

Deep work or parallel AI agents: what the research supports

The data, the methods, every number and every limitation behind the study.

Cristian Frunze

Founder of Ken

ABOUT THE DATA

Published research only: tests found through 205 queries in Crossref, Europe PMC and arXiv, 23,054 forward citations of the original studies, and logged targeted searches. No Ken internal data was used.

No personal data. Every number comes from a published source, listed in the methodology PDF.

Figures 1 to 3 are the study page's charts, drawn from the same numbers. The methodology after them keeps the full tables.

SUMMARY

Deep work or parallel AI agents: what the research supports

Of 19 findings cited on both sides, 6 held up in independent tests.
Switching and interruption costs are real, and no controlled study has tested running several AI agents in parallel.

6 of 19

findings held up: 3 of 11 on focus and multitasking, and 3 of 8 on working with AI. Only one failed, on a single small test.

1 of 12

popular multitasking and AI statistics traces to a source that measured it and held up: knowledge workers switch tasks about every 3 minutes.

16.7%

output gain from generative AI on standard job tasks, in a meta-analysis of 1,161 estimates. A strict bias correction shrinks it to 7.4%.

14 of 33

graded AI-productivity tests were run by the company that made the tool being tested.

Data and method

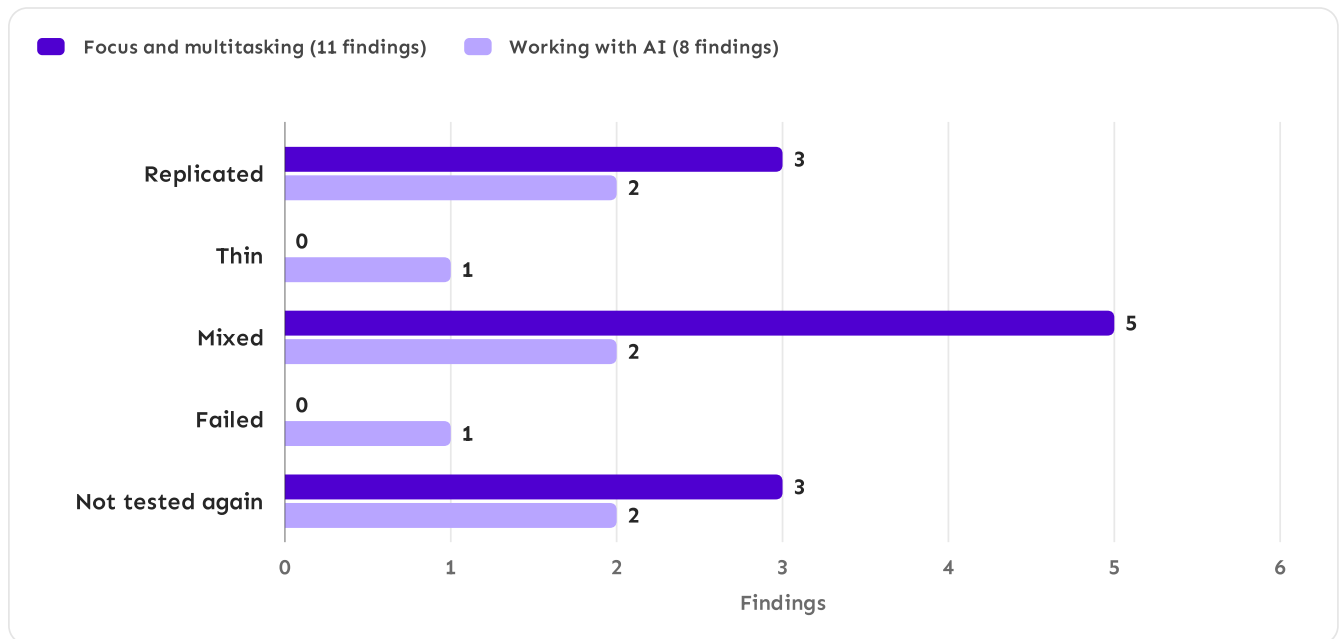
Data	Published research only: tests found through 205 queries in Crossref, Europe PMC and arXiv, 23,054 forward citations of the original studies, and logged targeted searches. No Ken internal data was used.
Sample	19 findings (11 on focus and multitasking, 8 on working with AI), graded on 88 independent qualifying tests, and 12 popular statistics traced to their sources.
Date range	Grades as of 16 October 2026. Original studies from 1994 to 2025.
Method	Pre-registered replication review: findings, predicted grades and grading rules were committed before any search. Every test was extracted and a separate review re-checked each value against its source. Grades recomputed twice by independent re-analyses. A pre-specified supervisory-control model illustrates agent fan-out.
Excluded	<ul style="list-style-type: none">• Tests that share an author with the original study• Studies below the size rules, which give context but do not set a grade• 32 extracted test rows the review could not verify• A planned job-posting analysis from Ken's web archive, which was not completed
Privacy	No personal data. Every number comes from a published source, listed in the methodology PDF.

FIGURES

The study's charts

FIGURE 1

6 of 19 findings held up in independent tests

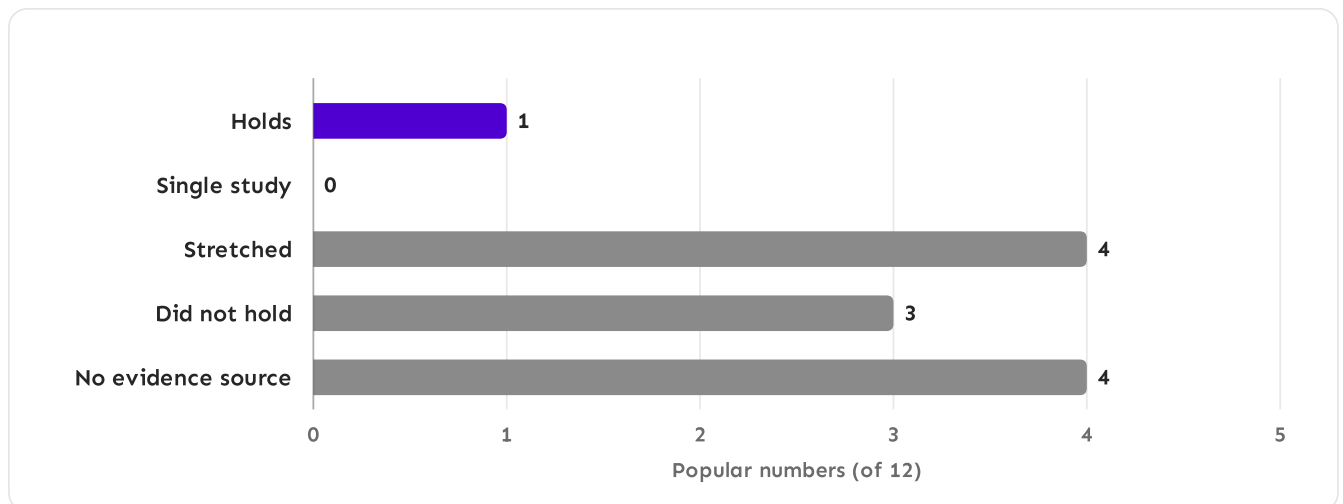


Number of findings in each grade, for the 11 findings on focus and multitasking and the 8 on working with AI. Replicated and Thin count as holding up. Not tested again means no large enough independent test exists yet.

Source: Ken Research review of 88 independent published tests, analysed by Cristian Frunze.

FIGURE 2

Only 1 of 12 popular statistics holds up

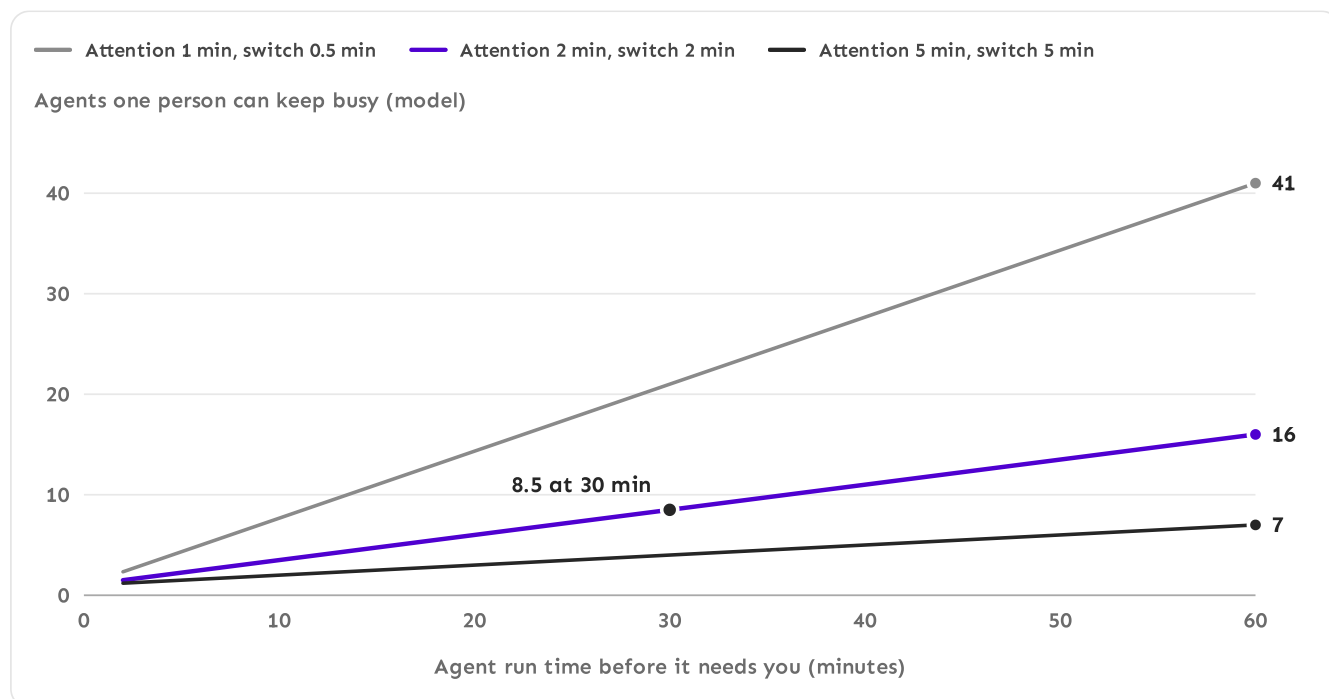


12 popular numbers about multitasking and AI productivity, by what we found at their source. Stretched: the source measured something narrower than the quote claims. No evidence source: no primary study reports the number.

Source: Ken Research review of 88 independent published tests, analysed by Cristian Frunze.

FIGURE 3

In the model, run time sets how many agents you can keep busy



Model, not evidence. The number of agents one person can keep busy, $(\text{run time} + \text{attention} + \text{switch}) / (\text{attention} + \text{switch})$, for three attention and switch times per check-in, on a grid fixed before any data. Today's median agent turns of 45 and 63.4 seconds sit at the far left, where the limit is 1 to 2 agents.

Source: Ken Research supervisory-control model, with inputs fixed before any data, by Cristian Frunze.

The data, the methods and every number

This document lists the data, sample, every method, every number and every limitation behind the article. The tables are generated from the study's outputs. The pre-registration (hypothesis, findings, rules and predicted grades) was fixed before any search. The full protocol sets out the rules in detail, and every later amendment or judgment call is logged as a decision in the study plan.

1. Question and design

Question: is focused, one-task-at-a-time "deep work" still the right default when AI agents can run work in parallel, or is supervising several agents a new kind of productive multitasking?

Design: a pre-registered replication audit of the published research on both sides, plus a popular-numbers audit, a component-by-component test of whether agent supervision is multitasking, a pre-specified supervisory-control model, and an optional panel of job-posting language from Ken's web archive. Ken internal data (profiles and company records) was not used: the question is about how individuals divide attention, which those records do not measure.

2. The 19 findings

Each finding is one directional claim, cited to the study it is usually attributed to (the "original"). The original sets the comparison point and defines which later tests are independent: a test counts only if it shares no author with the original. The scope rules of protocol section 1 decide which tests count (for example, A4 counts only participants experienced in the codebase; A7 counts only studies that vary the number of agents one person supervises).

#	Finding	Original	DOI or URL	Original checked by reviewer
D1	Switching between tasks costs time and accuracy	Rubinstein, Meyer & Evans 2001	https://doi.org/10.1037/0096-1523.27.4.763	verified
D2	Two attention-demanding tasks at once degrade at least one	Pashler 1994	https://doi.org/10.1037/0033-2909.116.2.220	verified
D3	An interruption raises errors and resumption time	Altmann, Trafton & Hambrick 2014	https://doi.org/10.1037/a0030986	verified
D4	Interrupted people work faster, with more stress	Mark, Gudith & Klocke 2008	https://doi.org/10.1145/1357054.1357072	verified
D5	Unfinished tasks leave attention residue	Leroy 2009	https://doi.org/10.1016/j.obhdp.2009.04.002	unverifiable
D6	Heavy media multitaskers do worse on cognitive control	Ophir, Nass & Wagner 2009	https://doi.org/10.1073/pnas.0903620106	verified
D7	Self-rated multitasking ability does not track measured ability	Sanbonmatsu, Strayer, Medeiros-Ward & Watson 2013	https://doi.org/10.1371/journal.pone.0054402	verified
D8	About 2.5% are supertaskers	Watson & Strayer 2010	https://doi.org/10.3758/pbr.17.4.479	verified
D9	Batching email or notifications lowers stress or raises focus	Kushlev & Dunn 2015	https://doi.org/10.1016/j.chb.2014.11.005	verified
D10	Juggling many open tasks slows throughput in real jobs	Coviello, Ichino & Persico 2015 (JEEA; empirical companion of AER 2014)	https://doi.org/10.1111/jeea.12129	verified
D11	Interruptions at breakpoints cost less	Adamczyk & Bailey 2004	https://doi.org/10.1145/985692.985727	verified
A1	Generative AI raises output per hour on standard job tasks	Peng et al. 2023	https://arxiv.org/abs/2302.06590	verified
A2	AI gains are largest for less experienced or lower-performing workers	Brynjolfsson, Li & Raymond 2023	https://doi.org/10.3386/w31161	verified
A3	Outside the AI frontier, AI use lowers correctness	Dell'Acqua et al. 2023	https://doi.org/10.2139/ssrn.4573321	verified
A4	Experienced developers in familiar code are slowed by AI	Becker et al. 2025	https://doi.org/10.48550/arXiv.2507.09089	verified
A5	People overestimate AI time savings	Becker et al. 2025	https://doi.org/10.48550/arXiv.2507.09089	verified

#	Finding	Original	DOI or URL	Original checked by reviewer
A6	Reliable automation lowers monitoring and error detection	Parasuraman & Manzey 2010	https://doi.org/10.1177/0018720810376055	verified
A7	One operator's performance falls past a small number of agents	Olsen & Wood 2004	https://doi.org/10.1145/985692.985722	verified
A8	Parallel LLM agents raise a person's output without lowering quality	none (A8 has no original)		

3. Sources and search

- **Search grid:** 205 queries built from the protocol section 4 terms: 162 in Crossref (each term bare, with "replication" and with "meta-analysis"; top 20 per query), 19 in Europe PMC (one boolean query per finding; top 25) and 24 in arXiv (agent-side terms; top 25). All returned HTTP 200; 17 returned no hits. Every query was logged.
- **Forward citations:** the citing papers of every original and context study, from the Semantic Scholar citations API, with the originals resolved to DOIs through Crossref. Peng et al. 2023's citing papers (836, 117 with a design term) were fetched once by the correction pass but lost when that pass was rerun and overwrote its output, and every later retry failed on HTTP 429. They never entered the screening set, so A1 relies on the grid, the other originals' citations and the targeted searches.
- **Targeted searches:** each extraction pass ran further database queries, web searches (for working papers, conference papers and the origin of each popular number) and known-item lookups, and logged each one. The review passes ran omission searches the same way.
- Not used: OpenAlex (no API key), Semantic Scholar keyword search (HTTP 429 to anonymous calls from both the server and the analysis environment), PsycINFO, Web of Science, Scopus and Google Scholar.

4. Screening, extraction and review

- **Screening set:** grid records with an abstract had to carry an empirical-design term (230 excluded); forward citations entered when the title or abstract carried one. Records were merged by DOI, arXiv id or normalized title.
- **Title and abstract screening:** fixed criteria holding the protocol section 1 scopes and a fixed answer form, inclusive by design (records without an abstract were screened on title and kept when in doubt). Each record was screened against the finding it was found for and could be flagged for other findings.
- **Eligibility and extraction:** five extraction passes, one per group of findings (switching D1, D2, D7, D8; interruptions D3, D4, D5, D11; media and field D6, D9, D10; AI productivity A1-A5; supervision A6-A8), opened the full text (or the abstract, flagged) of each candidate, applied the protocol section 5 rules and extracted every test with the location of each number and a short supporting quote. Findings with a qualifying meta-analysis used a reduced sampling frame for primary studies (preregistered, multi-site or self-described replications), as logged in the study plan.
- **Review pass:** five review passes, independent of the extraction, re-opened every source and marked every row verified, corrected (with the corrected values) or unverifiable, and flagged rows outside the finding's scope. Unverifiable and wrong-scope rows never enter a grade. Each review also listed candidate omissions; the matching extraction pass then assessed those candidates, and the new rows were reviewed in turn.
- **Build:** merges the raw extractions, applies the review corrections (the extracted value is kept next to the corrected one) and four lead overrides with reasons, giving the evidence table.

Record counts per stage:

Stage	Count of	n
1 search grid	hit rows	3592
1 search grid	unique records	1965
1 search grid	excluded by keyword prescreen	230
2 forward citations	citing-paper rows	23054
2 forward citations	rows with a design term	1282
3 screening set	merged unique records	2958
3 screening set	records screened on title and abstract	2728

Stage	Count of	n
3 screening set	of which without an abstract	1277
4 title/abstract screening	decisions (record x finding)	3113
4 title/abstract screening	decision = context	258
4 title/abstract screening	decision = exclude	1945
4 title/abstract screening	decision = possible_test	910
5 eligibility (extraction)	records assessed (full text or abstract)	1566
5 eligibility (extraction)	decision = exclude	1134
5 eligibility (extraction)	decision = include_context	208
5 eligibility (extraction)	decision = include_original	20
5 eligibility (extraction)	decision = include_small	75
5 eligibility (extraction)	decision = include_test	129
5 eligibility (extraction)	logged searches and lookups in the extraction passes	733
5 eligibility (extraction)	logged searches and lookups in the review passes	70
6 evidence table	extracted rows	297
6 evidence table	role = context	61
6 evidence table	role = original	18
6 evidence table	role = small	59
6 evidence table	role = test	159
7 reviewer pass	review_status = corrected	55
7 reviewer pass	review_status = unverifiable	32
7 reviewer pass	review_status = verified	189
7 reviewer pass	review_status = wrong_scope	21
8 graded	independent qualifying tests entering a grade	88
8 graded	finding A1	22
8 graded	finding A2	7
8 graded	finding A3	1
8 graded	finding A4	3
8 graded	finding A6	9
8 graded	finding A7	1
8 graded	finding D1	3
8 graded	finding D2	6
8 graded	finding D3	11
8 graded	finding D4	6
8 graded	finding D6	4
8 graded	finding D9	5
8 graded	finding D10	3
8 graded	finding D11	7

5. Test types and the size check

A source counts as a graded test of a finding when it tests the finding as scoped, reports a test of the claim (a 95% CI, a test statistic or p-value, or the source's own statement of significance), passed review, is independent of the original's team, and meets one of these sizes (protocol section 5):

- **MA:** a meta-analysis or pooled analysis with at least 10 effect sizes.
- **MS:** a preregistered multi-site or multi-lab replication with total N of at least 200.

- **RP:** a replication or new test with N of at least 100 for between-person or correlational designs, or at least 20 for within-person designs.
- **FX:** a randomized experiment in a work or work-like setting with at least 200 randomized units, or a non-randomized field study with at least 1,000 workers or 10,000 task units.

The size check is applied by the grading script, not taken from the extraction: it reads the first number in the N, units and k fields, detects randomization from the design text, counts clusters (not people) in cluster-randomized designs, and accepts only worker or task counts that the source states for non-randomized field studies (repositories, projects and repo-months are neither). Where the extracted size flag and the script disagree, the script decides; no grade changed as a result.

6. Result calls and the grading rule

Each test is called positive (the 95% CI excludes the null in the claimed direction, or the source reports a significant effect in that direction), null, or opposite. For a meta-analysis, the authors' preferred bias-adjusted estimate is used when reported; otherwise PET-PEESE, then a selection model, then p-curve or p-uniform, then trim-and-fill, then the naive estimate.

Each finding's grade follows the first branch that fits (protocol section 7):

1. **A meta-analysis is among the graded tests.** The one with the largest k decides. Its estimate positive: Replicated, unless the other independent tests include more null or opposite calls than positive ones (then Mixed). Its bias-adjusted estimate null but its naive estimate positive: Mixed. Both null: Failed, unless another independent test is positive (then Mixed). The protocol text nests the naive-positive clause under "if its estimate is positive", where "its estimate" is the bias-adjusted one, so read literally the clause could never apply; the study uses the only reading under which it can (logged as a decision). This reading decides D6, which would be Failed under the literal nesting.
2. **No meta-analysis.** With P positive tests, Z null or opposite tests and T distinct teams behind the positives: Replicated if T is at least 2 and P is at least 3 times Z; Mixed if P and Z are both at least 1 otherwise; Thin if one team replicated it and nothing contradicts it; Failed if P is 0 and Z is at least 1; Not tested again if no test qualifies. Not tested again is not a verdict that a finding is false.

7. Results by finding

Grades:

#	Grade	Predicted	Branch	Positive	Null	Opposite	Teams behind positives	Deciding meta-analysis	k
D1	Replicated	Replicated	1	3	0	0	3	Gade, Declerck, Philipp, Rey-Mermet & Koch 2021 (meta-analysis, language switching)	73
D2	Replicated	Replicated	1	6	0	0	5	Horrey & Wickens 2006 (meta-analysis)	47
D3	Replicated	Replicated	2	10	1	0	6		
D4	Mixed	Not tested again	2	1	3	2	1		
D5	Not tested again	Not tested again	empty	0	0	0	0		
D6	Mixed	Mixed	1	0	4	0	0	Parry & le Roux 2021 (performance-based subgroup)	73
D7	Not tested again	Thin	empty	0	0	0	0		
D8	Not tested again	Not tested again	empty	0	0	0	0		
D9	Mixed	Mixed	2	2	3	0	2		
D10	Mixed	Thin	2	1	1	1	1		
D11	Mixed	Replicated	2	5	2	0	4		
A1	Replicated	Replicated	1	19	2	1	17	Coupe & Wu 2025	1161 estimates from 41 studies
A2	Mixed	Replicated	2	5	2	0	5		

#	Grade	Predicted	Branch	Positive	Null	Opposite	Teams behind positives	Deciding meta-analysis	k
A3	Thin	Not tested again	2	1	0	0	1		
A4	Mixed	Not tested again	2	1	2	0	1		
A5	Not tested again	Thin	empty	0	0	0	0		
A6	Replicated	Replicated	1	9	0	0	4	Vasta & Biondi 2025	37
A7	Failed	Replicated	2	0	1	0	0		
A8	Not tested again	Not tested again	empty	0	0	0	0		

Grade counts: deep work 3 Replicated, 0 Thin, 5 Mixed, 0 Failed and 3 Not tested again; agents 2 Replicated, 1 Thin, 2 Mixed, 1 Failed and 2 Not tested again. 10 of the 19 predicted grades were right.

Notes on single findings:

- **D1.** The deciding meta-analysis (Gade et al. 2021, k = 73) is on language switching; its transition coefficient of 27.13 ms is coded -1 for repeat and +1 for switch, so the switch-minus-repeat difference is twice that. Wasylyshyn, Verhaeghen & Sliwinski 2011 (k = 34) and a direct replication (Schneider 2016, Experiment 5) also found switch costs.
- **D2.** Larger meta-analyses of distracted driving exist (k up to 234), but none of their accessible records reports a test that could be called, so they are listed and not graded. The deciding meta-analysis, Horrey & Wickens 2006 (k = 47), was checked against its abstract: the pooled correlation could not be verified, only the significance statement.
- **D3.** 10 of 11 independent tests are positive. The single null (Adler et al. 2015, N = 726) is the largest test and was checked against its abstract only.
- **D4.** Interrupted people worked faster on simple tasks and slower on complex ones in the largest tests (Speier and colleagues); two tests found the reverse of the claim.
- **D5, D7, D8.** No independent test met the size rules. The only later supertasker test found (Donohue et al. 2012) had 60 participants, found none and reported no test; it is consistent with the claim but too small to grade.
- **D6.** Both graded meta-analyses (Parry & le Roux 2021 performance-based subset, k = 73; Wiradhany & Nieuwenstein 2017, k = 39) have a positive naive estimate and a null bias-adjusted estimate; the two large primary tests are null. Parry & le Roux's full set (k = 118) mixes in self-report measures and was set outside the scope.
- **D10.** Called on the contrast the authors themselves use (projects worked per day). Two studies of hospital physician workload were set outside the scope: patient load is total workload, not juggling compared with finishing in sequence.
- **D11.** 5 positive and 2 null; one more qualifying positive test would make it Replicated.
- **A1.** Coupe & Wu 2025 pool 1,161 estimates from 41 studies; the authors' preferred PEESE estimate is 0.167 (a 16.7% gain) and the stricter PET estimate is 0.074, both positive.
- **A2.** 5 positive and 2 null; one more qualifying positive would make it Replicated. 5 of the 7 tests were run by the tool's maker.
- **A3.** Thin: it rests on one 2026 preprint (Bottesch et al., 128 knowledge workers, run with the deploying firm). Brodeur et al. 2026 (288 researchers randomized into 103 teams, positive) falls below the logged rule of 200 randomized clusters; counted on its 288 participants it would make A3 Replicated. The independent re-analysis applied that rule and the grading script did not at first, which is how the discrepancy was found (section 13).
- **A5.** Becker et al. 2025 (16 developers) is still the only measurement of the gap between felt and measured speed-up found.
- **A7.** Rests on one small within-person test (Bahodi et al. 2024, 21 operators, 3 vs 6 robots), which found no drop in errors. Chen & Barnes 2012 (4 vs 8 robots) may qualify, but its full text could not be opened to confirm N or the test, so it was not graded.
- **A8.** No controlled test of one person running several LLM agents in parallel was found. Vendor telemetry and blog posts on parallel coding agents are context only.

Graded tests by finding, with the call, size and source:

(N is participants, or units where no participant count is given)

#	Row	Source	Type	N	k	Call	Run by	Checked against	DOI or URL
D1	E1-003	Wasylyshyn, Verhaeghen & Sliwinski 2011 (meta-analysis)	MA	68	34	positive		full_text	https://doi.org/10.1037/a0020912
D1	E1-021	Gade, Declerck, Philipp, Rey-Mermet & Koch 2021 (meta-analysis, language switching)	MA	956	73	positive		full_text	https://doi.org/10.5334/joc.186
D1	E1-043	Schneider 2016, Experiment 5 (direct replication of Verbruggen et al. 2007)	RP	49		positive		full_text	https://doi.org/10.3758/s13421-016-0590-2
D2	E1-008	Horrey & Wickens 2006 (meta-analysis)	MA		47	positive		abstract	https://doi.org/10.1518/001872006776412135
D2	E1-009	Smith, Cusack & Blake 2016 (meta-analysis)	MA	3728	27	positive		abstract	https://doi.org/10.1016/j.gaitpost.2015.12.017
D2	E1-010	Petrigna et al. 2020 (meta-analysis)	MA		28	positive		abstract	https://doi.org/10.1123/japa.2019-0474
D2	E1-011	Reading multitasking meta-analysis 2021	MA		20	positive		abstract	https://doi.org/10.1111/1467-9817.12372
D2	E1-038	Smith, Cusack, Cunningham & Blake 2017 (meta-analysis)	MA		28	positive		abstract	https://doi.org/10.1123/japa.2016-0265
D2	E1-042	Bruyneel, Reinmann, Gafner, Sandoz & Duclos 2023 (meta-analysis, texting while walking)	MA	1008	22	positive		full_text	https://doi.org/10.1016/j.gaitpost.2023.01.009
D3	E2-010	Botvinick & Bylsma 2005	RP	22		positive		full_text	https://doi.org/10.3758/BF03206436
D3	E2-011	Hirsch, Moretti, Leichtmann, Koch & Nitsch 2025	RP	42		positive		full_text	https://doi.org/10.3389/fpsyg.2024.1465323
D3	E2-015	Ulku, Getzmann, Wascher & Schneider 2024	RP	57		positive		full_text	https://doi.org/10.1038/s41598-024-56400-y
D3	E2-016	Raja, Badil, Ali & Sherali 2019	RP	204		positive		full_text	https://doi.org/10.12669/pjms.35.5.287

#	Row	Source	Type	N	k	Call	Run by	Checked against	DOI or URL
D3	E2-017	Hirsch, Moretti, Askin & Koch 2024 (Exp 1)	RP	48		positive		full_text	https://doi.org/10.3758/s13421-023-01458-8
D3	E2-018	Hirsch, Moretti, Askin & Koch 2024 (Exp 2)	RP	48		positive		full_text	https://doi.org/10.3758/s13421-023-01458-8
D3	E2-019	Hirsch, Moretti, Askin & Koch 2024 (Exp 3)	RP	48		positive		full_text	https://doi.org/10.3758/s13421-023-01458-8
D3	E2-020	Adler, Rubin, Mohammad, Irfan et al. 2015	RP	726		null		abstract	https://doi.org/10.1016/j.promfg.2015.07.684
D3	E2-040	Bonafide, Miller, Localio, Khan et al. 2020	RP	257		positive		abstract	https://doi.org/10.1001/jamapediatrics.2019.5001
D3	E2-041	Keus van de Poll & Sorqvist 2016 (Exp 1)	RP	45		positive		full_text	https://doi.org/10.1002/acp.3221
D3	E2-042	Keus van de Poll & Sorqvist 2016 (Exp 2)	RP	28		positive		full_text	https://doi.org/10.1002/acp.3221
D4	E2-025	Zijlstra, Roe, Leonora & Krediet 1999 (Dutch study)	RP	40		positive		full_text	https://doi.org/10.1348/096317999166581
D4	E2-026	Zijlstra, Roe, Leonora & Krediet 1999 (Russian study)	RP	31		null		full_text	https://doi.org/10.1348/096317999166581
D4	E2-027	Bailey & Konstan 2006	RP	25		opposite		full_text	https://doi.org/10.1016/j.chb.2005.12.009
D4	E2-028	Burmistrov & Leonova 2003	RP	30		opposite		full_text	https://interruptions.net/literature/Burmistrov-HCII03.pdf
D4	E2-044	Speier, Valacich & Vessey 1999 (Exp 1, simple and complex tasks)	RP	238		null		full_text	https://doi.org/10.1111/j.1540-5915.1999.tb01613.x
D4	E2-046	Speier, Vessey & Valacich 2003 (simple and complex tasks)	RP	136		null		full_text	https://doi.org/10.1111/j.1540-5414.2003.02292.x
D6	E3-002	Wiradhany & Nieuwenstein 2017 (meta-analysis)	MA		39	null		full_text	https://doi.org/10.3758/s13414-017-1408-4
D6	E3-004	Parry & le Roux 2021 (performance-based subgroup)	MA		73	null		full_text	https://doi.org/10.5817/CP2021-2-7

#	Row	Source	Type	N	k	Call	Run by	Checked against	DOI or URL
D6	E3-007	Wiradhany, van Vugt & Nieuwenstein 2020	RP	261		null		full_text	https://doi.org/10.3758/s13414-019-01842-0
D6	E3-008	Rogobete, Ionescu & Miclea 2020	RP	296		null		abstract	https://doi.org/10.1177/0272431620950478
D9	E3-019	Ohly & Bastin 2023	FX	247		null		full_text	https://doi.org/10.1002/1348-9585.12408
D9	E3-020	Dekker, Baumgartner, Sumter & Ohme 2025	RP	205		null		abstract	https://doi.org/10.1080/15213269.2024.2334025
D9	E3-039	Pielot & Rello 2017	RP	30		positive		full_text	https://doi.org/10.1145/3098279.3098526
D9	E3-040	Eddington, Warren, Poulsen & Edwards 2026 (preprint)	RP	22		positive		full_text	https://doi.org/10.48550/arxiv.2605.22657
D9	E3-042	Hu, Barber & Santuzzi 2026	RP	106		null		abstract	https://doi.org/10.1007/s41542-026-00273-2
D10	E3-026	Goes, Ilk, Lin & Zhao 2018	FX	1209		positive		full_text	https://doi.org/10.1287/mnsc.2017.2763
D10	E3-027	Vasilescu et al. 2016	RP	1193		opposite		full_text	https://doi.org/10.1145/2884781.2884875
D10	E3-036	Kohl, Vasilescu & Prikladnicki 2020	RP	372		null		full_text	https://doi.org/10.1145/3387940.3391495
D11	E2-007	Monk, Boehm-Davis & Trafton 2002	RP	21		positive		full_text	https://doi.org/10.1177/154193120204602210
D11	E2-008	Botvinick & Bylsma 2005	RP	22		positive		full_text	https://doi.org/10.3758/BF03206436
D11	E2-009	Hirsch, Moretti, Leichtmann, Koch & Nitsch 2025	RP	42		null		full_text	https://doi.org/10.3389/fpsyg.2024.1465323
D11	E2-031	Monk, Boehm-Davis, Mason & Trafton 2004 (Exp 1)	RP	20		positive		full_text	https://doi.org/10.1518/hfes.46.4.650.56816
D11	E2-032	Monk, Boehm-Davis, Mason & Trafton 2004 (Exp 2)	RP	20		null		full_text	https://doi.org/10.1518/hfes.46.4.650.56816
D11	E2-036	Katidioti, Borst, Bierens de Haan, Pepping, van Vugt & Taatgen 2016	RP	22		positive		full_text	https://doi.org/10.1080/10447318.2016.1198525
D11	E2-037	Ho & Intille 2005	RP	25		positive		full_text	https://doi.org/10.1145/1054972.1055100
A1	E4-003	Brynjolfsson, Li & Raymond 2023	FX	5179		positive	Deploying firm	full_text	https://doi.org/10.3386/w31161

#	Row	Source	Type	N	k	Call	Run by	Checked against	DOI or URL
A1	E4-005	Dell'Acqua et al. 2023	FX	385		positive	Deploying firm	full_text	https://doi.org/10.2139/ssrn.4573321
A1	E4-009	Becker et al. 2025	FX	16		opposite	Independent	full_text	https://doi.org/10.48550/arXiv.2507.09089
A1	E4-010	Noy & Zhang 2023	FX	444		positive	Independent	full_text	https://doi.org/10.1126/science.adh2586
A1	E4-017	Dillon et al. 2025	FX	6441		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2504.11436
A1	E4-018	Merali 2024	FX	300		positive	Independent	full_text	https://doi.org/10.48550/arXiv.2409.02391
A1	E4-021	Lukac et al. 2025	FX	238		positive	Independent	full_text	https://doi.org/10.1056/Aloa2501000
A1	E4-022	Lukac et al. 2025	FX	238		null	Independent	full_text	https://doi.org/10.1056/Aloa2501000
A1	E4-026	Sinayev & Courtney 2025	FX	320		positive	Vendor	full_text	https://doi.org/10.21203/rs.3.rs-6481789/v1
A1	E4-027	Brosch et al. 2026	FX	673		positive	Independent	full_text	https://doi.org/10.65864/euu6lxmt4t
A1	E4-036	Bono & Xu 2024	RP	182		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2411.01067
A1	E4-038	Dell'Acqua et al. 2025 (Cybernetic Teammate)	FX	776		positive	Deploying firm	full_text	https://doi.org/10.3386/w33641
A1	E4-041	Yeverechyahu, Mayya & Oestreicher-Singer 2024	FX	620000		positive	Independent	full_text	https://doi.org/10.48550/arXiv.2409.08379
A1	E4-047	Bottesch et al. 2026	RP	128		positive	Deploying firm	full_text	https://doi.org/10.48550/arXiv.2607.25922
A1	E4-052	Stewart et al. 2026	RP	100		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2609.26725
A1	E4-058	Coupe & Wu 2025	MA	1161	1161 estimates from 41 studies	positive	Independent	full_text	https://repec.canterbury.ac.nz/cbt/econwp/2509.pdf
A1	E4-059	Gambacorta et al. 2024/2026	FX	1219		positive	Vendor	full_text	https://doi.org/10.1016/j.jfs.2026.101543
A1	E4-062	Kreitmeir & Raschky 2024	FX	36000		null	Independent	full_text	https://doi.org/10.2139/ssrn.4745624
A1	E4-066	Maier et al. 2026	MA		16 effect sizes from 14 studies	positive	Independent	full_text	https://doi.org/10.48550/arXiv.2605.04779
A1	E4-067	Bono, Cheng & Lozano 2025	RP	162		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2511.13865
A1	E4-084	Afshar et al. 2025	RP	66		positive	Deploying firm	full_text	https://doi.org/10.1056/Aloa2500945
A1	E4-088	Ni et al. 2026	FX	5940		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2603.29888
A2	E4-011	Noy & Zhang 2023	FX	444		positive	Independent	full_text	https://doi.org/10.1126/science.adh2586
A2	E4-015	Cui et al. 2025/2026	FX	1746		null	Vendor	full_text	https://doi.org/10.1287/mnsc.2025.00535
A2	E4-037	Bono & Xu 2024	RP	182		null	Vendor	full_text	https://doi.org/10.48550/arXiv.2411.01067

#	Row	Source	Type	N	k	Call	Run by	Checked against	DOI or URL
A2	E4-043	Xu et al. 2025	FX	1699		positive	Independent	full_text	https://doi.org/10.48550/arXiv.2510.10165
A2	E4-056	Sinayev & Courtney 2025	FX	320		positive	Vendor	full_text	https://doi.org/10.21203/rs.3.rs-6481789/v1
A2	E4-060	Gambacorta et al. 2024/2026	FX	1219		positive	Vendor	full_text	https://doi.org/10.1016/j.jfs.2026.101543
A2	E4-089	Ni et al. 2026	FX	5940		positive	Vendor	full_text	https://doi.org/10.48550/arXiv.2603.29888
A3	E4-048	Bottesch et al. 2026	RP	128		positive	Deploying firm	full_text	https://doi.org/10.48550/arXiv.2607.25922
A4	E4-016	Cui et al. 2025/2026	FX	1746 (about half long tenure)		null	Vendor	full_text	https://doi.org/10.1287/mnsc.2025.00535
A4	E4-044	Xu et al. 2025	FX	1699		positive	Independent	full_text	https://doi.org/10.48550/arXiv.2510.10165
A4	E4-061	Gambacorta et al. 2024/2026	FX	1219		null	Vendor	full_text	https://doi.org/10.1016/j.jfs.2026.101543
A6	E5-002	Vasta & Biondi 2025	MA	1482 (all 41 studies)	37	positive	Independent	full_text	https://doi.org/10.1177/00187208251323132
A6	E5-003	Marcoccia, Quattrociochi & Capraro 2026 (Study 2)	RP	812		positive	Independent	full_text	https://doi.org/10.48550/arxiv.2607.13562
A6	E5-004	Marcoccia, Quattrociochi & Capraro 2026 (Study 3)	RP	843		positive	Independent	full_text	https://doi.org/10.48550/arxiv.2607.13562
A6	E5-005	Marcoccia, Quattrociochi & Capraro 2026 (Study 4)	RP	853		positive	Independent	full_text	https://doi.org/10.48550/arxiv.2607.13562
A6	E5-006	Warning about AI error 2026 (Experiment 3)	RP	243		positive	Independent	full_text	https://doi.org/10.1186/s41235-026-00726-w
A6	E5-007	Jabbour et al. 2023	FX	457		positive	Independent	full_text	https://doi.org/10.1001/jama.2023.22295
A6	E5-008	Jabbour et al. 2025 (selective prediction)	RP	259 (125 in Clinician + AI arm)		positive	Independent	full_text	https://doi.org/10.48550/arxiv.2508.07617
A6	E5-010	Vicente & Matute 2023 (Experiment 2)	RP	199		positive	Independent	full_text	https://doi.org/10.1038/s41598-023-42384-8
A6	E5-012	Agudo, Liberal, Arrese & Matute 2024 (Experiment 2)	RP	260 (132 + 128)		positive	Independent	full_text	https://doi.org/10.1186/s41235-023-00529-3
A7	E5-031	Bahodi, van Berkel, Skov & Merritt 2024	RP	21 (22 recruited, 1 excluded)		null	Independent	full_text	https://doi.org/10.1145/3686038.3686069

Graded tests by check, publication type and test type:

Measure	Value	n
checked_against	abstract	10
checked_against	full_text	78
pub_type	conference	8
pub_type	journal	49
pub_type	preprint	22
pub_type	working_paper	9
test_type	FX	27
test_type	MA	13
test_type	RP	48

8. Effect sizes

Each estimate is reported in its native metric: milliseconds, seconds, percent change, standardized mean differences, Fisher's z or log points. The planned conversion of the best independent estimate and the original's estimate to a common d, and the shrinkage ratio between them, was not computed: for most findings the original and the best test use different metrics or designs (for example a within-person switch cost against a meta-analytic correlation), so a ratio would compare unlike quantities. This is a deviation from protocol section 8.

#	Original	Original estimate (metric)	Best independent evidence	Estimate (metric)	Bias-adjusted	Call
D1	Rubinstein, Meyer & Evans 2001		Gade, Declerck, Philipp, Rey-Mermet & Koch 2021 (meta-analysis, language switching) (deciding meta-analysis (largest k))	27.13 [21.58, 32.68] (ms (beta for transition, contrast coded -1 repeat / +1 switch))		positive
D2	Pashler 1994		Horrey & Wickens 2006 (meta-analysis) (deciding meta-analysis (largest k))			positive
D3	Altmann, Trafton & Hambrick 2014	5.8 vs 1.8 (post-interruption vs positions 2-5) (percent of trials)	Adler, Rubin, Mohammad, Irfan et al. 2015 (largest graded test by N)	abstract: no significant differences in conditions (score)		null
D4	Mark, Gudith & Klocke 2008	22.77 min baseline vs 20.31 (same context) and 20.60 (different context); stress 6.92 vs 9.46 and 9.13 (minutes; NASA-TLX style 1-20 scale)	Speier, Valacich & Vessey 1999 (Exp 1, simple and complex tasks) (largest graded test by N)	simple 90.8 vs 110.3 s (interrupted faster); complex 760.8 vs 608.3 s (interrupted slower) (seconds)		null
D5						
D6	Ophir, Nass & Wagner 2009	167 (ms difference HMM minus LMM)	Parry & le Roux 2021 (performance-based subgroup) (deciding meta-analysis (largest k))	0.091 [0.044, 0.139] (Fisher's z)	0.032 trim-and-fill	null
D7	Sanbonmatsu, Strayer, Medeiros-Ward & Watson 2013	0.08 (r)				
D8	Watson & Strayer 2010	2.5 (percent)				
D9	Kushlev & Dunn 2015	0.37 (Cohen's d (within-person, ESCI conversion))	Ohly & Bastin 2023 (largest graded test by N)	0.06 (direct effect of intervention on irritation (total effect not reported); indirect effect via interruption frequency in notes)		null

#	Original	Original estimate (metric)	Best independent evidence	Estimate (metric)	Bias-adjusted	Call
D10	Coviello, Ichino & Persico 2015 (JEEA; empirical companion of AER 2014)	2.1 (% change in duration per 1% increase in task juggling (IV))	Goes, Ilk, Lin & Zhao 2018 (largest graded test by N)	16.15 [15.423, 16.878] (matched difference in seconds, multitasking vs not)		positive
D11	Adamczyk & Bailey 2004	annoyance 56% lower at best vs worst moment (TLX-style rating 0-5)	Hirsch, Moretti, Leichtmann, Koch & Nitsch 2025 (largest graded test by N)	515 (boundary, before subtask 4) vs 485 (before 2) and 639 (before 3) (ms)		null
A1	Peng et al. 2023	55.8 [21, 89] (% reduction in completion time)	Coupe & Wu 2025 (deciding meta-analysis (largest k))	0.167 (proportional change)	0.167 (PEESE 3L-VCV, SE 0.028 ***); PET 0.074 (SE 0.028 *) PEESE (authors preferred); PET also reported	positive
A2	Brynjolfsson, Li & Raymond 2023	34 (% change (log points))	Ni et al. 2026 (largest graded test by N)	0.072 (log points; interaction Treat x Q5)		positive
A3	Dell'Acqua et al. 2023	-0.245 (difference in share correct, GPT+Overview vs control (linear probability, Table 4))	Bottesch et al. 2026 (largest graded test by N)	qualitative 1.55 vs 1.86; quantitative 2.50 vs 2.88 (expert rating means)		positive
A4	Becker et al. 2025	19 [2, 39] (% change in completion time)	Cui et al. 2025/2026 (largest graded test by N)	13 (% change)		null
A5	Becker et al. 2025	measured +19; forecast -24; post-hoc estimate -20 [2, 39] (% change in completion time)				
A6	Parasuraman & Manzey 2010		Vasta & Biondi 2025 (deciding meta-analysis (largest k))	-0.513 [-0.733, -0.294] (standardized mean change (SMC))	Begg rank correlation test for funnel asymmetry nonsignificant (tau = 0.099, p = .398); no adjusted estimate	positive
A7	Olsen & Wood 2004		Bahodi, van Berkel, Skov & Merritt 2024 (largest graded test by N)	errors chi2(3) = 16.695, p < .001, significant pairs involve alert condition only; auto-resolves chi2(3) = 38.957, p < .001, 6 robots without alerts higher than 3 robots without alerts (p = .001) (Friedman and Bonferroni-corrected Wilcoxon tests)		null
A8						

9. Popular-numbers audit

For each of the 12 numbers fixed in the protocol, the primary source was found and checked for the number and what was measured, and a review pass re-checked it; the audit script applies the five categories in protocol order. For the four numbers with no graded finding behind them (Q4, Q6, Q7, Q12), an independent measurement counts only if its value was read in the source itself.

#	Popular number	Primary source	DOI or URL	Category	Basis	What was measured
Q1	Switching between tasks costs up to 40% of productive time	APA web page "Multitasking: Switching costs" (citing remarks by David Meyer); Rubinstein, Meyer & Evans 2001 checked at abstract level	https://www.apa.org/topics/research/multitasking; 10.1037/0096-1523.27.4.763	No evidence source	no primary evidence source reports the number	Rubinstein 2001: per-switch time costs in lab tasks (APA page: a few tenths of a second per switch); no measurement of share of productive time
Q2	It takes about 23 minutes (23 minutes 15 seconds) to get back to a task after an interruption	Gallup Business Journal interview with Gloria Mark, 8 June 2006 (origin); peer-reviewed counterpart Mark, Gonzalez & Harris 2005 (CHI)	https://news.gallup.com/businessjournal/23146/too-many-interruptions-work.aspx ; 10.1145/1054972.1055017	No evidence source	no primary evidence source reports the number	Elapsed time from interruption until an interrupted working sphere was resumed, among work resumed the same day, in shadowing of information workers; about two other working spheres intervened
Q3	Only about 2.5% of people can multitask without losing performance	Watson & Strayer 2010	10.3758/pbr.17.4.479	Stretched	Measured in one sample of 200 University of Utah undergraduates, on one task pairing (simulated driving with auditory OSPAN), in one session, with no test-retest. The quote extends it to all people and all multitasking. The source rule also counted people with a decrement on 1 of 4 measures (4 of the 5 supertaskers), so without losing performance overstates even the source classification.	5 of 200 undergraduates met supertasker criteria (top quartile single-task on all 4 measures and no decrement on 3 or 4 of 4 measures) for simulated driving with auditory OSPAN
Q4	Juggling email and messages lowers IQ by 10 points, more than smoking cannabis	Wilson, G. 'The Infomania Study' (author's note on drglennwilson.com; HP/Porter Novelli press release 2005)	http://drglennwilson.com/Infomania_experiment_for_HP.doc	No evidence source	no primary evidence source reports the number	matrices-type IQ test, quiet vs phones ringing and emails arriving, within-person

#	Popular number	Primary source	DOI or URL	Category	Basis	What was measured
Q5	Heavy media multitaskers are worse at filtering distractions and switching tasks	Ophir, Nass & Wagner 2009	10.1073/pnas.0903620106	Did not hold	D6 Mixed	lab filter task (change detection with distractors) and cued task-switch cost in extreme groups of a media multitasking index
Q6	Attention on a screen lasts about 47 seconds on average	Mark, Iqbal, Czerwinski, Johns & Sano 2016 (CHI)	10.1145/2858036.2858202	Did not hold	independent measurement within 2x: no	duration an application window stays in the foreground before a window switch (logged), averaged per person
Q7	Knowledge workers switch tasks about every 3 minutes	Gonzalez & Mark 2004 (CHI)	10.1145/985692.985707	Holds	independent measurement within 2x: yes	shadowing observation of continuous 'events' (any activity unit, e.g. a call or spreadsheet use) before switching
Q8	Developers finish tasks 55% faster with GitHub Copilot	Peng et al. 2023	https://arxiv.org/abs/2302.06590	Stretched	Source measured one standardized task (implement an HTTP server in JavaScript) by 95 Upwork freelancers recruited online (p3, p5); p7 says it examines a standardized programming task instead of collaborative work on large projects. The quote presents this as how fast developers finish tasks in general (best case as typical).	Completion time for one standardized task (HTTP server in JavaScript)

#	Popular number	Primary source	DOI or URL	Category	Basis	What was measured
Q9	AI tools made experienced developers 19% slower, though they believed they were 20% faster	Becker et al. 2025	https://arxiv.org/abs/2507.09089	Stretched	Source is titled Early-2025 AI on experienced open-source developers, 246 issues in mature repositories they know (p1-2). Section 4.1 says: We caution readers against overgeneralizing on the basis of our results; Table 2 (p17) lists claims the study does not support, incl. that AI does not speed up many or most developers. The usual quote states a general effect of AI tools on experienced developers (time frame and population extrapolation). Units also shift: source reports 19% more time and a post-hoc estimate that AI reduced time by 20%; the quote says 19% slower and 20% faster.	Completion time per issue (+19%); post-hoc self-estimate (-20%)
Q10	Consultants using GPT-4 finished 12.2% more tasks, 25.1% faster, at 40% higher quality	Dell'Acqua et al. 2023	10.2139/ssrn.4573321	Stretched	The 12.2%, 25.1% and 40% figures hold only for 18 tasks selected to be inside the AI frontier (WP p4; Tables 2-3). On the outside-frontier task AI users were 19 pp less often correct (WP p18; Table 4 GPT+Overview -0.245). The quote presents the inside-frontier best case as general consultant performance; the peer-reviewed version reports quality +32% (Org Sci 2026 p11) and more than 30% (p17), not 40%.	Tasks completed, time and quality on 18 inside-frontier tasks

#	Popular number	Primary source	DOI or URL	Category	Basis	What was measured
Q11	Generative AI raised customer-support productivity by 14%, and by 34% for novices	Brynjolfsson, Li & Raymond 2023	10.3386/w31161	Did not hold	A1 Replicated; A2 Mixed	Resolutions per hour, staggered rollout
Q12	Workers toggle between apps about 1,200 times a day and lose about 4 hours a week reorienting	Murty, Dadlani & Das 2022, HBR article (Soroco authors; Scout work-graph data)	https://hbr.org/2022/08/how-much-time-and-energy-do-we-waste-toggling-between-applications	No evidence source	no primary evidence source reports the number	app and website toggles logged by vendor software; reorientation time estimated

Counts: 1 Holds, 0 Single study, 3 Did not hold, 4 Stretched, 4 No evidence source. Notes: Q2's 23 minutes 15 seconds appears only in a 2006 Gallup interview; the peer-reviewed counterpart (Mark, Gonzalez & Harris 2005) reports 25 min 26 s, measured as time until an interrupted project was resumed, with about two other projects worked in between. Q6's independent value (Reeves et al. 2021, a median of 20 seconds between content switches on 30 students' laptops) is a different unit from Mark et al.'s mean time per foreground window; it also falls just outside a factor of two of Mark's own median (40.2 s). Q12's figures could not be read in the paywalled article; the category does not depend on them.

10. Who ran the agent-side tests

Each agent-side test is flagged Vendor (an author works for the tool's maker, or the maker ran or funded it), AI industry, Deploying firm or Independent. Of the 33 graded A1-A5 and A8 tests, 14 are Vendor (share 0.4242), 0 AI industry, 6 Deploying firm and 13 Independent; over all 36 qualifying agent-side tests (those that report a test of the claim) before the independence filter, 17 are Vendor or AI industry (0.4722).

11. Is supervising agents multitasking?

Components (protocol 10a): a component's pre-registered verdict is supported when every finding it rests on grades Replicated or Thin, partly supported when some do.

Component	Findings (grades)	Pre-registered verdict	Status
Switching between agent threads	D1 Replicated; D3 Replicated; D5 Not tested again	Multitasking: switch costs apply per switch	partly supported
Answering agent prompts mid-task	D3 Replicated; D9 Mixed; D11 Mixed	Multitasking: interruption costs apply, smaller at breakpoints	partly supported
Waiting while agents run	D2 Replicated	Not multitasking when waits are idle or filled with unrelated work; the cost is idle time (D2 applies only if the person watches while doing other work)	supported
Watching agents run	A6 Replicated	Different kind of work: vigilance and complacency costs, not switch costs	supported
Reviewing finished output	A3 Thin; A6 Replicated	Different kind of work: review errors and over-reliance	supported
Writing specs and delegating		Different kind of work: closer to deep work	no graded finding

Fan-out (protocol 10b): from every A7 study (any size) whose capacity value the reviewer verified, the number of agents at which performance peaked or began to fall. Only one A7 study meets the size rules, and it reports no capacity point, so the synthesis runs over small studies and is labelled as such.

Study	Agents per operator tested	Operators	Capacity	Basis	Review
Crandall & Cummings 2007	2; 4; 6; 8	12	5.0	decline_point	verified

Study	Agents per operator tested	Operators	Capacity	Basis	Review
Cummings & Guerlain 2007	8; 12; 16	42	12.0	decline_point	verified
Velagapudi et al. 2008	4; 8; 12	not verified	8.0	peak	verified
Zheng, Glas, Kanda, Ishiguro & Hagita 2011	1 to 8	15	3.0	peak	verified

Median 6.5, min 3, max 12, over 4 studies. Studies that varied the number of agents without locating a limit: Bahodi et al. 2024 (3 vs 6 robots, no drop in errors), Squire & Parasuraman 2010 (4 vs 8, no change in switch costs; abstract), Shoffner & Feng 2024 (1 vs 2, accuracy comparable; abstract), Dahiya et al. 2023 (4 vs 9), Hunt et al. 2026 (swarm size changes, no capacity point) and Mancuso et al. 2015 (span of control did not vary with agents available).

Model (protocol 10c): each agent works alone for a neglect time N, then needs interaction time I, and each move to its thread adds a switch cost R. One person is fully busy at $n^* = (N + I + R) / (I + R)$. Past n^* , agents queue. The share of supervision time spent switching at the limit is $R / (I + R)$. If the person also has their own task, each check-in that pulls them out costs R_{own} , and checking k agents in one block costs one R_{own} instead of k (the study sets $R_{own} = R$). The grid was fixed before any data: N in {2, 5, 10, 20, 30, 60} minutes, I in {1, 2, 5, 10}, R in {0.1, 0.5, 2, 5, 10}. On it, n^* runs from 1.1 to 55.545; at N = 30, I = 2 and R = 2 it is 8.5, with a switch share of 0.5. In that cell, with 5 agents, checking each separately means 8.824 exits from one's own task per hour (17.647 minutes lost), and checking all 5 in one block 1.765 (3.529 minutes).

Literature values: lab resumption lags of about 0.5 to 2 seconds (Monk et al. 2002 and 2004; Hirsch et al. 2025; Katidioti et al. 2016), longer resumption phases in simulated office work (14.3 s and 112.0 s; Zijlstra et al. 1999) and field times to return to an interrupted task of 8 to 29 minutes (Iqbal & Horvitz 2007; Mark, Gonzalez & Harris 2005). The field values include other work done in between, so they are upper bounds on R. Agent run times: median turn lengths of 45 seconds (McCain et al. 2026, Anthropic) and 63.4 seconds (Liu et al., Microsoft), both vendor telemetry, and a 110-minute task horizon at 50% success (Kwa et al. 2025, METR), which measures human task length a model can complete, not how long agents run unattended.

n^* at the literature run times, over the I and R grid:

N (minutes)	Source	Evidence status	n^* min	n^* max
0.75	McCain et al. 2026 (Anthropic)	vendor or non-peer-reviewed	1.038	1.682
110.0	Kwa et al. 2025 (METR)	peer reviewed or preprint	6.5	101
1.0567	Liu, Qiu, Goiri, Fonseca, Bianchini & Choukse 2026 (Microsoft)	vendor or non-peer-reviewed	1.053	1.961

12. Employer language in job postings (Ken's web archive)

Not completed. The server running the fetch stopped answering on 2026-10-16, during the second fetch run, and was still unreachable on 2026-10-17. At the last checks before the outage, 6,353 postings had been fetched and samples were saved for 12 of the 18 year-and-board cells. These counts were read directly from the fetch log; the aggregation never ran on the full panel. No share is reported: the aggregate tables were never retrieved, and a partial panel would mix complete and missing years. The pre-registered web archive predictions are not assessed, and the article and post do not use this data.

This part of the study will be finished and published once the posting panel is complete.

13. Verification

Independent re-analysis (literature numbers)

An independent re-analysis was written from the study plan alone (hypothesis, plan and logged decisions), the protocol and four reviewed extracts, without seeing our code. It was told to ignore the audit categories proposed during extraction and wrote its own code. It did not see any study output.

Headline number	Study	Verifier	Match
Deep-work grades (Replicated, Thin, Mixed, Failed, Not tested again)	3, 0, 5, 0, 3	3, 0, 5, 0, 3	yes
Agent grades	2, 1, 2, 1, 2 (after the fix below; first run 3, 0, 2, 1, 2)	2, 1, 2, 1, 2	yes, after the fix
Grade of each of the 19 findings	the study's grades	same 19 grades	yes, after the fix
Predicted grades right	10	10	yes

Headline number	Study	Verifier	Match
Audit category of Q1-Q12 (Holds, Single study, Did not hold, Stretched, No evidence source)	1, 0, 3, 4, 4, same category for every number	same	yes
Vendor share of graded A1-A5 and A8 tests	14 of 33, 0.4242 (14 of 34, 0.4118, until the second verification excluded Xu et al. 2025)	14 of 33, 0.4242	yes, after the second verification
Fan-out: studies, median, min, max	4, 6.5, 3, 12	4, 6.5, 3, 12	yes
Model grid n* min and max; n* at N = 30, I = 2, R = 2	1.1, 55.545; 8.5	1.1, 55.5455; 8.5	yes
n* at the literature run times (45 s, 63.4 s, 110 min)	1.038-1.682; 1.053-1.961; 6.5-101	1.0375-1.6818; 1.0528-1.9606; 6.5-101	yes

Why the first run did not match, and what was fixed:

- 1. Brodeur et al. 2026 (A1 and A3).** The logged decisions say cluster-randomized experiments need at least 200 randomized clusters, a rule logged for this study by name. Brodeur et al. randomized 288 researchers into 103 teams. The verifier applied the rule; the grading script applied it only to rows coded FX, and this row is coded RP, so it counted the 288 people. The script was fixed (logged as a decision). A3 moved from Replicated to Thin, A1 lost one null test, and the vendor share moved from 14 of 36 to 14 of 34. Counted on its 288 participants, Brodeur et al. would make A3 Replicated (the verifier's sensitivity check shows the same).
- 2. Xu et al. 2025 (E4-042, A1).** The verifier excluded it, reading its analysed unit as 2,755 repositories. The study first kept it, reading the reviewer's note as confirming a contributor count for the analysed sample. The second verification below excluded it on the same grounds, and the row's reviewer note says the project-level test uses 2,755 repositories, not the separate 1,699-contributor sample, so it is now excluded (logged as a decision). No grade depends on it.
- 3. Counting convention.** For branch-1 findings the verifier reports P, Z and T over the non-meta-analytic tests only; the study's tables count the meta-analyses as tests too. Grades are unaffected.

The verifier also listed the remaining judgment calls: the branch-1 reading (D6 is Mixed under the study's reading, Failed under the literal nesting; Q5 is Did not hold either way); Crandall & Cummings' "four to six" entered as 5 (the fan-out median is 6.0 if read as 4, 7.0 if read as 6); and the McCain et al. 99.9th-percentile turn (above 45 minutes, a lower bound), which the study's model table leaves out as a lower bound and which gives n* of 3.25 to 41.9091.

Second independent re-analysis

A second independent re-analysis, run later and again written without seeing our code, got the study plan's question, hypothesis and plan, the rule text of the logged decisions (not the stated reasons, which name some resulting grades), the protocol and the same four reviewed extracts. A comparison script checks 179 of its numbers with the study's outputs at the reported precision: the grade counts, each finding's grade, test counts and teams, the include decision for each of the 159 test rows, the 12 audit categories, the vendor shares, the fan-out, the model grid, the batching values and n* at the literature run times.

- **Against the outputs after the first verification,** 168 of 179 matched. The 11 misses came from three places where the code did not apply a logged rule as written: Xu et al. 2025 (above); the vendor share over all qualifying tests counted five rows that report no test of the claim (protocol section 5, item 4); and the team grouping dropped author names written as "Surname Initials", so E5-006 (Vicente & Matute) counted as its own team in A6. It also found the cluster-rule error independently. Each fix is logged as a decision; no grade changed.
- **After the fixes, 179 of 179 match,** including the vendor shares (14 of 33, 0.4242; 17 of 36, 0.4722), A1's 19 positive of 22 graded tests and A6's 4 teams.
- **Calls:** 80 of the 88 graded calls could be recomputed mechanically from the reviewed numbers, and all 80 match; the 5 naive calls of bias-adjusted meta-analyses also match. The other 8 (E2-009, E2-015, E2-028, E2-041, E2-042, E2-044, E2-046, E5-031) rest on significance statements without a stated direction or on contrasts that disagree, so the reviewed call is used. Its sensitivity check: E2-009 positive would make D11 Replicated, E5-031 positive would make A7 Thin, and two of E2-015, E2-041 and E2-042 not positive would make D3 Mixed.
- **Falsifiers:** none fired in either recomputation.

Web archive shares

Not run: the panel did not complete (section 12).

Alternative explanations

- 1. Access, not evidence, sets some grades.** Closed publishers kept some tests out (32 rows unverifiable) and 10 of the 88 graded tests were checked against an abstract only. Several grades sit one test from a different grade: D11 and A2 need one more positive for

Replicated; A4 Mixed rests on one positive test and would be Failed without it; A7 Failed rests on one null test; A3 is Thin only because Brodeur et al. falls below the cluster rule. These are stated as fragile in the report rather than ruled out.

2. **The size rules decide what counts.** Small studies are never graded, which is why D5, D7, D8 and most multi-robot studies are Not tested again or ungraded, and why the fan-out median comes from small studies. This is a feature of the pre-registered design and is named as a limitation.
3. **Publication bias inflates the Replicated grades.** Branch 1 uses bias-adjusted estimates where reported: A1 stays positive under PET (0.074), and D6 is Mixed precisely because its bias-adjusted estimate is null. A6's deciding meta-analysis reports no bias adjustment, and D3 rests on primary tests with no meta-analysis, so both remain exposed to unpublished nulls. Named as a limitation.
4. **Vendor-run tests drive the AI grades.** The deciding A1 meta-analysis is independent, and without Vendor or AI-industry tests A1 still has 12 positive against 3 null or opposite tests. The vendor tests matter elsewhere: both null A2 tests and both null A4 tests are vendor-run, so independent tests alone would grade A2 Replicated (2 positive, 0 null) and A4 Thin. The report states this.
5. **Lab and robot evidence may not transfer to people supervising LLM agents.** Not ruled out. The component mapping and the model are arguments by analogy, the model is labelled as a model, and A8 (the direct test) is Not tested again.
6. **Errors in screening, extraction and review.** Every value carries a location and quote and was re-opened in a separate review pass, and the headline numbers were recomputed independently; errors that extraction and review share would survive. Named as a limitation.
7. **Scale.** The replicated switching costs (D1-D3) are milliseconds to seconds per switch in lab tasks, while an agent check-in takes minutes, so these costs may be small next to review and waiting time. The study does not measure that ratio; the model only shows how the answer depends on run time, check-in time and switch cost. Limitation.
8. **Search gaps.** No Semantic Scholar keyword grid, OpenAlex, PsycINFO, Web of Science or Scopus, and a single screening pass on titles and abstracts. Forward citations and the 733 logged targeted searches and lookups reduce this. Limitation.
9. **No head-to-head test.** No graded study compares a focused schedule with a parallel-agent schedule for the same person, so the grades alone cannot pick a default; the guidance rests on the component mapping and the model, and Not tested again is not a verdict that a finding is false. Limitation.

14. Limitations

- **Coverage.** The search used Crossref, Europe PMC, arXiv, Semantic Scholar citations, web search and targeted lookups. PsycINFO, Web of Science, Scopus, Google Scholar and OpenAlex were not available, Semantic Scholar's keyword search was blocked, and Peng et al. 2023's citing papers were lost to an overwritten correction pass. Tests indexed only in those sources may be missing.
- **Access.** Many publishers block automated access. 10 of the 88 graded tests were checked against an abstract only, 32 extracted rows were unverifiable and dropped, and some checks used author copies or preprints instead of the version of record. Several possibly qualifying studies (for example Chen & Barnes 2012 for A7, Gurlek & KC 2026 and Zhang & Narayandas 2026 for A2) could not be graded for lack of access.
- **Screening, extraction and review.** Every value was traced to a location and checked in a separate review pass; values the review could not verify were dropped. Errors that extraction and review share would not be caught. Some search-log fields (web-search result counts and times) were reconstructed after the fact rather than captured live, so the logs show which searches ran, not exact hit counts.
- **Thresholds and counting.** The grading rule counts tests rather than weighting them by size, and its 3-to-1 bar means a single test can move a grade: D11 and A2 are each one positive test short of Replicated. Small studies below the size rules are listed but never graded, which is why D5, D7, D8 and most of the multi-robot literature are Not tested again or ungraded.
- **Team independence** is judged by shared authorship with the original only. Tests by the same later team count as separate tests (for example the two software-developer studies under D10: Vasilescu et al. 2016 on open-source projects and Kohl et al. 2020 on one company's GitLab).
- **Who ran the tests.** 14 of the 33 graded AI-productivity tests are by the tool's maker; 21 of the 43 agent-side tests are preprints and 9 working papers, not yet peer reviewed.
- **Transfer.** Most deep-work evidence comes from lab tasks and students; the AI evidence from coding, customer support, consulting and writing; the multi-agent evidence from robot and drone operators. None of it measures people supervising several LLM agents, so the component mapping and the model are arguments by analogy.
- **The model** is an illustration with fixed inputs, not an estimate. Agent run times come from vendor telemetry, and field resumption times are upper bounds.
- **The audit** tests the numbers as they are usually quoted, in wording fixed before the search; other phrasings of the same number could fare differently.

- **Web archive panel not completed.** The server running the fetch became unreachable, so no employer-language shares are reported and the pre-registered web archive predictions are not assessed (section 12).
- **Grades are as of 2026-10-16.** "Failed" means the qualifying tests did not find the effect, not that it is zero in every setting; "Not tested again" means no qualifying test exists yet.

15. Sources

Originals, graded tests and audit sources are listed with their DOI or URL in the tables above. Context sources quoted in the article and this document: McCain et al. 2026, Anthropic, on the turn lengths of its coding agent (vendor; context only) <https://www.anthropic.com/research/measuring-agent-autonomy>; Liu et al., Microsoft, on Copilot agent turns (vendor; context only) <https://arxiv.org/abs/2608.00101>; Kwa et al. 2025, METR, measuring AI ability to complete long tasks, <https://arxiv.org/abs/2503.14499>; the APA topic page on multitasking, <https://www.apa.org/topics/research/multitasking>; the 2006 Gallup Business Journal interview, <https://news.gallup.com/businessjournal/23146/too-many-interruptions-work.aspx>; Murty, Dadlani & Das 2022, Harvard Business Review, <https://hbr.org/2022/08/how-much-time-and-energy-do-we-waste-toggling-between-applications>. Replication status: every graded finding's status is its grade in section 7; context sources are not replication evidence.

Ken Research · Deep work or parallel AI agents: what the research supports · Cristian Frunze, Founder of Ken