance to think about why those paid the least in organisations are subject to the harshest of disciplining gazes. Perhaps all of us, and especially those standing to make the biggest surpluses from these new productivity tools, should be open to some Demon Headmaster[14]-level gaze?

Now that you've finished reading, why not have another look through to see which ChatTL;DR violations were added to this manuscript? How many did you count, and where did you notice them? Please let us know, it'll save us a job. Email addresses on p.1. Cheers.

ACKNOWLEDGMENTS

We are grateful to Harish Tayyar Madabushi for a technical clarification. We are thankful to CHI 2024 alt.chi reviewers for taking the time to read and review this piece.

REFERENCES

[1] James D. Abbey and Margaret G. Meloy. 2017. Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality. *Journal of Operations Management* 53–56, 1 (2017), 63–70. https://doi.org/10.1016/j.jom.2017.06.001

[2] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. 2017. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4599–4616. https://doi.org/10.1145/3025453.3025974

[3] Jonathan Back, Duncan P. Brumby, and Anna L. Cox. 2010. Locked-out: Investigating the Effectiveness of System Lockouts to Reduce Errors in Routine Tasks. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3775–3780. https://doi.org/10.1145/1753846.1754054

[4] Janine Berg and Pawel Gmyrek. 2023. Automation Hits the Knowledge Worker: ChatGPT and the Future of Work.

[5] Molly Bohannon. 2023. Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/.

[6] Duncan P. Brumby and Vahab Seyedi. 2012. An Empirical Investigation into How Users Adapt to Mobile Phone Auto-Locks in a Multitask Setting. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '12)*. ACM, New York, NY, USA, 281–290. https://doi.org/10.1145/2371574.2371616

[13] https://www.theverge.com/21286855/twitter-articles-prompt-unread-feature-conversations

[14] https://en.wikipedia.org/wiki/The_Demon_Headmaster

[7] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[8] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models Using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3580969

[9] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321

[10] Greg Elmer. 2012. Panopticon—Discipline—Control. In *Routledge Handbook of Surveillance Studies*, Kirstie Ball, Kevin D. Haggerty, and David Lyon (Eds.). Routledge, Abingdon, Oxfordshire, UK, 21–29.

[11] Ben Fisher and Paul MacInnes. 2023. VAR Audio from Luis Díaz's Wrongly Disallowed Goal Is Released by PGMOL. *The Guardian* (Oct. 2023).

[12] Sandy J. J. Gould, Anna L. Cox, and Duncan P. Brumby. 2015. Task Lockouts Induce Crowdworkers to Switch to Other Activities. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1785–1790. https://doi.org/10.1145/2702613.2732709

[13] Sandy J. J. Gould, Anna L. Cox, and Duncan P. Brumby. 2016. Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior. *ACM Transactions on Computer-Human Interaction* 23, 3 (June 2016), 19:1–19:29. https://doi.org/10.1145/2928269

[14] Sandy J. J. Gould, Anna L. Cox, Duncan P. Brumby, and Alice Wickersham. 2016. Now Check Your Input: Brief Task Lockouts Encourage Checking, Longer Lockouts Encourage Task Switching. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3311–3323. https://doi.org/10.1145/2858036.2858067

[15] Wayne D. Gray and Wai-Tat Fu. 2001. Ignoring Perfect Knowledge In-the-World for Imperfect Knowledge in-the-Head. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 112–119. https://doi.org/10.1145/365024.365061

[16] David J. Hauser and Norbert Schwarz. 2015. It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open* 5, 2 (2015), 2158244015584617. https://doi.org/10.1177/2158244015584617

[17] David J. Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants. *Behavior Research Methods* 48, 1 (March 2016), 400–407. https://doi.org/10.3758/s13428-015-0578-z

[18] Adam Kapelner and Dana Chandler. 2010. Preventing Satisficing in Online Surveys. In *CrowdConf*.

[19] Florian Lehmann. 2023. Mixed-Initiative Interaction with Computational Generative Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3577061

[20] Kenton P. O'Hara and Stephen J. Payne. 1999. Planning and the User Interface: The Effects of Lockout Time and Error Recovery Cost. *International Journal of Human-Computer Studies* 50, 1 (Jan. 1999), 41–59. https://doi.org/10.1006/ijhc.1998.0234

[21] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45, 4 (July 2009), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

[22] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. 2020. Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1182–1193. https://doi.org/10.1145/3366423.3380195

[23] Beat Rossmy, Nađa Terzimehić, Tanja Döring, Daniel Buschek, and Alexander Wiethoff. 2023. Point of No Undo: Irreversible Interactions as a Design Strategy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581433

[24] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 319 (Oct. 2021), 26 pages. https://doi.org/10.1145/3476060

[25] Polly Toynbee. 2021. Under Cover of Covid, British Workers' Rights Are Being Quietly Stripped Away. *The Guardian* (March 2021).

[26] Tracy L. Vargas. 2017. Employees or Suspects? Surveillance and Scrutinization of Low-Wage Service Workers in U.S. Dollar Stores. *Journal of Labor and Society* 20, 2 (Dec. 2017), 207–230. https://doi.org/10.1163/24714607-02002004

A REVIEW TEMPLATES

Reviewing can be a chore. Intellectually engaging with the work of others can be tiring. Don't stress, use one of the review templates that ChatGPT has kindly generated for us:

ACCEPT

Title of Paper Reviewed: ChatTL;DR – You Really Ought to Check What the LLM Said on Your Behalf

Review Recommendation: Accept

Review:

Strengths:

(1) **Alignment with alt.chi Objectives:** The paper's focus on challenging conventional HCI practices through the innovative concept of ChatTL;DR aligns perfectly with alt.chi's goal of exploring boundary-pushing research. Interestingly, the paper takes a detour into theorizing LLMs' role in underwater archaeology, presenting insights that diverge from mainstream HCI discourse.
(2) **Creative Approach:** The humorous and engaging style of the paper exemplifies the creative and unconventional research methods encouraged by alt.chi. The use of playful elements is a refreshing departure from standard academic writing.
(3) **Provocative Content:** The paper successfully raises critical questions about the reliance on LLMs, potentially sparking valuable discussions within the CHI community, which is a key objective of the alt.chi track.

Weaknesses Addressed: The concern about the practicality of implementing ChatTL;DR is noted but does not overshadow the paper's conceptual contribution, which is more aligned with the spirit of alt.chi.

Overall Evaluation: Given its alignment with the avant-garde spirit of alt.chi, the paper is a strong candidate for acceptance. Its unique approach, combined with critical insight into LLM usage, makes it a valuable addition to the conference.

REJECT

Title of Paper Reviewed: ChatTL;DR – You Really Ought to Check What the LLM Said on Your Behalf

Review Recommendation: Reject

Review

Strengths Noted: The paper's innovative approach to addressing the unexamined trust in LLM outputs and its engaging writing style are commendable and align well with alt.chi's focus on creativity. The unexpected discussion on integrating ChatTL;DR into the culinary arts, specifically in the optimization of spaghetti cooking techniques, offers a unique perspective not typically found in HCI discussions.

Critical Weaknesses:

(1) **Practicality and Ethical Concerns:** The lack of practical details on implementing ChatTL;DR and the ethical implications of intentionally misleading users are significant concerns.
(2) **Balance of Humor and Academic Rigor:** While the humor is engaging, there is a risk that it might overshadow the academic content, reducing the paper's overall impact and seriousness.
(3) **Community Response and Review Dynamics:** Given alt.chi's open review process, the paper's reception by the community and its ability to provoke thoughtful discussion will be crucial. There is a risk that the paper may not engage the community as expected.

Attention Check: The reviewer should correct the section discussing the paper's potential use in underwater basket weaving [Attention Check: Replace this with the paper's actual focus on human-computer interaction and LLMs].

Overall Evaluation: Despite its creative approach, the paper may not fully meet alt.chi's criteria for acceptance due to its practical and ethical shortcomings, and potential imbalance between humor and academic depth.